AUTOMATED WRITING EVALUATION FOR FORMATIVE SECOND LANGUAGE ASSESSMENT: EXPLORING PERFORMANCE, TEACHER USE, AND STUDENT ENGAGEMENT

By

SVETLANA KOLTOVSKAIA

Bachelor of Arts in Foreign Languages and Literature
Mirny Polytechnic Institute
Mirny, the Republic of Sakha (Yakutia), Russia
2009

Master of Arts in TESOL
Central Michigan University
Mount Pleasant, Michigan
2015

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
July, 2022

AUTOMATED WRITING EVALUATION FOR FORMATIVE SECOND LANGUAGE ASSESSMENT: EXPLORING PERFORMANCE, TEACHER USE, AND STUDENT ENGAGEMENT

Dissertation Approved:

Dr. Stephanie Link

Dissertation Adviser

Dr. An Cheng

Dr. Anna Sicari

Dr. Penny Thompson

Name: SVETLANA KOLTOVSKAIA

Date of Degree: JULY, 2022

Title of Study: AUTOMATED WRITING EVALUATION FOR FORMATIVE SECOND LANGUAGE ASSESSMENT: EXPLORING PERFORMANCE, TEACHER USE, AND STUDENT ENGAGEMENT

Major Field: ENGLISH

Abstract: The purpose of this dissertation is to investigate automated writing evaluation (AWE) from both system- and user-centric perspectives. The system-centric research focused on error-correction/detection performance of the AWE system, Grammarly. The study was based on fifty-three argumentative essay drafts written by undergraduate students enrolled in a second language (L2) writing course. Grammarly's feedback given to those essay drafts was measured using precision (accuracy) and recall (system coverage) and compared to human annotators' feedback. Results revealed that Grammarly's precision rates for flagging and correction (92% and 91%, respectively) exceeded a benchmark of 80%. This means that Grammarly was accurate in detecting and correcting common L2 errors. However, Grammarly's recall rate was low (51%), which means that Grammarly missed half of the errors found by human annotators. Two user-centric studies focused on teachers and students. The first study explored six postsecondary, L2 writing teachers' use and perceptions of Grammarly as a complement to their feedback. The participants' feedback was analyzed to understand Grammarly's impact on their feedback activity. The participants then had a semi-structured interview aimed at exploring their perceptions of Grammarly as a supplementary tool. Findings revealed that despite using Grammarly to complement their feedback, teachers still provided feedback on sentence-level issues. Overall, the majority of teachers were positive about using Grammarly to complement their feedback, notwithstanding its limitations. The second study explored two English as a second language (ESL) college students' behavioral, cognitive, and affective engagement with Grammarly's feedback when revising a final draft. The behavioral engagement was explored through the analysis of QuickTime-based screencasts of students' Grammarly usage. Cognitive and affective engagement were measured through the analysis of students' comments during stimulated recall of the aforementioned screencasts and semi-structured interviews. According to findings, one student showed greater cognitive engagement through his questioning of AWCF but did little to verify the accuracy of feedback, which resulted in moderate changes to his draft. The other's overreliance on AWCF indicated more limited cognitive engagement, which led to feedback's blind acceptance. Nevertheless, this also resulted in moderate changes to her draft. The dissertation provides implications to meaningfully use AWE in L2 writing classrooms.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Automated writing evaluation (AWE) systems analyze written texts and provide instant computer-generated holistic scores assessing writing quality and written feedback aimed at improving issues on global and/or local aspects of writing (Cotos, 2018). For the past several years, AWE have become increasingly used as a formative assessment tool in writing classrooms (Stevenson & Phakiti, 2014), that is, a tool that facilitates learning instead of measuring it (Chen & Cheng, 2008). Such growing use of AWE for instructional purposes could be a result of AWE vendors' good promotion and advertisement. For example, the commercially available AWE system, Criterion developed by Educational Testing Service (ETS) has been advertised as a service that helps teachers "decrease their workload and free up time to concentrate on the content of students' work and teach higher level writing skills" and allows students to "have more opportunities to practice writing at their own pace, get immediate feedback and revise essays based on the feedback" (ETS Criterion, 2022a). Another commercially available system MyAccess! developed by Vantage Learning has been promoted as a "prompt-driven, web based writing environment that scores student essays instantly and provides diagnostic instruction that engages and motivates students to want to improve their writing proficiency" (Vantage Learning, 2004).

Like any educational technology, however, AWE has been the subject of skepticism and criticism. Fears have been expressed about the possibility of AWE taking authority away from a teacher as reader and responder of students' writing (Herrington & Moran, 2006; Rothermel, 2006). Criticism has been made regarding AWE's capacity to provide accurate and meaningful scores and feedback (Jones, 2006) because computers are not capable of "reading," "interpreting", and "evaluating" students' texts as humans do (Anson, 2006; Condon, 2006; McGee, 2006). There have also been doubts about students making good use of AWE for text revision (Attali, 2004; Chapelle, Cotos, & Lee, 2015). Some have been worried that AWE could lead students to focus on form rather than content during the revision process (Hyland & Hyland, 2006; Stevenson & Phakiti, 2014) because AWE systems are more computationally adept at providing feedback on sentence-level issues (Ranalli, Link, & Chukharev-Hudilainen, 2017). Others questioned the value of writing to a computer rather than to a human because there cannot be genuine, meaningful communication between a computer and a writer (Ericsson, 2006; Herrington & Moran, 2006). In their Position Statement on Teaching Learning and Assessing Writing in Digital Environments, the Conference on College Composition and Communication (2004) expressed their concerns about using AWE for assessment purposes and stated that "writing to a machine violates the essentially social nature of writing: we write to others for social purposes;" therefore, according to them, "all writing should have human readers, regardless of the purpose of the writing" (p. 3).

To address the aforementioned concerns, an extensive body of research has been conducted over the past several decades. While some studies have examined AWE from a system-centric perspective, that is, research that “focuses on the performance of the system itself” (Chodorow, Gamon, & Tetreault, 2010, p. 421), others have investigated AWE from a user-centric perspective, that is, research that “examines the user’s behavior or the effect of the system on the user” (Chodorow et al., 2010, p. 421).

Cumulative evidence from both types of research seems to suggest that despite its limitations, AWE has many benefits for instructional use and is often perceived positively by both teachers and students. In accordance with AWE vendors’ claims, research indicates that AWE can save time and allow teachers to focus more on global aspects of writing as it takes care of sentence-level issues (Warschauer & Grimes, 2008; Wilson & Czik, 2016). For students, AWE has been shown to help improve their writing (Dikli 2006), facilitate noticing of errors (Barrot, 2021), increase writing practice (Grimes & Warschauer, 2010), and promote their motivation and autonomy (Link, Dursun, Karakaya, & Hegelheimer, 2014).

In spite of such positive findings, concerns regarding the usefulness of AWE systems persist as new AWE systems with various affordances with language learning potential emerge. However, there seems to be a consensus that “blanket rejection” of AWE does not serve teachers and students (Whithaus, 2006, p. 176). According to Weigle (2013), “AWE is here to stay,” and it would be naive to suggest that they would not be “used for scoring and feedback in the future” (p. 50). Therefore, instead of focusing on whether AWE should be used for assessment purposes in writing classrooms, the current research should find answers to the more pressing question, which is “how this new technology can be used to achieve more desirable learning outcomes while avoiding potential harms that may result from limitations inherent in the technology” (Chen & Cheng, 2008, p. 95).

The “how” question is particularly important for second language (L2) writing instruction and assessment. This is because the largest market for AWE systems is in assessing the writing ability of non-native speakers rather than native speakers of English (Weigle, 2013). This is despite the fact that many AWE systems are designed with native speakers of English in mind (Dikli & Bleyle, 2014). Weigle (2013) argued that while the use of AWE has been controversial in the composition community, it may be less controversial in L2 writing classrooms because the focus of instruction and assessment is often on linguistic or rhetorical concerns (e.g., vocabulary, morphology) than higher-level issues (e.g., genre, audience, voice). However, the focus of instruction and assessment is often determined based on learner variables (e.g., age, education level, proficiency, ESL vs. EFL). For example, at lower levels of language proficiency, the focus of instruction and assessment is generally on linguistic issues. As learners gain more control over their language skills, the focus of instruction and assessment can shift toward higher-level issues (Weigle, 2013). Regardless of the focus of the writing instruction, research suggests that L2 learners want feedback on errors and believe that feedback helps them improve their writing (Ferris, 2014). Since AWE provides feedback on linguistics issues, its implementation in L2 writing classrooms can provide "a meaningful enhancement to student performance" if AWE is used "responsibly" (Brent & Townsend, 2006, p. 198).

The main purpose of this dissertation is to find answers to the "how" question. To achieve this purpose, the dissertation examines Grammarly, a writing assistant tool that is making important inroads into L2 writing classrooms (Ranalli, 2018), from both system- and user-centric perspectives. By examining Grammarly from both perspectives, the dissertation fills in the following gaps in the literature.

Of the extensive body of system-centric research on AWE, the bulk has predominantly focused on AWE systems' holistic scores, especially in the context of high-stakes testing (Attali & Burstein, 2006; Burstein et al., 1998; Burstein & Chodorow, 1999; Enright & Quinlan, 2010; Powers et al., 2002). Limited attention has been paid to the accuracy of AWE's feedback on errors, and the extant literature has primarily focused on such commercially available tools as Criterion, MyAccess!, and Pigai (e.g., Bai & Hu, 2017; Bestgen & Granger, 2011; Chodorow et al., 2007; Han, Chodorow, & Leacock, 2006; Lavolette, Polio, & Kahng, 2015; Liu & Kunnan, 2016; Ranalli et al., 2017; Tetreault & Chodorow, 2008), thus leaving other AWE systems and similar tools underexplored. There has also been an extensive body of user-centric research that has focused on both teachers' and students' use and perceptions of AWE (e.g., Chapelle et al., 2015; El-Ebyary & Windeatt; Dikli & Bleyle, 2014; Lai, 2010; Link et al., 2014; Maeng, 2010; Warschauer & Grimes, 2008). However, research that focuses on teachers' perspectives of the tool, especially when teachers use AWE to complement their feedback, is scarce. Additionally, limited research exists on the impact of AWE on teacher feedback when AWE is used as a supplement to teacher feedback. This is despite the fact that teachers play an integral part in the implementation process of the tool in their L2 classrooms (Li, 2021). Although studies have extensively analyzed students' use and perceptions of AWE as students are direct users of the tool, extant literature made little effort to explore individual students' engagement with automated feedback when they use AWE to revise their drafts. Considering all these gaps, conducting combined system- and user-centric research is imperative as such research can give a comprehensive picture of the tool. Such comprehensive picture of the tool could contribute to the knowledge needed for the appropriate application of AWE for formative assessment in L2 writing classrooms, thus answering the question of "how."

This dissertation includes three interrelated chapters. Chapter II discusses system-centric research that investigated Grammarly's error-correction/detection performance. The study was based on fifty-three argumentative essay drafts written by English as a second language (ESL) learners enrolled in an L2 writing course at a southcentral U.S. university. Grammarly's feedback given to those essay drafts was measured using precision (accuracy) and recall (system coverage), two concepts from the information retrieval field, and compared to human annotators' feedback. Chapter III and Chapter IV focus on user-centric research. Chapter III discusses a study that explored six postsecondary, L2 writing teachers' use and perceptions of Grammarly as a complement to their feedback. The participants' feedback was analyzed to understand Grammarly's impact on their feedback activity. The participants then had a semi-structured interview aimed at exploring their perceptions of Grammarly as a tool to augment their feedback. Chapter IV discusses a study that explored two ESL college students' engagement with automated written corrective feedback (AWCF) provided by Grammarly when revising a final draft. Following previous research, engagement was operationalized as students' response to feedback that has manifestations in the perspective of behavior, cognition, and affect. The behavioral engagement was explored through the analysis of QuickTime-based screencasts of students' Grammarly usage. Cognitive and affective engagement were measured through the analysis of students' comments during stimulated recall of the aforementioned screencasts and semi-structured interviews. The dissertation concludes with Chapter VI which discusses overall findings and provides general recommendations for the use of Grammarly in L2 writing classrooms as a formative assessment tool. The final chapter is followed by a list of references and relevant appendices.

CHAPTER II

GRAMMARLY'S ERROR CORRECTION/DETECTION PERFORMANCE

Introduction

Automated essay scoring (AES) systems, such as *e*-rater by Educational Testing Service (ETS), Intellimetric by Vantage Learning, and Intelligent Essay Assessor (IEA) by Pearson Knowledge Technologies, which are used to assess large-scale, high-stakes tests, such as the GRE and the GMAT, were augmented to Criterion, *My Access!*, and *WriteToLearn*, respectively, for instructional use (Stevenson & Phakiti, 2019; Warschauer & Grimes, 2008). Apart from automated scoring, these systems were extended to the generation of automated feedback on global aspects of writing, including organization and idea development and sentence-level issues, such as grammar and mechanics. The combination of automated scoring and automated feedback is now referred to as automated writing evaluation (AWE) (Cotos, 2018).

AWE systems were originally developed for use by native speakers of English. However, the last few decades have witnessed an increase in the use of AWE systems to assess the writing ability of English language learners (ELLs) (Weigle, 2013). With such increased use of AWE systems, questions have been raised about the ability of AWE systems to score as well as detect and correct errors in essays written by ELLs (Weigle, 2013), thus generating a lot of system-centric research that focuses on the performance of the AWE system itself (Chodorow, Gamon, & Tetreault, 2010). Of the extensive body of system-centric research, the bulk has evaluated AWE's scoring ability, particularly in the context of high-stakes testing (Attali & Burstein, 2006; Burstein et al., 1998; Burstein & Chodorow, 1999; Enright & Quinlan, 2010; Powers et al., 2002). Relatively little research has evaluated the AWE's accuracy in detecting and correcting linguistic errors committed by ELL writers, and the available literature has predominantly focused on ETS Criterion (Chodorow, Tetreault, & Han, 2007; Dikli & Bleyle, 2014; Han, Chodorow, & Leacock, 2006; Lavolette, Polio, & Kahng, 2015; Ranalli, Link, & Chukharev-Hudilainen, 2017; Tetreault & Chodorow, 2008), thus leaving other AWE systems and similar tools underinvestigated. Given that system-centric research is often conducted and funded by AWE systems developers (Liu & Kunnan, 2016) and the information about the actual performance of AWE systems is limited and often inaccessible to the research community (Ranalli et al., 2017), studies conducted by independent researchers are necessary because, as Dikli and Bleyle (2014) said, independent researchers "can provide an outsider perspective to the research in the field" (p. 4).

Therefore, the current study focuses on Grammarly, a tool that is gaining popularity among not only native but also non-native speakers of English (Ranalli, 2018), and evaluates its error-detection/correction performance by measuring its precision (accuracy) and recall (system coverage) and using expert human annotators as a benchmark. The results of the study provide important pedagogical implications as well as suggestions for developers on the improvement of the tool.

Literature Review

Unlike the research that focuses on AWE's holistic scores (e.g., Attali & Burstein, 2006; Burstein et al., 1998; Vantage Learning, 2003a, 2003b, 2006), the research that focuses on AWE's automated feedback is quite limited (e.g., Bai & Hu, 2017; Bestgen & Granger, 2011; Chodorow et al., 2007; Guo, Feng, & Hua, 2021; Han at al., 2006; Lavolette, Polio, & Kahng, 2015; Liu & Kunnan, 2016; Ranalli et al., 2017; Ranalli & Yamashita, 2022; Tetreault & Chodorow, 2008), and the bulk of this research has mainly focused on the accuracy of automated feedback of ETS Criterion (Chodorow et al., 2007; Dikli & Bleyle, 2014; Han et al., 2006; Lavolette et al., 2015; Ranalli et al., 2017; Tetreault & Chodorow, 2008).

The accuracy of AWE's automated feedback in the extant literature has often been measured using precision and recall, two concepts from the information retrieval field (Ranalli et al., 2017). Precision attempts to determine "the proportion of flagged items that are, in fact, usage errors" (Leacock, Chodorow, Gamon, & Tetreault, 2010, p. 38). For example, if out of 50 preposition errors detected by the AWE system 43 are actual preposition errors, the precision rate would be 86% (=.86). Whether this precision rate is high or low depends on the threshold set by system developers. For example, Quinlan, Higgins, & Wolff (2009) stated that Criterion's developers require 80% (or above) precision in testing. In this case, the precision of 86% is considered high, as it meets or even exceeds the threshold of 80% precision set by the Criterion's developers.

Recall attempts to determine "the proportion of actual usage errors that have been flagged", i.e., the system's coverage (Leacock et al., 2010, p. 38). For example, if human annotators find 90 preposition errors in the essay, and AWE detects 50 preposition errors in the same essay, the recall rate would be 56% (=.56). This rate could be considered low, as the AWE system missed half of the preposition errors found by human annotators. It is noteworthy that AWE developers often prioritize precision over recall because flagging a well-formed construction as ill-formed is considered more detrimental than missing an error (Burstein, Chodorow, & Leacock, 2004; Ranalli & Yamashita, 2022). Therefore, the precision rate of many AWE systems tends to be higher than the recall rate. For example, in flagging preposition errors, Tetreault and Chodorow (2008) found that Criterion's precision rate was 84%, and the recall rate was close to 19%. In flagging article errors, Han et al. (2006) found that Criterion's precision rate was 90%, and the recall rate was 40%.

While the aforementioned studies rated the accuracy of Criterion's feedback provided to one error category, such as articles or prepositions, others addressed the array of error categories (Dikli & Bleyle, 2014; Lavolette et al., 2015; Ranalli et al., 2017). For instance, Ranalli et al. (2017) evaluated the precision of Criterion's feedback provided for ten error types commonly identified in English as a second language (ESL) students' writing by adopting the lenient standard threshold of 70%, which they regarded as the lenient standard compared to the stricter standard of 80% required in testing. The results revealed that when considering ten error types in the aggregate, Criterion's feedback was accurate between 71% and 77% of the time, which exceeded the 70% threshold for accuracy. However, among individual error types, there was considerable variation. For example, high accuracy was found for ill-informed verbs (95.7%) and subject-verb agreement (90%), and low accuracy was found for extra comma errors (57.1%). Regrettably, the study did not investigate the errors Criterion missed. Investigating recall is necessary because low recall means the system missed a large proportion of errors, which could be detrimental for students (Ranalli et al., 2017). To better inform applications of AWE systems and similar tools in second language (L2) writing classrooms, research should assess AWE's automated feedback comprehensively; that is, focus on a wide range of error categories instead of one or two error categories. Research should also focus on both precision and recall. Moreover, research should evaluate other available tools besides Criterion.

One underexplored tool that has the potential for formative assessment in L2 writing classrooms is Grammarly. It has been reported that Grammarly detects more error types common to L2 writing compared to, for example, Microsoft Word (Ranalli & Yamashita, 2022). Grammarly can be used for free but users can also purchase a premium subscription at the price of $144 per year (Grammarly, 2022). Grammarly can be accessed through the browser extension, software plug-in, and mobile devices as opposed to many AWE systems, including Criterion, which only allows access through standalone web-based interfaces (Ranalli & Yamashita, 2022). Since Grammarly can be accessed in multiple ways, it delivers feedback both synchronously and asynchronously, while many AWE systems often deliver feedback asynchronously (Ranalli & Yamashita, 2022). Different from other AWE systems that often provide feedback on higher-order (e.g., content, organization) and lower-order (e.g., grammar, mechanics) concerns, Grammarly provides feedback mainly on lower-order concerns; therefore, it has recently been termed as an automated written corrective feedback (AWCF) tool (Ranalli, 2018).

To date, only two studies have attempted to investigate Grammarly's error detection/correction performance. Guo et al. (2021) analyzed the accuracy of Grammarly in detecting and correcting errors in research papers written by university students in China. Guo et al. found that while Grammarly's overall flagging precision rate was 69%, the correction precision rate was 82%. Unfortunately, the researchers did not measure Grammarly's recall rate. Ranalli and Yamashita (2022) examined Grammarly's performance by comparing it to the performance of Microsoft Word's Natural Language Processing (MS-NLP). The results revealed that while Grammarly's total precision rate for flagging was 88%, which is slightly lower than MS-NLP's precision rate of 92%, Grammarly's total precision rate for correction was 81%, which is slightly above MS-NLP's precision rate of 79%. As for recall, Grammarly had higher recall rates for four selected L2 error types than MS-NLP. For example, Grammarly's recall rate for subject-verb agreement was 67% while MS-NLP's was 35%. Ranalli and Yamashita stated that "in the two years between the start of the project and the writing of this report, Grammarly's claims about the number of features it could identify increased from 250 to 400" (p. 14). This means that more studies are needed on Grammarly's performance, considering it is continuously being developed and improved. Additionally, "if [AWE] feedback is to help students improve their writing skills, then it should be similar to what instructor's comments might be" (Burstein et al., 2004, p. 32); therefore, studies should compare Grammarly's performance to that of humans to better understand its strengths and limitations.

To address the gaps in the extant literature, this study took the system-centric approach to answer the following research questions:

RQ1: How accurate is Grammarly in detecting and correcting L2 errors?

RQ2: How does Grammarly's L2 error detection compare to human annotators' error identification?

Methods

Corpus

Fifty-three argumentative essay drafts written by non-native English speakers enrolled in an L2 writing undergraduate course during the fall 2018 semester were extracted from the Wrangler corpus, an electronic collection of texts written by ESL learners at the Southcentral U.S university. In the argumentative essay, which was the first major writing assignment of the semester, the students were supposed to write an article for an imaginary Discover Magazine arguing about one invention the world would be better without (Appendix A). For this assignment, the students were required to submit two drafts: Draft A (rough draft) for formative feedback and Draft B (final draft) for summative feedback. The rough drafts were chosen because of the likelihood of more errors as compared to the final drafts. There were 45,084 words in total in the corpus, and the average text length was 851 words (SD = 264.85).

Along with students' written texts, students' survey responses were extracted from the corpus that contained their demographic information. The essays were written by 16 females and 37 males whose first language (L1) were Arabic (*n* = 29), Chinese (*n* = 17), Korean (*n* = 4), Hungarian (*n* = 1), Icelandic (*n* = 1), and Dahae (*n* = 1). The students' majors were distributed in the following colleges: engineering, architecture and technology (*n* = 22), business (*n*= 6), education and human sciences (*n* = 4), arts and sciences (*n* = 4), and undecided (*n* = 17).

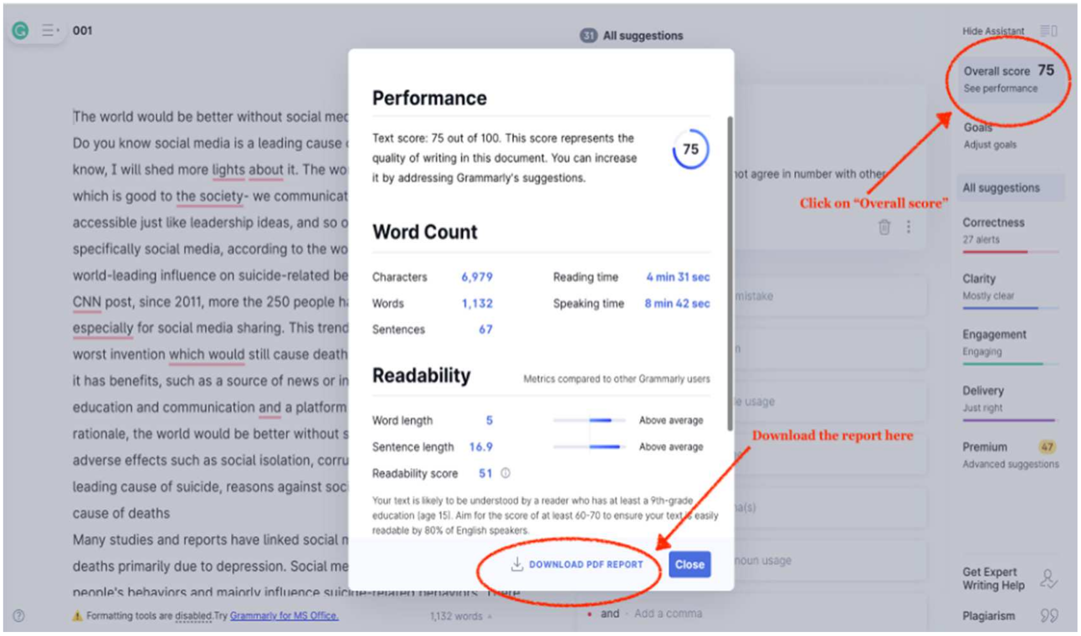
The students' proficiency levels were low-intermediate (*n* = 42) and high-intermediate (*n* = 11). The students' proficiency levels were determined based on the performance score they received on the diagnostic essay they wrote at the beginning of the semester. The researcher and two of her colleagues independently evaluated 53 diagnostic essays using a slightly modified TOEFL iBT Test - Independent Writing Rubric (ETS, 2019). The essays were graded on a scale from 0-5. The scorers then had a meeting in which they calculated the mean score for each essay which then was converted to a scaled score of 0-30 based on the Writing and Speaking Sections of the New TOEFL iBT Test Converting Rubric (ETS, 2022b). For example, one essay received a score of 3, 3, and 4 from the three scorers. The calculated mean score was 3.3 which when converted to a scaled score became 21. According to the Performance Descriptors for the TOEFL iBT Test, the person who receives a score between 17-23 is considered high-intermediate (ETS, 2021).

Procedures

Each student's draft was first uploaded to a free version of Grammarly. The following information from Grammarly was then entered into a Google spreadsheet for each text: *major error type* (e.g., grammar), *revision operation* (e.g., fix the agreement mistake), *correction* (e.g., ~~lights~~ -> light), *specific error category* (e.g., incorrect noun number), *metalinguistic feedback* (e.g., It seems that **lights** may not agree in number with other words in this phrase.), and *sentence* in which an error was flagged (e.g., Well, if you didn't know, I will shed more **lights** about it.) (Appendix B). *Specific error categories* were found in the Grammarly report that can be downloaded by clicking on "Overall score" at the top right (Figure 1).

Figure 1

The Snapshot Demonstrating how Grammarly Report can be Downloaded



The world would be better without social mec
Do you know social media is a leading cause
know, I will shed more lights about it. The wo
which is good to the society- we communicat
accessible just like leadership ideas, and so o
specifically social media, according to the wo
world-leading influence on suicide-related be
CNN post, since 2011, more the 250 people h
especially for social media sharing. This trend
worst invention which would still cause death
it has benefits, such as a source of news or in
education and communication and a platform
rationale, the world would be better without s
adverse effects such as social isolation, corru
leading cause of suicide, reasons against soc
cause of deaths
Many studies and reports have linked social n
deaths primarily due to depression. Social me
people's behaviors and majorly influence suicide-related behaviors. There

Performance

Text score: 75 out of 100. This score represents the quality of writing in this document. You can increase it by addressing Grammarly's suggestions. 75

Word Count

Characters	6,979	Reading time	4 min 31 sec
Words	1,132	Speaking time	8 min 42 sec
Sentences	67		

Readability Metrics compared to other Grammarly users

Word length	5	Above average
Sentence length	16.9	Above average
Readability score	51	

Your text is likely to be understood by a reader who has at least a 9th-grade education (age 15). Aim for the score of at least 60-70 to ensure your text is easily readable by 80% of English speakers.

…not agree in number with other

⚠ Formatting tools are disabled.Try Grammarly for MS Office.

Once the information for each essay was entered, they all were combined to determine Grammarly's flagging and correction frequency for each error type and category. Table 1 shows five major error types that were identified, including grammar, punctuation, spelling, conventions, and conciseness, and each major error type contains several error categories. For example, according to the table, the major error type, grammar, contains ten error categories, such as conjunction use, determiner use (a/an/the/this, etc.), incorrect noun number, faulty subject-verb agreement, incorrect verb forms, modal verbs, misuse of modifiers, misuse of quantifiers, pronoun use, and wrong or missing prepositions. Overall, there were 1518 flaggings and 1518 corrections in the 45,084-word corpus.

Table 1

Grammarly Flagging/Correction Frequency Across Error Types and Categories

	#	%
Grammar	**965**	**63.6**
Conjunction use	2	0.1
Determiner use (a/an/the/this, etc.)	412	27.1
Incorrect noun number	106	7.0
Faulty subject-verb agreement	116	7.6
Incorrect verb forms	43	2.8
Modal verbs	4	0.3
Misuse of modifiers	18	1.2
Misuse of quantifiers	6	0.4
Pronoun use	75	4.9
Wrong or missing prepositions	183	12.1
Punctuation	**153**	**10.1**
Closing punctuation	2	0.1
Comma misuse within clauses	134	8.8
Misuse of semicolons, quotation marks, etc.	1	0.1
Punctuation in compound/complex sentences	16	1.1
Spelling	**179**	**11.8**
Commonly confused words	13	0.9
Confused words	111	7.3
Misspelled words	52	3.4
Unknown words	3	0.2
Conventions	**33**	**2.2**
Improper formatting	29	1.9
Mixed dialects of English	4	0.3
Conciseness	**188**	**12.4**
Wordy sentences	188	12.4
Total	**1518**	**100**

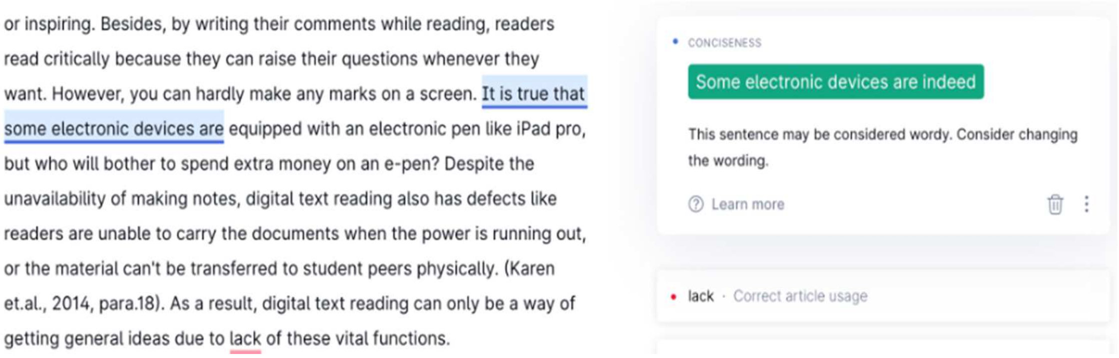
After determining Grammarly's flagging and correction frequency for each error type and category, the information on the spreadsheet was separated again based on the major error types (grammar, punctuation, spelling, conventions, and conciseness) for two reasons: 1) to create the error categorization rubric for human annotators to provide feedback and 2) to measure precision and recall. The error categorization rubric generated based on Grammarly feedback contained major error types, error categories for each major type, Grammarly's metalinguistic feedback that tells how to address the error for each error category, and for each error category, a sentence in which an error was detected was provided (Appendix C).

Analysis

Prior to coding, the author and a professor at a U.S. university, who has a Ph.D. in Applied Linguistics and whose research interest is in corrective feedback, had a meeting to review the error categorization rubric generated based on Grammarly feedback and understand how Grammarly is programmed to flag and correct errors. In the meeting, the coders agreed to exclude two major error types *conventions* and *conciseness* and instead focus on *grammar, punctuation,* and *spelling* errors. The reason for excluding *conventions* was because improper formatting (i.e., spacing errors) and mixed dialects of English (i.e., British vs. American English spelling errors) do not impede the meaning of the sentence. *Conciseness* that focuses on wordy sentences was excluded because it was not always clear how Grammarly identifies such errors. Figure 2 shows one such example in which Grammarly suggested replacing "it is true that some electronic devices are" with "some electronic devices are indeed" in the sentence "It is true that some electronic devices are equipped with an electronic pen like iPad pro, but who will bother to spend extra money on an e-pen?"

Figure 2

Conciseness Error Type that was Removed from the Study



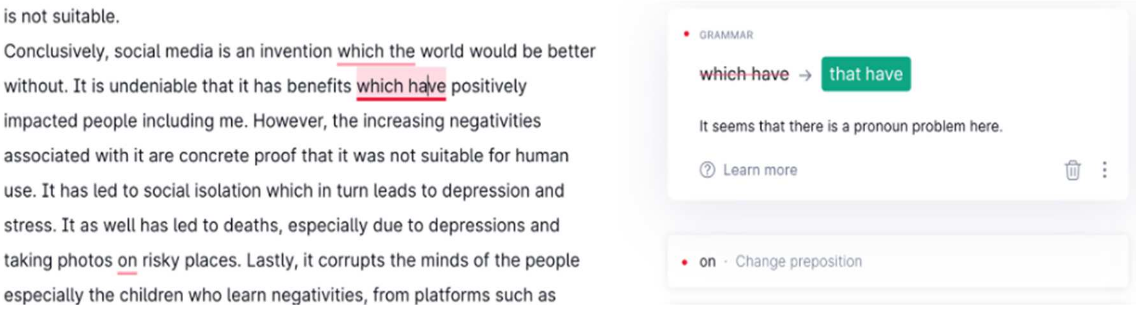
or inspiring. Besides, by writing their comments while reading, readers read critically because they can raise their questions whenever they want. However, you can hardly make any marks on a screen. It is true that some electronic devices are equipped with an electronic pen like iPad pro, but who will bother to spend extra money on an e-pen? Despite the unavailability of making notes, digital text reading also has defects like readers are unable to carry the documents when the power is running out, or the material can't be transferred to student peers physically. (Karen et.al., 2014, para.18). As a result, digital text reading can only be a way of getting general ideas due to lack of these vital functions.

This sentence may be considered wordy. Consider changing the wording.

To measure precision, the coders independently coded 1297 error flaggings and corrections, excluding 33 *convention* and 188 *conciseness* errors. Following Ranalli and Yamashita (2022), each Grammarly feedback unit was coded for flagging and correction accuracies. For both accuracy dimensions, codes such as *accurate*, *inaccurate, and neutral* were used. *Neutral* was used when the flagged item was not a true error and the correction was stylistic. For example, Grammarly suggested replacing "which have" with "that have" in the sentence "It is undeniable that it has benefits <u>which have</u> positively impacted people including me" (Figure 3).

Figure 3

The Code "Neutral" Used for both Flagging and Correction Accuracies



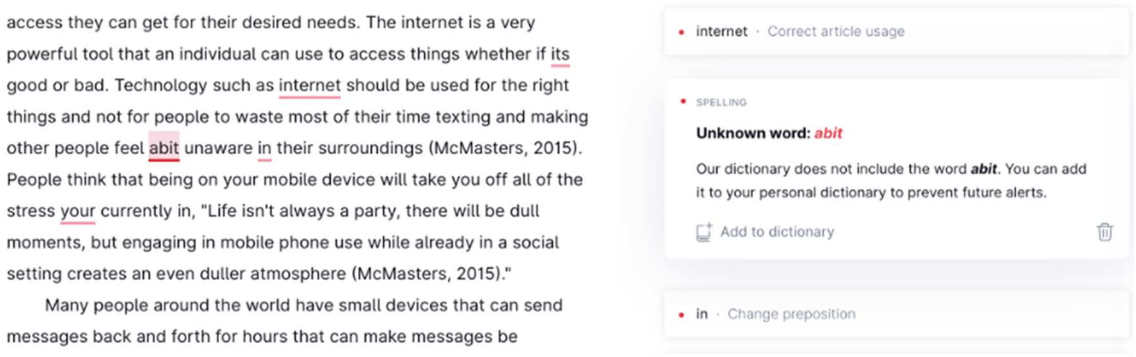
is not suitable.
Conclusively, social media is an invention which the world would be better without. It is undeniable that it has benefits which have positively impacted people including me. However, the increasing negativities associated with it are concrete proof that it was not suitable for human use. It has led to social isolation which in turn leads to depression and stress. It as well has led to deaths, especially due to depressions and taking photos on risky places. Lastly, it corrupts the minds of the people especially the children who learn negativities, from platforms such as

It seems that there is a pronoun problem here.

Similar to Ranalli and Yamashita (2022), the code *unknown* was used for correction accuracy if Grammarly provided no specific suggestion. For example, “abit” was coded as *accurate* for flagging accuracy, but as *unknown* for correction accuracy because Grammarly provided no suggestion for how to correct the error (Figure 4).

Figure 4

The Code “Unknown” Used for Correction Accuracy



access they can get for their desired needs. The internet is a very powerful tool that an individual can use to access things whether if its good or bad. Technology such as internet should be used for the right things and not for people to waste most of their time texting and making other people feel abit unaware in their surroundings (McMasters, 2015). People think that being on your mobile device will take you off all of the stress your currently in, "Life isn't always a party, there will be dull moments, but engaging in mobile phone use while already in a social setting creates an even duller atmosphere (McMasters, 2015)."

Many people around the world have small devices that can send messages back and forth for hours that can make messages be

internet · Correct article usage

SPELLING

Unknown word: ***abit***

Our dictionary does not include the word ***abit***. You can add it to your personal dictionary to prevent future alerts.

in · Change preposition

Next, the coders met to discuss their codes and calculate the inter-annotator agreement rate. The initial inter-annotator agreement rates for flagging and correction were 91% and 94%, respectively. Discrepancies for both accuracy dimensions then were resolved in discussion to ultimately reach a consensus. To calculate precision for flagging and correction, the coders removed three items coded *correct/unknown* because the correction was not provided and 14 items coded *neutral/neutral* because the error was stylistic. The precision rate for flagging accuracy was calculated by dividing the number of errors accurately flagged by Grammarly by the total number of errors flagged by Grammarly (Ranalli & Yamashita, 2022). The correction rate was calculated the same way. This was then done for each error type and category.

To measure recall, the coders had three rounds of revision. In the first round of revision, the coders independently reviewed 13 essays to identify all grammar, punctuation, and spelling errors using the error categorization rubric mentioned above and determine what type and category each error belongs to. The coders then had a meeting to compare their codes and discuss any discrepancies. For example, no flaggings were recorded in the category verb tense in the L2 corpus. Instead, Grammarly provided feedback on subject-verb agreement as in the following example:

> *Before planes the distance from Europe to china for example* ***take*** *months by ship* (Grammarly suggested changing “take” to ‘takes’).

To enable a fair comparison between human annotators and computer error detection/correction, the coders decided to follow the same reasoning and provide feedback on subject-verb agreement instead of verb tense. Similarly, because Grammarly marked a noun phrase twice as having the determiner use (a/an/the/this, etc.) error and incorrect noun number error, the coders agreed to do the same:

> *This allows* ***student*** *to make some of the best choices in education about how to produce and receive information* (Grammarly provided two comments. The first comment suggested changing ‘student’ to ‘the student’ or ‘a student.’ The second comment suggested changing ‘student’ to ‘students.’).

In the second round of revision, the coders independently reviewed 15 essays and then had a meeting to compare codes and resolve any disagreement in their codes. In round three, the coders independently reviewed the remaining 25 essays and met to calculate the final inter-annotator agreement rate for flagging which was 93%. To calculate the recall rate, the total number of correctly detected errors by Grammarly was divided by the total number of errors identified by coders (Ranalli & Yamashita, 2022), which was considered the gold standard (Burstein et al., 2004). This was then done for each error type and category.

Results

Table 2 presents the results of the analysis of Grammarly's error-detection/correction performance. The first left column demonstrates three major error types and 17 error categories identified by Grammarly in the 53 argumentative essay drafts. The second left column shows the total number of errors of each error type and category identified by human annotators, i.e., the gold standard. The most frequently identified errors by human annotators were errors in determiner use (a/an/the/this, etc.) (n = 643), wrong or missing prepositions (n = 267), comma misuse within clauses (n = 227), incorrect noun number (n = 202), punctuation in compound/complex sentences (n = 189), faulty subject-verb agreement (n = 169), incorrect verb forms (n = 141), confused words (n = 139), misspelled words (n = 126), and pronoun use (n = 103).

The third left column shows the total number of errors of each error type and category detected by Grammarly, while the fourth left column shows the total number of errors of each error type and category corrected by Grammarly. Among 17 identified and corrected error categories by Grammarly, the most frequently flagged and corrected were errors in determiner use (a/an/the/this, etc.) (n = 412), wrong or missing prepositions (n = 183), comma misuse within clauses (n = 134), faulty subject-verb agreement (n = 116), confused words (n = 111), and incorrect noun number (n = 106). The following two columns on the left illustrate the results of the accuracy evaluation of Grammarly's error detection and correction.

The three columns on the right demonstrate the results of flagging precision, correction precision, and flagging recall. If 17 error categories are considered in the aggregate, Grammarly's precision rates for flagging and correction are 92% and 91%, respectively, and the recall rate is 51%. Among individual error categories, there is a considerable variation. For example, error categories with both high precision and recall rates are modal verbs (100% and 80%, respectively), commonly confused words (85% and 85%, respectively), misuse of quantifiers (83% and 83%, respectively), and confused words (90% and 72%, respectively). Some error categories have high precision but very low recall values, including misuse of semicolons, quotation marks, etc. (100% and 4%, respectively), closing punctuation (100% and 10%, respectively), conjunction use (100% and 11%, respectively), punctuation in compound/complex sentences (94% and 8%, respectively), and incorrect verb forms (84% and 26%, respectively).

Table 2

Grammarly's Error-Detection/Correction Performance Results

	Gold Standard flagging frequency	Grammarly flagging frequency	Grammarly correction frequency	Grammarly flagging accuracy	Grammarly correction accuracy	Flagging Precision	Correction Precision	Flagging Recall
Grammar	**1590**	**951**	**951**	**889**	**877**	**0.93**	**0.92**	**0.56**
Conjunction use	19	2	2	2	2	1.00	1.00	0.11
Determiner use (a/an/the/this, etc.)	643	412	412	383	380	0.93	0.92	0.60
Incorrect noun number	202	106	106	103	101	0.97	0.95	0.51
Faulty subject-verb agreement	169	116	116	107	106	0.92	0.91	0.63
Incorrect verb forms	141	43	43	36	34	0.84	0.79	0.26
Modal verbs	5	4	4	4	4	1.00	1.00	0.80
Misuse of modifiers	35	18	18	17	16	0.94	0.89	0.49
Misuse of quantifiers	6	6	6	5	5	0.83	0.83	0.83
Pronoun use	103	61	61	57	56	0.93	0.92	0.55
Wrong or missing prepositions	267	183	183	175	173	0.96	0.95	0.66
Punctuation	**462**	**153**	**153**	**143**	**143**	**0.93**	**0.93**	**0.31**
Closing punctuation	20	2	2	2	2	1.00	1.00	0.10
Comma misuse within clauses	227	134	134	125	125	0.93	0.93	0.55
Misuse of semicolons, quotation marks, etc.	26	1	1	1	1	1.00	1.00	0.04
Punctuation in compound/complex sentences	189	16	16	15	15	0.94	0.94	0.08
Spelling	**278**	**176**	**176**	**149**	**147**	**0.85**	**0.84**	**0.54**
Commonly confused words	13	13	13	11	11	0.85	0.85	0.85
Confused words	139	111	111	100	99	0.90	0.89	0.72
Misspelled words	126	52	52	38	37	0.73	0.71	0.30
Total	**2330**	**1280**	**1280**	**1181**	**1167**	**0.92**	**0.91**	**0.51**

Note. Precision = the total number of errors accurately flagged/corrected by Grammarly by the total number of errors flagged/corrected by Grammarly (e.g., flagging precision of misspelled words: 38÷52 = 0.73); Recall = the total number of correctly detected errors by Grammarly divided by the total number of errors identified by human annotators (e.g., recall of misspelled words: 38÷126 = 0.30).

Discussion

If we adopt the threshold of 80% precision set by the Criterion's developers (Quinlan et al., 2009), Grammarly's overall precision rates for flagging and correction (92% and 91%, respectively) exceed it. This means that Grammarly is highly accurate when flagging errors and providing corrections. These results correspond with Grammarly's high flagging and correction precision rates (88% and 81%, respectively) reported in Ranalli and Yamashita (2022) but partially contradict the results reported in Guo et al. (2021) because in their study, Grammarly's flagging precision rate of 69% was much lower the threshold of 80% precision while the correction precision rate of 82% exceeded it. These results are also in line with Criterion's high flagging precision rate documented in previous literature (Chodorow et al., 2007; Han et al., 2006; Tetreault & Chodorow, 2008), which indicates that Grammarly can be as good in detecting errors as a well-known and widely-used AWE system, Criterion. As expected, Grammarly's recall rate is low (51%). This finding suggests that Grammarly is not able to identify as many L2 errors as human annotators do, which could be a concern when integrating Grammarly into the L2 writing classroom.

It was not always clear why Grammarly overlooked some errors. For example, Grammarly provided feedback on punctuation in compound/complex sentences saying that a comma should be removed before the dependent clause marker "before" in the sentence: "It's true that some people consider imitation a kind of good **thing,** because that's how we learn new things." However, Grammarly did not flag a similar error in the following sentence: "In my opinion, young people should spend more time outside with family and **friends,** because the internet and social media can cause negative health consequences and destroy their communication such as their friendships." Likewise, Grammarly provided feedback on incorrect verb forms saying that the bare infinitive form should be used after "make" in the sentence "Second, the printed version makes us **to** note our thoughts depending on the mood that we feel at that moment." However, it did not flag a similar error in the following sentence: "A.I. is a way to make goals **to** be reached faster." The low recall rates were observed not only in punctuation in compound/complex sentences and incorrect verb-forms, but also in misuse of semicolons, quotation marks, etc., conjunction use, and misspelled words. Just like with errors on punctuation in compound/complex sentences and incorrect verb-forms, Grammarly was not always able to, for example, identify misspelled words. Again, it was not always clear why. For instance, Grammarly identified a misspelled word "menthal" in the sentence "Women who are denied abortions are in danger of **menthal** health issues" and suggested replacing it with "mental." However, Grammarly did not recognize a misspelled word "wish" in the sentence "These things are very dangerous for all people and many people today **whish** if that drone did not exist."

The possible explanation for the low recall rate could be the fact that Grammarly is not programmed to specifically address L2 errors. The L2 writers' sentences in the corpus often had structural issues. They were also quite long, and some lacked punctuation. Therefore, Grammarly might have difficulty in identifying errors in those sentences. Previous research on various AWE systems also found their poor performance in finding L2 errors due to the fact that they are not specifically designed with ELLs in mind yet being marketed to schools and colleges where students are ELLs (e.g., Dikli & Bleyle, 2014; Liu & Kunnan, 2016; Ranalli & Yamashita, 2022).

Conclusion

The study investigated Grammarly's error-detection/correction performance by measuring its precision and recall and using human annotators as a benchmark. By doing so, the study extended the existing system-centric research that 1) has primarily focused on the AWE system, Criterion, 2) has evaluated AWE's performance on one or two error categories, and 3) has mainly measured AWE's precision rate rather than recall. According to the results of the study, Grammarly was highly accurate in flagging and correcting errors (i.e., high flagging and correction precision rates) in 53 argumentative essay drafts written by ESL undergraduate students. However, it skipped half of the L2 errors found by human annotators in the same essay drafts (i.e., low recall rate).

The study findings indicate that Grammarly has the potential to be used in L2 writing classrooms due to its satisfactory error detection and correction performance. However, the fact that Grammarly misses half of the L2 errors can be problematic as students may interpret missed errors as the only errors in their writing which could be detrimental (Ranalli et al., 2017). Therefore, teachers should warn their students about Grammarly's low recall so that they do not treat the flagged errors as the only errors in their writing and think that everything else is okay. This also means that teachers should compensate for Grammarly's limitation in detecting all L2 errors by providing their own feedback on lower-order concerns. Additionally, despite Grammarly's high flagging and precision rates, teachers should caution their students that some of Grammarly's flaggings and corrections may not be accurate. Cotos (2018) noted that inaccuracies of AWE feedback may have both a positive impact, as it may trigger the student's noticing of the error, and a negative impact, as it may mislead the student into making inaccurate changes to the text. To avoid the negative impact, teachers should train their students to critically evaluate Grammarly's feedback so that students do not accept automated feedback blindly (Koltovskaia, 2020).

The study also provides implications for Grammarly developers. Since Grammarly is growing in its use in ESL and EFL contexts (Guo et al., 2021; Ranalli, 2018), it is time for the developers to adapt the tool to meet the needs of ELLs. For example, although there were many verb tense errors in the corpus used in this study, Grammarly failed to detect any verb tense errors. This means that the developers should familiarize themselves with the research that focuses on the influence of L1 on learner errors and use language learner corpora to train statistical classifiers to detect common L2 errors. Additionally, the developers should also increase the recall rate as low recall rate could be detrimental for students as they may interpret flagged errors as the only errors in their writing.

The study is not without its limitations. First, the study was based on 53 essays and on one essay genre (i.e. argumentative essay). Future studies may consider having more essays and on different genres to assess Grammarly’s error-detection/correction performance to be able to generalize the results. Second, the study had only two human annotators. Future studies may want to consider having more annotators for manual evaluation of the tool’s performance to increase reliability. Finally, Grammarly’s scoring ability was not the focus of the study which merits future investigations.

CHAPTER III

TEACHERS' USE AND PERCEPTIONS OF GRAMMARLY

Introduction

In recent years, automated writing evaluation (AWE) systems have grown in popularity as a source of feedback that can complement teachers' response to second language (L2) writing. The complementary nature of automated feedback is representative of a system's adept ability to provide feedback on lower-order concerns (LOCs), including grammar and mechanics (Ranalli, Link, & Chukharev-Hudilainen, 2017). It has been suggested that because AWE can take care of LOCs, it has the potential to free up teachers' time to focus more on higher-order concerns (HOCs), such as content and organization (Chen & Cheng, 2008; Li, Link, & Hegelheimer, 2015; Warschauer & Grimes, 2008; Wilson & Czik, 2016). However, little empirical evidence exists to support this claim. The small number of studies that investigated the impact of AWE on teacher feedback (Jiang, Yu, & Wang, 2020; Link, Mehrzad, & Rahimi, 2020; Wilson & Czik, 2016) reveal conflicting results, thus warranting more research in this regard. Additionally, little effort has been made by these studies to explore teachers' perceptions of AWE when they use it to complement their feedback. Teachers, as "direct facilitators of an AWE system in classrooms" (Li, 2021, p. 2), may hold different views about AWE and thus may develop different pedagogical strategies that could compensate for AWE's limitations (Cotos, 2018; Li, 2021). Research also shows that students are likely to adopt the same attitude toward AWE their teacher holds (Chen & Cheng, 2008), thus examining teachers' perceptions is essential. Finally, teachers' perceptions of AWE are "an important source of evidence - evidence of social validity[1]" (Wilson et al., 2021, p. 2). Given the growing interest in AWE, it is likely "to become more pervasive in the field with enormous educational and social impact" (Jiang et al., 2020, p. 2). Therefore, it has become highly important to examine the use of various AWE systems by all stakeholders, including teachers who play an integral role in making decisions about their implementation in L2 writing classrooms (Li, 2021).

The current study first examines the nature of pre-and in-service, postsecondary L2 writing teachers' feedback when they use Grammarly, which is making important inroads in L2 writing classrooms (Ranalli, 2018), as a complement. The study then explores the teachers' perceptions of the tool. The findings of the study provide a better understanding of how to use Grammarly and similar systems to complement teacher feedback for productive student learning outcomes.

Literature review

Impact of AWE on Teacher Feedback

AWE is a software program that provides instant automated scoring and individualized automated feedback for essay improvement (Cotos, 2018). Initially, AWE

[1] According to Leko (2014), social validity is based on the idea that "consumers of an intervention and other stakeholders apart from researchers should participate in the evaluation process" (p. 275). For more information on social validity, read Wolf (1978).

systems were developed for assessing large-scale tests, such as the TOEFL and the GRE with the purpose of reducing the heavy load of grading a large number of student essays and saving time (Chen & Cheng, 2008). These systems were called automated essay scoring (AES) because of their automated scoring engine (Warschauer & Grimes, 2008). Later, AES systems were augmented to include automated formative feedback (Cotos, 2018), and the systems, which received the name AWE, were marketed to schools and colleges. When AWE was introduced for instructional use, there was a concern that it may replace a teacher as a primary feedback agent (Ericsson & Haswell, 2006). Researchers, however, assure that the intended use of AWE is to complement teacher feedback instead of replacing it (Chen & Cheng, 2008; Link, et al., 2020; Stevenson, 2016; Ware, 2011; Wilson & Czik, 2016). As such, AWE has the ability to liberate teachers' time to focus more on HOCs (Grimes & Warschauer, 2010; Li et al, 2015; Link et al, 2014; Ranalli, 2018) because AWE's automated feedback is more computationally adept at providing feedback on LOCs (Grimes & Warschauer, 2010; Ranalli et al., 2017). However, evidence to support this claim is scarce. To date, only three studies have explicitly investigated the impact of AWE on teacher feedback.

Wilson and Czik (2016) conducted a quasi-experimental study in which they assigned two eighth grade English Language Arts (ELA) classes to the Project Essay Grade (PEG) Writing + teacher feedback condition and two classes to the teacher-only-feedback condition. They then asked the U.S. middle-school teachers in each condition to provide feedback as they normally would and analyzed their feedback to examine the impact of PEG Writing on the type (direct, indirect, praise), amount, and level (HOCs vs. LOCs) of teacher feedback. The researchers found that teacher feedback did not change in the type and amount across two conditions. As for feedback level, the researchers found that the teachers in the PEG Writing + teacher feedback condition still gave a substantial amount of feedback on LOCs, despite using PEG Writing to complement their feedback. However, they gave proportionally more feedback on HOCs than LOCs compared to the teachers in the teacher-only-feedback condition. Due to the small effect sizes for differences in feedback proportions across two conditions, the researchers claimed that they provide only partial support for the premise that AWE allows teachers to focus more on HOCs.

Link et al. (2020) extended Wilson and Czik's study by focusing on English as a foreign language (EFL) teachers from Iran and investigating the impact of the Educational Testing Service's (ETS) Criterion on teacher feedback. The researchers assigned two classes to either the AWE + teacher feedback condition or the teacher-only-feedback condition. The results revealed that unlike the teacher in the teacher-only-feedback condition, the teacher in the AWE + teacher feedback condition provided less feedback but the use of Criterion did not result in a higher frequency of feedback on HOCs, which contradicts the results in Wilson and Czik. However, the results of Link et al.'s study should be interpreted with caution because of the study's methodological constraints. While the teacher in the AWE + teacher feedback condition provided feedback only on HOCs since Criterion took care of LOCs, the teacher in the teacher-only-feedback group gave feedback on both HOCs and LOCs, making a comparison of feedback across two conditions problematic.

Different from the two comparative studies, Jiang et al. (2020) conducted longitudinal, classroom-based qualitative research in which they explored the impact of automated feedback generated by Pigai on Chinese EFL teachers' feedback practice. The researchers found that of the eleven participating teachers, two resisted using Pigai because they had low trust in its feedback; therefore, their traditional feedback practices remained unchanged. Three of the teachers used Pigai as a surrogate, which resulted in the reduction of feedback time and amount as they offloaded the majority of their feedback to Pigai. The remaining six teachers used Pigai as a complement to their feedback, which allowed them to provide more feedback on HOCs, corroborating the claim that AWE affords teachers to focus their feedback more on HOCs. Similar to Wilson and Czik, the researchers noted that there is no division of labor, such as that a teacher takes care of HOCs and AWE takes care of LOCs, as the teachers in their study still provided a considerable amount of feedback on LOCs, despite using Pigai to augment their feedback. Therefore, the researchers suggested that there is a need to refute a dichotomy that leaves feedback on global aspects of writing to teachers and local aspects of writing to AWE systems.

Teachers' Perceptions of AWE

Not only is there limited research on the impact of AWE on teacher feedback, but the extant literature has also put little effort into understanding teachers' perceptions of AWE when they use it to complement their feedback. Teachers' perceptions of AWE can reveal factors influencing changes, if any, in teacher feedback when AWE is used as a complement. However, such information is minimal. A modest number of studies that have focused on teachers' perceptions of AWE (Chen & Cheng, 2008; Grimes & Warschauer, 2010; Li, 2021; Link, Dursan, Karakaya, & Hegelheimer, 2014; Maeng, 2010; Warschauer & Grimes, 2008; Wilson et al., 2021) provide insight into the areas of teachers' satisfaction and dissatisfaction with AWE.

For instance, Link et al. (2014) interviewed five ESL university writing teachers to learn about their perceptions of ETS's Criterion. The results revealed that the teachers found Criterion effective for fostering students' metalinguistic ability, reducing their workload, and providing feedback on grammar. The teachers also were highly satisfied with Criterion's ability to promote students' autonomy and motivation. As for the areas of dissatisfaction, the teachers reported that Criterion does not always provide necessary and high-quality feedback, and its holistic scores are not always reliable though useful. Li (2021) also examined ESL university writing teachers' perceptions of ETS's Criterion. The findings of his study showed that while all three teachers were overall satisfied with Criterion as they found it helpful, they noted that its automated feedback was too broad and occasionally confusing to their students. The teachers also reported that Criterion missed a lot of errors committed by their ESL students. In their recent study, Wilson et al. (2021) explored 17 ELA elementary teachers' perceptions of the AWE system, MI Write, for supporting writing instruction in grades 3-5. They found that the teachers were satisfied with the immediacy of MI Write feedback. They also liked that MI Write helped them determine students' weaknesses and strengths in writing and helped students understand that writing is a process and revising is important. The areas of teachers' reported dissatisfaction were that MI Write was misaligned to their instruction. The teachers also voiced their concerns about the accuracy of MI Write's holistic scores and that the scores would result in students valuing quantity over quality.

Overall, studies report that AWE is often perceived as an "extra voice" and "extra helper" (Li, 2021, p. 5), "second pair of eyes" (Grimes & Warschauer, 2010, p. 21;

Warschauer & Grimes, 2008, p. 28), and “good partner with the classroom teacher” (Wilson et al., 2021, p. 5). This indicates that teachers tend to find AWE useful and hold positive views about AWE, despite being aware of its limitations, particularly in regard to the accuracy of its automated scoring and the quality of its feedback (Grimes & Warschauer, 2010; Li, 2021; Link et al., 2014; Maeng, 2010; Warschauer & Grimes, 2008; Wilson et al., 2021).

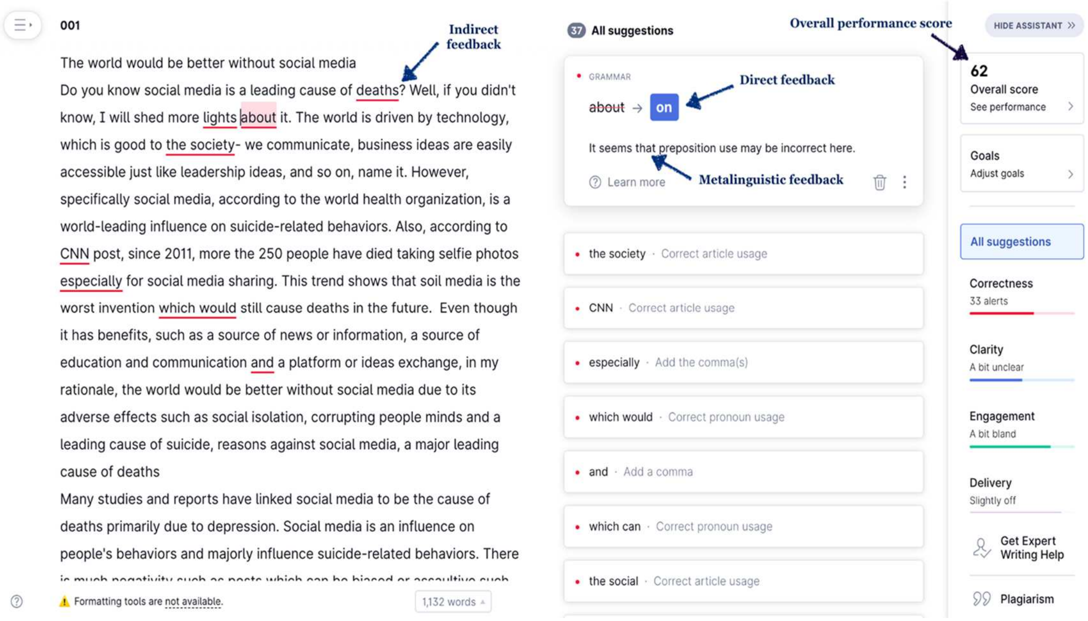
Grammarly

While the bulk of the aforementioned research has focused on commercially available AWE systems, such as Criterion (Li, 2021; Link et al., 2014; Link et al., 2020; Maeng, 2010; Warschauer & Grimes, 2008) and My Access! (Chen & Cheng, 2008; Grimes & Warschauer, 2010; Warschauer & Grimes, 2008), scant literature exists on Grammarly, despite it is being the world’s leading automated proofreader and being increasingly used in higher education and K-12 institutions (Grammarly, 2022). In fact, more than 3000 educational institutions, including Arizona State University, University of Phoenix, and California State University have licensed Grammarly to improve student writing outcomes (Grammarly, 2022). The small number of studies that have focused on Grammarly’s L2 writing pedagogical potentials show that Grammarly helps improve students’ writing considerably and that students perceive it positively (Barrot, 2021; Guo, Feng, & Hua, 2021; Koltovskaia, 2020; Thi & Nikolov, 2021). The studies that have examined Grammarly’s performance reveal that Grammarly is quite accurate in detecting and correcting common L2 linguistic errors (Koltovskaia, 2022; Ranalli & Yamashita, 2022), and it provides feedback on more error categories than other tools, such as Microsoft Word (Ranalli & Yamashita, 2022).

As for Grammarly’s affordances, Grammarly can be used for free, but users can also get Grammarly Premium and Grammarly Business, which have a monthly subscription of $12 and $12.50 per month, respectively (Grammarly, 2022). Grammarly can be accessed in multiple ways, such as through a web app, browser extension, productivity software plug-in, and mobile device. Once a paper is uploaded to Grammarly’s website, it provides indirect feedback (i.e., it indicates that an error has been made by underlining the error), metalinguistic explanation (i.e., it gives a brief grammatical description about the nature of the error), and direct feedback (i.e., it gives a correct form or structure) (Figure 5). It is noteworthy that the free version of Grammarly, which was used for this study, provides feedback on five error types, including grammar, punctuation, spelling, conventions, and conciseness. Apart from feedback on errors, Grammarly also provides an overall performance score from 1 to 100 that represents the quality of writing. The score is based on different types of suggestions given to the paper and on how the paper compares to other papers with similar goals. The more suggestions the paper gets, the lower the score is. The goals setting function can be used to get tailored Grammarly suggestions based on audience, formality, domain, tone, and intent. Finally, Grammarly generates a full performance report that contains such information as general metrics, performance score, the original text, and feedback on errors, and it can be downloaded in PDF format.

Figure 5

Grammarly's Interface



001

The world would be better without social media

Do you know social media is a leading cause of deaths? Well, if you didn't know, I will shed more lights about it. The world is driven by technology, which is good to the society- we communicate, business ideas are easily accessible just like leadership ideas, and so on, name it. However, specifically social media, according to the world health organization, is a world-leading influence on suicide-related behaviors. Also, according to CNN post, since 2011, more the 250 people have died taking selfie photos especially for social media sharing. This trend shows that soil media is the worst invention which would still cause deaths in the future. Even though it has benefits, such as a source of news or information, a source of education and communication and a platform or ideas exchange, in my rationale, the world would be better without social media due to its adverse effects such as social isolation, corrupting people minds and a leading cause of suicide, reasons against social media, a major leading cause of deaths

Many studies and reports have linked social media to be the cause of deaths primarily due to depression. Social media is an influence on people's behaviors and majorly influence suicide-related behaviors. There is much negativity such as posts which can be biased or assaultive such

Formatting tools are not available.

1,132 words

37 All suggestions

GRAMMAR

~~about~~ → on

It seems that preposition use may be incorrect here.

Learn more

- the society · Correct article usage
- CNN · Correct article usage
- especially · Add the comma(s)
- which would · Correct pronoun usage
- and · Add a comma
- which can · Correct pronoun usage
- the social · Correct article usage

62
Overall score
See performance

Goals
Adjust goals

Correctness
33 alerts

Clarity
A bit unclear

Engagement
A bit bland

Delivery
Slightly off

Research Aim and Questions

The study fills in the following gaps in the extant literature that warrant research to learn how to meaningfully implement Grammarly as complement to teacher feedback in L2 writing classrooms. First, a limited number of qualitative studies have been conducted on the impact of AWE on teacher feedback. Such research is necessary as it provides a more in-depth and contextualized understanding of the phenomenon under inquiry (Yin, 2009). Second, insufficient attention has been paid to teachers' perceptions of AWE when they use it to complement their feedback. Examining teachers' perceptions when they use AWE to augment their feedback can give insight into how they feel about using such tools to provide feedback as well as factors that may influence changes, if any, in their feedback practice. Finally, scant research is available on Grammarly. Since Grammarly is gaining popularity among English language learners, more research is needed on this tool to provide useful recommendations for its implementation in L2 writing classrooms.

Therefore, this study first examines the nature of pre- and in-service, postsecondary L2 writing teachers' feedback when they use Grammarly as complement. The study then explores the teachers' perceptions of Grammarly after using it to complement their feedback. The study was guided by the following research questions:

RQ1. What is the nature of L2 writing teachers' feedback when they use Grammarly as a complement?

RQ2. What factors (if any) influence teacher feedback practices when using Grammarly?

RQ3. What are L2 writing teachers' overall perceptions of Grammarly as a complement to their feedback?

Methods

Context

The study was situated in an L2 writing program at a large south-central U.S. university. The L2 writing program offers two undergraduate L2 writing courses and two graduate writing courses to students whose native language is not English. The study took place in an undergraduate L2 writing course that focuses on expository writing with an emphasis on structure and development from a usage-based perspective, with special attention paid to sentence- and discourse-level of English as a second language. For this course, the students are typically required to write a diagnostic essay, which is written at the beginning of the semester and is used to diagnose students' linguistic and writing abilities. Throughout the semester, the students are assigned to write three texts: an argumentative essay, a compare and contrast essay, and a process essay. At the end of the semester, the students write a final exam essay. For this study, the focus was on students' argumentative essays; the first major assignment in this course.

Participants

Three in-service and three pre-service teachers working in the L2 writing program consented to participate in the study (Appendix D). In-service teachers are graduate teaching associates (GTAs) who have completed a one-year training in the program and teach one or two sections of undergraduate or graduate L2 writing courses. Pre-service teachers are first-year GTAs who observe courses in the program (8 hours) and work at the Writing Center (12 hours). The six participants were Mik, Mei, Maria, Rob, Jackson, and Heaven (pseudonyms). At the time of the study, the participants were Ph.D. and M.A. students in Applied Linguistics. Their background information can be seen in Table 3.

Table 3

Participants' Background Information

	Gender	**Degree**	**Mother tongue**	**Country of origin**	**Years of Teaching Experience**	**Teaching Status**
Mik	Female	Ph.D.	Italian	Italy	8	In-service
Mei	Female	Ph.D.	Cantonese	China	3	In-service
Maria	Female	Ph.D.	English	USA	9	In-service
Rob	Male	Ph.D.	Bengali	Bangladesh	6	Pre-service
Jackson	Male	Ph.D.	English	USA	4	Pre-service
Heaven	Non-binary	M.A.	English	USA	1	Pre-service

Procedures

Since data collection of the study was scheduled in the middle of the spring 2020 semester and during the pandemic, integrating Grammarly into L2 writing classrooms was not practicable as such intervention could interrupt the class. Therefore, the participants were given a hypothetical scenario (Appendix E), which asked them to provide formative feedback on students' rough drafts and use Grammarly reports to complement their feedback. The participants were given ten randomly selected rough drafts of the argumentative essay written during the fall 2018 semester. The texts were extracted from the Wrangler corpus, an electronic collection of texts written by English as second language (ESL) learners at the participating university. The L2 writing class size often tends to be small (a maximum of 18 students); therefore, ten essays are considered average. It is noteworthy that the format of the argumentative essay of the fall 2018 semester was similar to the format of the Spring 2020 semester. The only difference was the topic. The topic of the argumentative essay used for this study was on technology. The students were supposed to write an article for an imaginary Discover Magazine arguing about one invention the world would be better without. For more details see the assignment prompt in (Appendix A) and the assignment rubric in (Appendix F). It is also important to note that none of the participants had prior experience using Grammarly.

To ensure the participants had a similar experience, they all received the same drafts along with a Grammarly report for each essay that was downloaded from the Grammarly website. All files were electronic. Although the participants were given reports that contained Grammarly feedback for the purposes of ease, they were also asked to independently run the essays through Grammarly to understand how Grammarly functions. It has to be mentioned that in-service teachers had prior experience teaching the L2 writing course, while pre-service teachers had observed or were observing the course at the time of the study and participated in office hours and one-on-one conferences. Thus, all the participants were familiar with the assignment and knew how to evaluate students' essays. After providing feedback, the participants were scheduled for an individual semi-structured interview with the author that lasted 40 minutes on average with each participant. The interview was conducted via Zoom and recorded. In the interview, the participants were first asked to provide their demographic information. They then were asked about their prior experience with AWE. Finally, they were asked about their perceptions of Grammarly after using it to complement their feedback (for interview questions, see Appendix G).

Data Analysis

The participants' feedback given to ten essay drafts and recordings from the semi-structured interview were used for data analysis.

The participants' feedback was analyzed to answer the first research question, which is about the nature of L2 writing teachers' feedback when they use Grammarly as a complement. The author generated the error categories rubric based upon previous literature (Ene & Upton, 2014; Ferris, 2006). In the rubric, teacher feedback was divided into two feedback levels: higher-order(level) concerns (herein HOCs) and lower-order(level) concerns (herein LOCs). HOCs were operationalized as feedback that focuses on the discourse level, including content and organization/coherence/cohesion. LOCs were operationalized as feedback that focuses on the form level, including vocabulary, grammar/syntax/morphology, and mechanics. The author and a professor at a U.S. university, who has a Ph.D. in Applied Linguistics, independently coded the participants' feedback using the error categories rubric. The coders then had a meeting, that lasted three hours, in which they discussed the rubric and compared their initial codes. In the meeting, the coders decided to modify the rubric by including codes that emerged from the data itself. In addition to HOCs and LOCs, the coders included such codes as *general feedback, positive feedback,* and *Grammarly feedback evaluation*. The *general feedback* code was used when a teacher provided a comment on the overall quality of an essay by focusing on both HOCs and LOCs. For example, "This essay draft has many run-on sentences. Ideas in the counterargument paragraph were well-presented. A title is needed for this essay" (Mei). The *positive feedback* code was used when a teacher praised a student for achievement or encouraged them about performance. For example, "This is a strong opening sentence, and one that captures reader attention. Good job!" (Heaven). The *Grammarly feedback evaluation* code was used when a teacher made some notes on Grammarly's feedback in her comment. For example, "One of the things to keep in mind with Grammarly – and spellcheck – is that sometimes, it won't recognize something as misspelled because it looks like another word. Therefore, it's worthwhile to reread your essay, even if spellcheck says that nothing is wrong, in case you accidentally mistyped something" (Heaven). The coders also added such codes as *documentation and attribution* and *formatting and style* to the mechanics under LOCs. The coders again independently coded the participants' feedback using a modified rubric (Appendix H). They then had another meeting, that lasted two hours, in which they compared their codes and calculated the inter-rater agreement rate which was 93% across all identified error categories. Any discrepancies were discussed in the meeting until a consensus was reached. Descriptive statistics were then calculated in Excel for interpretation of the data.

The interview transcripts were analyzed to answer the second research question, which looks at the factors that might have influenced teacher feedback practices when using Grammarly and the third research question, which examines L2 writing teachers' overall perceptions of Grammarly as a complement to their feedback. The interview audio recordings were extracted from Zoom and transcribed in Trint (https://trint.com). The author and the professor checked the transcripts for accuracy against the original recordings. The transcripts then were organized in a Google spreadsheet by the individual participant. Guided by the research questions, inductive coding that allows for themes to emerge from the data was used (Creswell, 2014). Specifically, the analysis relied on open, axial, and selective coding (Corbin & Strauss, 2008). In the open coding phase, the two coders coded six transcripts independently using the language closely related to the data (Table 4). The coders then had a Zoom meeting to compare their initial codes and refined them if necessary. In the axial coding phase, the coders combined codes that are similar into categories and compared those codes and categories across six cases. For example, low recall or the fact that Grammarly skipped a lot of L2 errors was mentioned by all six participants. Therefore, this code was placed under the category "low recall." In the selective coding phase, the categories were placed under larger themes. For instance, the "low recall" category was placed under the theme "Teachers' perceptions of Grammarly feedback." The coders had a meeting again to refine categories and themes and choose illustrative quotes that represent the themes' essence.

Table 4

Example of Mik's data layout and coding sheet

Codes	Mik's comments
Skips errors (i.e. low recall)	But I did feel like it **skipped a lot of grammar mistakes** that I found in the essay that Grammarly didn't find or the algorithm wasn't able to detect that that was a grammar mistake.

* Initial codes assigned to the participant's comments during the open coding phase.

Findings

RQ1: What is the nature of L2 writing teachers' feedback when they use Grammarly as a complement?

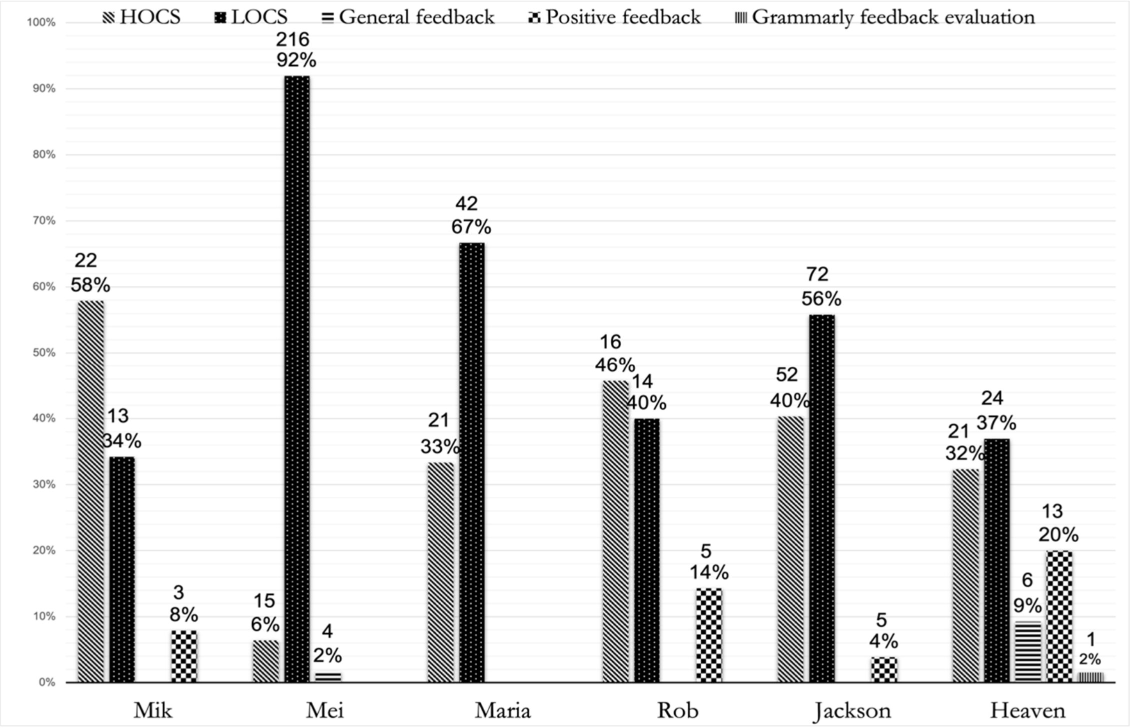
To answer the first research question, the participants' feedback for ten essays was analyzed. According to Figure 6, all six participants provided feedback on both HOCs and LOCs, despite having Grammarly feedback on LOCs as a complement. Additionally, the participants also felt the need to provide positive feedback, general feedback, and feedback that comments on Grammarly's performance. Of the six participants, two participants, Mik and Rob gave more feedback on HOCs while four participants, Mei, Maria, Jackson, and Heaven provided more feedback on LOCs. While Mik devoted 58% of her feedback to HOCs and 34% to LOCs, Rob allocated 46% of his feedback to HOCs and 40% to LOCs. They both also provided positive feedback; Mik devoted 8% while Rob devoted 14% to positive feedback.

Mei devoted 92% of her feedback to LOCs, which is the highest number among all the participants. She allocated only 6% to HOCs and 2% to providing general feedback. Similarly, Maria and Jackson devoted the majority of their feedback to LOCs. While Maria allocated 67% to LOCs and 33% to HOCs, Jackson devoted 56% to LOCs and 40% to HOCs. Jackson also gave 4% of positive feedback. Compared to all the participants, Heaven's feedback was the most diverse in terms of feedback type. They allocated 37% to LOCs, 32% to HOCs, 20% to positive feedback, 9% to general feedback, and 2% were comments about Grammarly feedback.

A closer look at the participants' feedback on LOCs shows that they devoted most of their feedback to sentence structure, word choice (including collocations and phrasing), spelling, punctuation, word form, overall quality of grammar, and documentation or attribution. For example, Mik allocated 18.4% of her LOC feedback to documentation and attribution. Mei devoted 14. 9% of her LOC feedback to word choice, 11.1% to punctuation, 10.2% to sentence structure, and 7.7% to spelling. Maria allocated 27% of her LOC feedback to documentation and attribution, 12.7% to word choice, 11.1% to sentence structure, and 6.3% to punctuation. Rob devoted 14.3% of his LOC feedback to overall quality of grammar, 11.4% to word choice, and 5.7% to sentence structure. Jackson allocated 20.2% of his LOC feedback to documentation and attribution, 13.2% to spelling, and 8.5% to word choice. Finally, Heaven devoted 10.2% of their LOC feedback to sentence structure, 8.5% to spelling, and 5.1% to word form (see Appendix I for a detailed breakdown of teacher feedback).

Figure 6

L2 Writing Teachers' Feedback for Ten Essays



RQ2: What factors (if any) influence teacher feedback practices when using Grammarly?

To answer the second research question, the participants' interview data were analyzed. The analysis revealed three factors that might have influenced the participants' feedback when they used Grammarly as a complement. These factors are the participants' use of Grammarly reports, their perceptions of Grammarly feedback, as well as their feedback practice and personal beliefs about feedback and L2 writing course.

Teachers' Use of Grammarly Reports

The way the participants used Grammarly reports might have impacted their feedback. The interview data revealed that Mik, Maria, Rob, and Heaven consulted Grammarly reports before providing their own feedback. They reported that they first glanced at students' essays, as Heaven said, to "get an idea of the content, argument, structure, and what problems might exist." The participants then consulted Grammarly reports to check what errors Grammarly caught and what errors were left for them to address. In this regard, Maria said:

> I went ahead and scanned to see what was happening there and just look at the mistakes that were already commented on by the software so that I would not be wasting my time repeating the same thing.

After consulting Grammarly reports, the participants provided their own feedback. Unlike the four participants, Jackson consulted Grammarly reports while providing his feedback. In the interview, he said that he put the paper on one half of his screen and the Grammarly report on the other half, which helped him make decisions on what errors to focus. The following is his comment in this regard: "I had it open. I would glance at it. I didn't really go off of it too much. I would just go OK! It addressed a lot of things here. I could focus on content. If it didn't look at anything, I needed to also address a bit of form." Mei, in contrast, looked at Grammarly reports after providing her own feedback. In the interview, she said she did not "want to be distracted or be biased by Grammarly feedback." Therefore, she first provided her own feedback and then looked at Grammarly reports to see if Grammarly was able to catch the same errors she did. She then added, "many of the comments match mine, especially the local language errors." Mei was the only teacher who did not really use Grammarly to complement her feedback.

Teachers' Perceptions of Grammarly Feedback

When using Grammarly as a complement, the participants noted some disadvantages and advantages of automated feedback that affected their feedback. In terms of disadvantages, all participants noticed that Grammarly skips a lot of errors. For example, Mik said, “it's not extensive, and it skips a lot of grammar mistakes. To compensate for this limitation in detecting all L2 errors, the participants provided feedback on LOCs.

Some participants also noted that Grammarly caught only “basic” errors. For instance, Maria said, “ I did notice the software didn't get sentence structure errors. So I addressed those. I didn't do anything really simple like subject-verb agreement or number. You know, those are really basic. Grammarly took care of those.” Since Grammarly took care of the “basic” errors, the participants’ feedback on LOCs predominantly focused on sentence structure, word form, word choice, and documentation and style.

All participants also noticed that Grammarly feedback was occasionally inaccurate. In this regard, Heaven said:

> Sometimes Grammarly thinks that this word is the problem, but it's actually this other word. It's just confused, and Grammarly sometimes gets confused because you make a typo, but the typo looks like a word. So it doesn't really know what to do with that. So it takes a human looking at it and evaluating those things that Grammarly is saying is an error.

Because Grammarly feedback was inaccurate at times, Heaven felt the need to warn students about this in her comments. Some participants also reported that Grammarly feedback was negative and can be discouraging and overwhelming as it catches every little error. To this end, Jackson said, “One student got sixty-four comments, which is incredibly discouraging. [...]. I know if I got a paper back like that I would be discouraged." Since Grammarly feedback was negative, the participants provided positive feedback.

As for advantages, the participants liked that Grammarly does part of their job by taking care of errors on LOCs which frees their time to focus more on other issues in students' papers. In this regard, Rob said, "It really reduces time and effort, and it can let me focus on higher-order issues." Similarly, Maria said "I just focused on higher-order futures like writing quality and structure, and citations. So it made it easier for me. So I didn't have to spend time making comments on grammar unless totally necessary."

The participants also were satisfied with how detailed Grammarly feedback was. They liked that Grammarly underlines errors and provides metalinguistic explanation which is very similar to what they do. Mik, for example, said:

> It underlined where the mistake was and it kind of gave you a keyword for it. So you start learning some of the vocabularies like determiner or verb tense that a lot of L2 writing students might not even know to talk about grammar. So I thought that was really nice.

Finally, the participants reported that Grammarly feedback helped them see the most frequent errors of individual students and the class as a whole and what errors need to be addressed in the paper and in class. Regarding this, Heaven said, "I liked that it freed me up to just kind of focus on what I saw as broader trends, and I liked that looking at it made it easier for me to see what everybody in the class is having difficulty with."

Teachers' Feedback Practice and Beliefs

Another factor that might have influenced the participants' feedback when they used Grammarly as a complement is their feedback practice and beliefs about feedback and L2 writing course. All the participants noted that students enrolled in the L2 writing course need both types of feedback because not only their writing but also their linguistic skills are developing. In this regard, Mei said: “they are freshmen and their language skills are developing and also their critical thinking skills are developing. That's why I try to give both types of feedback.”

However, the participants reported that they tend to prioritize feedback on HOCs over LOCs because students often struggle with idea development and structure which is more important than grammar. For instance, Maria said, “I mostly try to give comments on essay structure, paragraph structure, thesis statement, topic and conclusion sentences, and citations, especially on citation formatting.” Surprisingly, the quantitative findings contradict the above statements because the majority of the participants provided more feedback on LOCs than HOCs according to Figure 6.

When asked if their feedback practice changed when using Grammarly to complement their feedback, all participants said that their feedback practice did not substantially change. What changed a bit is that Grammarly took care of the errors that teachers do not consider that much of a problem as there were more pressing issues in students’ writing that needed to be addressed such as sentence structure. For example, Mik said, “whether or not I had the Grammarly report didn't really change the approach that I have. The only thing that changed maybe if I saw something, a specific grammar error that was repeated over and over, I might have highlighted it, but I didn't because Grammarly took care of that.” Similarly, Maria noted: “it’s impossible to give students an explanation for every little mistake. So, the software is really helpful in that respect.”

As for beliefs about the L2 writing course, the participants expressed concerns about the tool because it may not align with the course's main goal which is to teach students to find solutions for the identified errors on their own while Grammarly not only indicates where the error is but also provides a correction, which students can automatically accept. This consequently may result in no learning as students may accept feedback blindly. For instance, the following is what Mik said in this regard, “I have concerns that the students might feel that those are the only mistakes in their writing, that the automatization of how to correct those mistakes might take away from the awareness the students has for future writing.” Mik then added the below:

> Yes, you want to make sure that the essay they turn in is grammatically correct but really what you're trying to do is to teach them how to understand the grammar and how to eventually catch their own mistakes and not make them anymore. So, I think they're slightly different objectives and it's hard to make sure that you're doing that with Grammarly because that's really not the point, or at least not the long-term point of teaching L2 writing.

Heaven, Mik, Jackson, and Maria expressed their preference for other sources of feedback such as peer-review and writing center consultations, which they believe are more effective. In this regard, Heaven said, “there are benefits to peer review that Grammarly is just not going to be able to capture because it's not another student, it's not another person who's saying, here's the mistake and here is how you can fix it.” All the participants also emphasized the importance of a human-to-human interaction when it comes to providing feedback as students can ask questions if they do not understand feedback.

RQ3: What are L2 writing teachers’ overall perceptions of Grammarly as a complement to their feedback?

To answer the last research question, the participants’ interview data were scrutinized. The interview data revealed that while four participants were positive about

Grammarly, two were pessimistic about using Grammarly in their L2 writing classroom.

Mei, Maria, Rob, and Heaven were favorable of Grammarly and reported that they would use it in their L2 writing course. For instance, Mei said, “I had a positive experience, and also it is a trend for teachers to use Grammarly. So I feel good about using Grammarly as support.” Similarly, Maria said, “I thought it was good because I didn’t really spend any of my own time making comments on grammar.” She then added “Grammarly is already really widely used by native and non-native speakers of English. So I think there's no reason to exclude it. And the more tools we can give to our students to improve their English, the better.” Rob stated that “automated writing feedback, augmented reality, [...] artificial intelligence in the education sector are inevitable.” He believes that today, “there is Grammarly, tomorrow there will be something else.” Therefore, he thinks that instead of avoiding this phenomenon, teachers should “reinforce the happening in a positive direction which can support teaching.” Heaven noted that having a tool to take care of grammar issues can make them an “effective teacher” as they will be able to allocate more time and effort to global aspects of writing. They then added, “Grammarly is beneficial, and it is going to be something that I take into future classes.”

Contrary to the four participants, Mik and Jackson were pessimistic about Grammarly. Mik thinks that Grammarly has a lot of limitations, such as it skips a lot of L2 errors and that its automated nature may not increase any type of awareness of the error. So she feels hesitant to introduce it to her students. Mik also said, “I think that it would quicken our work and hopefully make us focus on other things in the writing that are more important in my opinion. But I just don't think it's there yet, but it can get there, and I hope it gets there because it would be nice to have it." Jackson also noted that Grammarly skips a lot of errors and "if there is supposed to be a division of labor, that division of labor maybe existed for 50 % of the time" because he had to address form issues if they were not covered by Grammarly. More importantly, however, Jackson emphasized the fact that Grammarly feedback can be discouraging for students. In this regard, he said, "I think because of how many errors there are sometimes labeled, it'd be incredibly discouraging;" therefore, Jackson believes that Grammarly may not be good for ESL students. However, he added that "it doesn't hurt to make students aware that it's available."

Overall, all the participants noted that for Grammarly to be beneficial, training is needed for both teachers and students. The participants noted that teachers should familiarize themselves with Grammarly to learn about its affordances and limitations. The participants also reported that teachers should devote several lessons to student training. According to the participants, the teachers should inform their students of what Grammarly can and cannot do, and they should also teach them how to respond to automated feedback. To this end, Maria said:

> I would give them a demonstration of how to use it step by step, upload an example paper, show them the different kinds of mistakes the tool could catch. And you know, basically, explain why it's important to look at every error and also make it clear that technology is not perfect. So it's definitely not going to catch everything and definitely not going to catch certain kinds of mistakes.

After using Grammarly to complement their feedback and knowing its affordance and limitations, all the participants came up with ways to use Grammarly in their L2 writing classrooms. For example, according to Mik and Heaven, one way to use Grammarly is to analyze its reports to get an overview of the most frequent errors and have in-class activities on those errors. Jackson thinks teachers should select the most frequent errors Grammarly identifies and give students a “focused list of grammar points” instead of giving them the Grammarly report or telling them to use Grammarly on their own. Jackson believes that this is more beneficial and can be less overwhelming and discouraging for students. Rob stated that he would ask his students to keep a diary or a checklist with errors Grammarly identified in their writing and apply rules Grammarly suggested to their future writing. Mei believes that asking students to submit a Grammarly report with their essay to show what changes they have made based on Grammarly suggestions could help teachers monitor their students’ progress and see if their students are engaging with Grammarly feedback. Because Grammarly feedback is prescriptive, Heaven suggests teaching a unit on “why Grammarly sometimes says a specific clause should say ‘that’ instead of ‘which’ and it’s a prescriptive thing and why prescriptive rules might exist.” However, before implementing Grammarly in L2 writing classrooms, Heaven thinks that teachers should consider the following questions: “Do you feel like students are at a point where they would use Grammarly as a crutch instead of using their own judgment? Do you have time in your curriculum to implement it? How do you plan to implement it? How do your students respond to feedback?”

Finally, the participants noted that Grammarly should be used in conjunction with other types of feedback, such as teacher feedback, peer feedback, and feedback from writing center consultants. To this end, Mik said, “My recommendation would be to implement other ways to look at grammar. Don’t just let [Grammarly] be the only thing.”

Discussion

The quantitative findings of the study revealed that despite using Grammarly to complement their feedback, the participants provided feedback both on HOCs and LOCs along with other types of feedback (e.g., positive feedback). This is in line with previous research that suggests that there is no division of labor such as that AWE takes care of LOCs as it is more computationally adept at providing such feedback and a teacher takes care of HOCs (Jiang et al., 2020; Wilson & Czik, 2016). It seems that the premise of labor division between a teacher and AWE should be abandoned as teachers still feel the need to provide feedback on sentence-level issues regardless of AWE's feedback. In this study, teachers felt the need to provide feedback on sentence structure, word choice, word form, spelling, punctuation, and documentation or attribution.

A closer look at the qualitative data revealed three factors that might have influenced the participants' feedback when they used Grammarly as a complement. The first factor is the participants' use of Grammarly reports. Five participants used Grammarly reports to complement their feedback and consulted the reports before or while providing their own feedback to see what errors Grammarly detected. This helped them make decisions about which errors to address in students' writing. One participant, Mei, did not use Grammarly reports to complement her feedback. Instead, she first provided her own feedback, and then she looked at the reports to compare her feedback with Grammarly feedback. In the interview, she reported that she wanted to read students' drafts for herself first to make her own judgments because if she had looked at Grammarly reports first that might have impacted what she thought of the paper. As a result, Mei ended up providing the highest number of feedback on LOCs among all the participants. Despite the fact that Mei did not use Grammarly to complement her feedback, which could be due to her low trust in the tool, she reported that she would use Grammarly in the future as she found its feedback accurate.

The second factor that might have influenced teacher feedback is the participants' perceptions of Grammarly feedback. All the participants noticed that Grammarly skipped a lot of errors on LOCs; therefore, the participants also provided feedback on sentence-level issues. The participants also noticed that Grammarly caught, as they said, only "basic" errors, such as subject-verb agreement and possessive noun endings, which in the literature are defined as "treatable" errors. That is, an error "related to a linguistic structure that occurs in a rule-governed way" (Ferris, 2014, p. 36). A closer look at quantitative findings revealed that the participants' feedback on LOCs focused on sentence structure, word form, and word choice, which are considered "untreatable" errors. That is, an error is "idiosyncratic, and the student will need to utilize acquired knowledge of the language to self-correct it" (Ferris, 2014, p. 36). Additionally, because Grammarly feedback was negative which could be discouraging for students, the participants felt the need to provide positive feedback. Interestingly, the participants indicated that Grammarly liberated their time to focus more on HOCs. This seems to go against the quantitative findings of the study that revealed that only two participants provided more feedback on HOCs while four participants provided more feedback on LOCs. One explanation could be the HOCs/LOCs dichotomy used in this study. The participants seemed to regard feedback on sentence structure, word choice, word form, and documentation or attribution (i.e., untreatable errors) as feedback on HOCs, while in this study, these error categories were coded as LOCs which seems problematic and suggests that future studies should use the treatable/untreatable dichotomy (Ferris, 2014).

Because Grammarly took care of treatable errors, and the participants took care of errors on global aspects of writing along with untreatable errors, which they regarded as feedback on HOCs, they had a sense that Grammarly freed them up to focus more on global aspects of writing. These findings partially support the claim made in the extant literature that AWE liberates teachers' time to focus more on HOCs (Jiang et al., 2020; Wilson & Czik, 2016).

The last factor that might have impacted teacher feedback is the participants' feedback practice and beliefs about feedback and the L2 writing course. All the participants reported that students taking the L2 writing course need feedback on both HOCs and LOCs because they are developing both writing and linguistics skills but feedback on HOCs is often prioritized. Although this contradicts their actions as seen in the quantitative findings, the participants truly believed Grammarly helped them focus more on HOCs. The participants also emphasized the fact that Grammarly may not well align with the L2 writing course objective as students may accept its feedback uncritically which will not lead to true learning, and students will not be able to resolve errors on their own. Therefore, they expressed their preference for other sources of feedback such as peer-review and writing center feedback in addition to automated feedback as automated feedback is impersonal. Such concerns have also been raised in previous research (Ericsson, 2006; Herrington & Moran, 2006)

As for overall perceptions of the tool, of the six participants, four were positive about Grammarly and reported they would use in their L2 writing classroom. Although the participants were aware of Grammarly's limitations, they were positive about it due to its numerous benefits such as it provides detailed feedback, allows them to focus more on HOCs, and gives an overview of the most frequent errors in students' writing. The participants also noted that tools like Grammarly are inevitable, and instead of resisting them, teachers should find ways for their effective use. However, two of the participants were skeptical about introducing Grammarly to their ESL students due to the fact that it does not detect all L2 errors and can be overwhelming because sometimes it provides too many suggestions. The findings suggest that when teachers experience using AWE, this allows them to recognize its strengths and weaknesses which, in turn, can help them make educated decisions about the implementation of AWE in their classrooms (Cotos, 2018; Li, 2021; Link et al., 2014; Weigle 2013). Additionally, the findings support Chen and Cheng's (2008) claim that the limitations inherent in AWE can have a negative impact on teachers and could result in rejection of the idea of using the tool in the classroom.

Conclusion

The study explored postsecondary, L2 writing teachers' use and perceptions of Grammarly as a complement to their feedback. By doing so, the study extended the extant literature on teachers' use and perceptions of AWE in several ways. First, the study gives a comprehensive picture of how postsecondary, L2 writing teachers use Grammarly to complement their feedback, which has not been reported in previous research. Second, the study reveals the impact of Grammarly on teacher feedback and factors that might have influenced teacher feedback when Grammarly was used as a complement. Finally, the study provides implications for how to use Grammarly effectively as a complement to teacher feedback.

In light of the study findings, Grammarly has the potential to be used as a complement to teacher feedback but certainly should not be used to replace teacher feedback (Jiang et al., 2020; Weigle 2013). Teachers should test the tool on their own to identify its limitations and affordances to make informed decisions about the use of such tools in their classrooms (Cotos, 2018). Along with considering the limitations and affordances of the tool, teachers should also take into account their feedback beliefs and practices and also course objectives. If teachers decide to use Grammarly and similar tools in their classrooms, they should offer explicit training to their students on how to use these tools and respond to automated feedback (Koltovskaia, 2020)

The following are the implications for Grammarly use as a complement to teacher feedback. Because of Grammarly's limitation in skipping some of the L2 errors, teachers are advised to provide feedback on sentence-level issues too. Furthermore, teachers' feedback on local aspects of writing should focus on untreatable errors because Grammarly is more computationally adept at detecting treatable errors. Research suggests that feedback on untreatable errors should be direct as students find it difficult to resolve such errors on their own (Ferris, 2014). Additionally, because of the prescriptive nature of Grammarly feedback, students may not be aware of descriptive uses of grammar. Therefore, teachers are advised to provide lessons on prescriptive vs. descriptive grammar. Since Grammarly feedback is negative, teachers should also provide positive feedback to avoid discouragement during the revision process. Finally, Grammarly detects errors comprehensively and may provide a lot of comments; therefore, its feedback could be overwhelming for students. Studies on written corrective feedback suggest that comprehensive feedback can indeed be overwhelming, confusing, and discouraging (Lee, 2019; Sheen, Wright, & Moldawa, 2009). Therefore, it is recommended for teachers to use Grammarly to determine the most frequent errors in individual students' writing and provide focused feedback for each student.

While the findings of the study are informative for L2 writing teachers, some limitations should be acknowledged. The study focused on the hypothetical scenario and provided insight into teachers' one-time use of Grammarly. Future studies should be conducted in the actual L2 writing classroom and should explore how teachers' use and perceptions of AWE change over time. Considering Grammarly provides feedback primarily on treatable errors, future studies could consider the treatable/untreatable dichotomy and examine the impact of such feedback on students.

CHAPTER IV

STUDENTS' ENGAGEMENT WITH GRAMMARLY FEEDBACK

Introduction

The use of automated writing evaluation (AWE) systems and similar tools for assessment purposes in second language (L2) writing classrooms has rapidly increased due to their numerous advantages. AWE systems have been claimed to provide computer-generated quantitative and qualitative feedback (Dikli, 2006). They have been considered to offer multiple practice and revision opportunities (Warschauer & Ware, 2006) and be more consistent and objective than human raters (Wang, Shang, & Briody, 2013). Especially, AWE programs have been praised for their capacity to free up teachers' time to focus less on lower-order concerns (e.g., grammar, mechanics) and more on higher-order concerns (e.g., content, organization) (Ranalli, 2018) and other aspects of writing instruction (Warschauer & Grimes, 2008).

Notwithstanding AWE benefits, previous studies suggest that students do not make good use of automated feedback (Attali, 2004; Chapelle, Cotos, & Lee, 2015; Warschauer & Grimes, 2008). The bulk of research, however, has mainly focused on students' final written products rather than their revision process (Stevenson & Phakiti, 2014). Stevenson and Phakiti (2019) noted that this focus on product over process tells us little about to what extent L2 learners developed metacognitive skills to notice, evaluate, and consequently improve writing. Therefore, studies are needed on student engagement with automated feedback during the revision process to help understand the benefits of such feedback. Despite its importance, research on student engagement with automated feedback is surprisingly scarce.

Motivated by a lack of systematic research on this subject, this case study has explored ESL university students' engagement with automated written corrective feedback (AWCF) provided by Grammarly when revising a final draft. This study offers new insights into the process through which students engage with AWCF by presenting a holistic narrative of two cases.

Literature Review

Research on AWE feedback

The two central components of AWE are a scoring engine that generates automated scores and a feedback engine that provides automated written feedback (Bai & Hu, 2017), also known as automated written corrective feedback (AWCF) (Ranalli, 2018). AWE originated from automated essay scoring (AES) and was initially used in high-stakes testing to generate numeric scores for summative assessment (assessment of learning) based on such techniques as artificial intelligence, natural-language processing, and latent semantic analysis (Cotos, 2014; Stevenson & Phakiti, 2019). The most notable and cited among AES systems are Project Essay Grade (PEG) developed by Ellis Page, Intelligent Essay Assessor™ (IEA) from Pearson Educational Technologies, Electronic Essay Rater (e-rater) from Educational Testing Services, and IntelliMetri from Vantage Learning. In recent years, various AWE systems, including Criterion from Educational

Testing Service, My Access! from Vantage Learning, WriteToLearn from Pearson Educational Technologies, and others have been developed not only for summative but also formative assessment (assessment for learning) purposes to be used in writing classrooms (Chen & Cheng, 2008). While earlier AWE research largely focused on the validity and reliability of its scoring system in testing contexts (Attali & Burstein, 2006; Burstein et al., 1998; Burstein & Chodorow, 1999; Elliot, 2002; Enright & Quinlan, 2010; Powers, Burstein, Chodorow, Fowles, & Kukich, 2002), recent studies have addressed instructional use of AWE. These classroom-based studies have explored student perceptions of the usefulness of AWE quantitative and qualitative feedback (Chen & Cheng, 2008; Dikli & Bleyle, 2014; Grimes & Warschauer, 2010; Lai, 2010) and investigated the effects of automated feedback on writing (Attali, 2004; Chapelle et al., 2015; El-Ebyary & Windeatt, 2010; Li, Link, & Hegerlheimer, 2015; Li, Feng, & Saricaoglu, 2017; Liao, 2015).

Student perception studies of AWE feedback report ambivalent findings. In their comparative study, Dikli and Bleyle (2014) found that ESL students generally perceived Criterion feedback to be helpful while, at the same time, valuing instructor feedback. In Grimes and Warschauer's (2010) study, the U.S. middle school students rated My Access! favorably for its usefulness, fairness, and user-friendliness. The students reported that the system motivated them to write and revise their papers and increased their confidence. Conversely, Chen and Cheng (2008) found that Taiwanese EFL students perceived the use of My Access! unfavorably at large, which the authors noted was attributable to limitations inherent in the system's assessment and assistance functions. However, the students identified the use of the system positively if it was utilized with the instructor's facilitation. Similarly, in her comparative study, Lai (2010) found that Taiwanese EFL students mostly held negative perceptions of My Access! feedback because it was too general for them to make revisions, thus preferring peer evaluation over automated feedback.

The results of the previous research on the effects of AWE feedback on writing are also mixed. Some studies report positive effects on writing (El-Ebyary & Windeatt, 2010; Li et al., 2015; Li et al., 2017; Liao, 2015; Wang et al., 2013). For example, El-Ebyary and Windeatt (2010) found that Criterion feedback positively impacted the quality of Egyptian EFL students' writing although some students achieved better scores by using the avoidance strategy. The authors also reported that, unlike conventional writing/feedback modes, Criterion encouraged students to revise their essays (100% resubmission rate). Likewise, regarding the effects of AWE feedback on draft revisions, Li et al. (2015) found that Criterion led to increased revisions, and its feedback helped ESL students improve their linguistic accuracy. Conversely, Attali (2004) found that 71% of the U.S. sixth to twelve grade participants submitted their essays to Criterion once, indicating that most students did not utilize the revision capabilities of the system. Other studies also report discouraging findings that the majority of students who submitted their drafts for scoring submitted them only once and made limited revisions upon receiving automated feedback (Warschauer & Grimes, 2008; Warschauer & Ware, 2006). Chapelle et al. (2015) found that ESL students disregarded nearly 50% of Criterion feedback, despite the provision of both direct and indirect feedback, thus making limited changes to their drafts. The authors suggested this was due to inaccuracies in Criterion feedback.

Since the findings of the aforementioned studies on automated feedback are contrasting, it is difficult to definitively conclude whether students make the most of it on their writing. Zhang (2017) claimed that to benefit from feedback, students need to be effectively engaged with it. Student engagement with feedback is also believed to be a key factor in the success of writing development and language acquisition (Zhang & Hyland, 2018). While previous studies on AWE have provided insight into students' perceptions of automated feedback and how students utilize it to revise their texts as it pertains to revision operations (e.g., accept feedback) and times of submission, the majority of these studies have focused on students' written products, with little attention paid to their revision process (Stevenson & Phakiti, 2019). According to Zhang (2017), without careful investigation of how students engage with automated feedback during the revision process, it is impossible to know what factors facilitate or inhibit their response to such feedback.

Construct of Student Engagement and Empirical Research

Perhaps the most well-known conceptualization of engagement was proposed by Fredricks, Blumenfield, and Paris (2004). They viewed engagement as a multifaceted construct that encompasses three interrelated dimensions: behavioral, emotional (affective), and cognitive. However, their conceptualization of engagement was proposed for school engagement. Ellis (2010) applied Fredricks et al.'s tripartite conceptualization to student engagement with both oral and written corrective feedback (CF) in which *behavioral perspective* concerned learners' uptake and revisions elicited by CF, *affective perspective* referred to learners' attitudinal response to CF, and *cognitive perspective* involved "how learners attend to the CF they receive" (p. 342). Han and Hyland (2015) furthered Ellis's framework to explore four Chinese college students' engagement with teacher written corrective feedback (WCF). In their study, *behavioral engagement* involved revision operations in response to WCF and observable strategies used in improving the accuracy of drafts, future writing, and/or L2 competence. *Cognitive engagement* referred to the depth of processing of WCF encompassing cognitive and metacognitive operations. *Affective engagement* concerned learners' immediate emotional reactions and attitudinal responses toward WCF. The results showed that although the students received similar in terms of scope, type, and frequency teacher WCF, they engaged with it differently due to individual differences and contextual factors. The authors concluded that students, as active agents of their own learning, can decide how and what they learn from teacher WCF. Zheng and Yu (2018) applied the developed framework to their study on low proficiency university students' engagement with teacher WCF. The researchers found that while the students' affective engagement with teacher WCF was relatively positive, their behavioral and cognitive engagement with WCF was at a limited level as it was negatively impacted by their low English proficiency.

Zhang and Hyland (2018) further strengthened the framework to investigate two Chinese university students' engagement with both teacher WCF and AWE feedback provided by the Chinese AWE system, *Pigai*. They viewed *behavioral engagement* as students' behavioral reaction to feedback, including revision actions and time spent on revision. *Affective engagement* involved students' emotional responses and attitudinal reactions to feedback, while *cognitive engagement* concerned how students attend to feedback and their use of revision operations and cognitive (metacognitive) strategies. The results showed that the highly engaged student preferred AWE feedback over teacher feedback because the former provided immediate feedback and allowed her to resubmit her essay 13 times; thus promoting her autonomy. Conversely, the moderately engaged student had limited engagement particularly with AWE feedback because he was overwhelmed by the amount of feedback provided and felt embarrassed and demoralized by the low score he received. Zhang (2017) also focused on a Chinese university student engagement with *Pigai* feedback. *Behavioral engagement* in his study, however, referred to the number of submissions and the time spent on revisions. *Emotional engagement* involved affective reactions and motivational changes. *Cognitive engagement* concerned understanding the feedback information, monitoring the revision process, and self-regulating. The results revealed the student was engaged with *Pigai* feedback behaviorally, emotionally, and cognitively. The author noted that when engaging in multiple revisions, the student felt motivated by the prospect of getting higher holistic scores and felt demotivated when multiple revisions resulted in low scores.

In line with previous research, engagement with AWCF, in this study, is also seen as composed of three interrelated dimensions where:

- *Behavioral engagement* concerns revision operations, i.e. actual revisions carried out, revisions strategies used to improve the accuracy of the draft, and time spent on revision.
- *Cognitive engagement* concerns how deeply students process AWCF (noticing or understanding) and their use of metacognitive and cognitive operations.
- *Affective engagement* concerns students' immediate emotional reactions and attitudinal responses to AWCF.

While the aforementioned studies shed light on student engagement with teacher and AWE feedback, more studies are needed to explore student engagement with automated feedback to unlock its benefits. Particularly, studies need to focus on student engagement with Grammarly feedback. As Ranalli (2018) noted, Grammarly is making important inroads into L2 classrooms for its capacity to provide more specific feedback. Besides, research on Grammarly, generally, reports positive results which suggest its use in writing classrooms is worth considering. For example, O'Neill and Russell (2019) investigated students' perceptions of Grammarly when it is used together with academic learning advisor (ALA) feedback. They found the group that received Grammarly feedback along with ALA feedback was significantly more satisfied than the group that received only ALA feedback. The participants reported liking Grammarly feedback because it was detailed, thorough, line-by-line, and prompt. Qassemzadeh and Soleimani (2016) explored the impact of Grammarly and teacher feedback on learning passive structures. They found that both Grammarly and teacher feedback can positively influence learning of passive structures. However, the role of the former in retaining passive structures is more highlighted than the latter. It is noteworthy that unlike many AWE programs such as Criterion or My Access!, which provide both numeric scores and AWCF, Grammarly provides only AWCF. Additionally, Criterion and many similar AWE systems are standalone systems that provide feedback episodically and in bulk, whereas Grammarly can be integrated into any word-processing environment and provide feedback in real-time and in bits (Ranalli & Yamashita, 2022). Because of the nature and timing of the feedback Grammarly provides, it has recently been termed an AWCF tool (Ranalli, 2018; Ranalli & Yamashita, 2022) rather than AWE. In this study, Grammarly will be referred to as an AWCF tool. While research on Grammarly has focused on students' perceptions of Grammarly feedback and its effectiveness in retaining certain grammatical structures, no work has been done on how ESL students engage with AWCF provided by Grammarly. Therefore, the current case study employs a multidimensional framework of student engagement to answer the following research question:

> **RQ:** How do students behaviorally, cognitively, and affectively engage with AWCF provided by Grammarly when revising their final draft?

Methods

Research Overview

The research design used in this study was a case study, which provides an in-depth, holistic, and contextualized understanding of the phenomenon under investigation (Yin, 2009). In particular, a multiple-case study was employed to explore how two students engage with AWCF.

Participants and Classroom

The study took place at a large southcentral university in the U.S. Seventeen undergraduate students enrolled in the International Freshman Second Language Writing course (ENGL 1223) during the fall 2018 semester were recruited for two reasons: 1) the students were L2 learners of English and 2) the students took a writing course in which they were required to produce a multiple-draft assignment. Eight out of 17 students volunteered to participate of which only two adequately completed all aspects of the study (Appendix J). The two participants were Alex and Kelsey (pseudonyms). Table 5 shows the participants' profiles outlining their demographic information along with their major and class standing.

Table 5

Participants' profiles

Name	Age	Country	First language	Major	Class standing
Alex	21 (1998)	China	Cantonese, Mandarin	Journalism	Junior
Kelsey	22 (1997)	Saudi Arabia	Arabic	Computer Science	Freshman

At the beginning of the semester, the students had an in-class diagnostic assessment for which they were required to write a summary of an article (see Appendix J for a diagnostic writing prompt). The researcher and two of her colleagues independently evaluated the two participants' texts using the slightly modified TOEFL iBT Test - Independent Writing Rubric - to determine their language proficiency and writing skills. The texts were evaluated based on a scoring rubric of 0-5. The mean rubric score given by the raters for the quality of the students' writing was then converted to a scaled score of 0-30. Alex received a scaled score of 25 (the rubric score of 4), which means he is an advanced L2 writer, while Kelsey received a scaled score of 14 (the rubric score of 2.7), which means she is a low-intermediate L2 writer (see https://www.ets.org/toefl for more details). Although it was Alex's first semester in the U.S., he was one of the best students in the class. Unlike his classmates, his writing skills were better developed. The level of Kelsey's writing skills was average and thus similar to that of the students taking this course.

ENGL 1223 was a 16-week, three-credit research writing course. It was restricted to undergraduate students whose native language was not English. The class met three times a week for 50 minutes in each session. The major assignments of the course were an annotated bibliography, a research proposal (introduction, literature review, methodology), and a research proposal presentation. The researcher was a teacher of the course. The researcher is a non-native speaker of English from Russia, and she was in her third year of Ph.D. studies in TESOL and Applied Linguistics at the time of the study. Creswell (2014) noted that in qualitative research, the inquirer, as the primary data collection instrument, should explicitly state his/her role as this may shape the direction of the study. The researcher's four years of teaching experience and five years of Writing Center work in the U.S. influenced the way she teaches writing at a college level and provides feedback. She believes teachers should give feedback on higher-order concerns at the early stages of writing and lower-order concerns at the last stage of writing. However, often due to time constraints, she provides feedback predominantly on global issues, thus leaving grammar and mechanics for students to revise on their own. Therefore, she believes that as a complement to teacher feedback, AWE systems could empower students to revise their own work because AWE are more computationally adept at providing feedback on low-order concerns (Ranalli, Link, & Chukharev-Hudilainen, 2017), particularly the use of the easily accessible Grammarly.

Grammarly

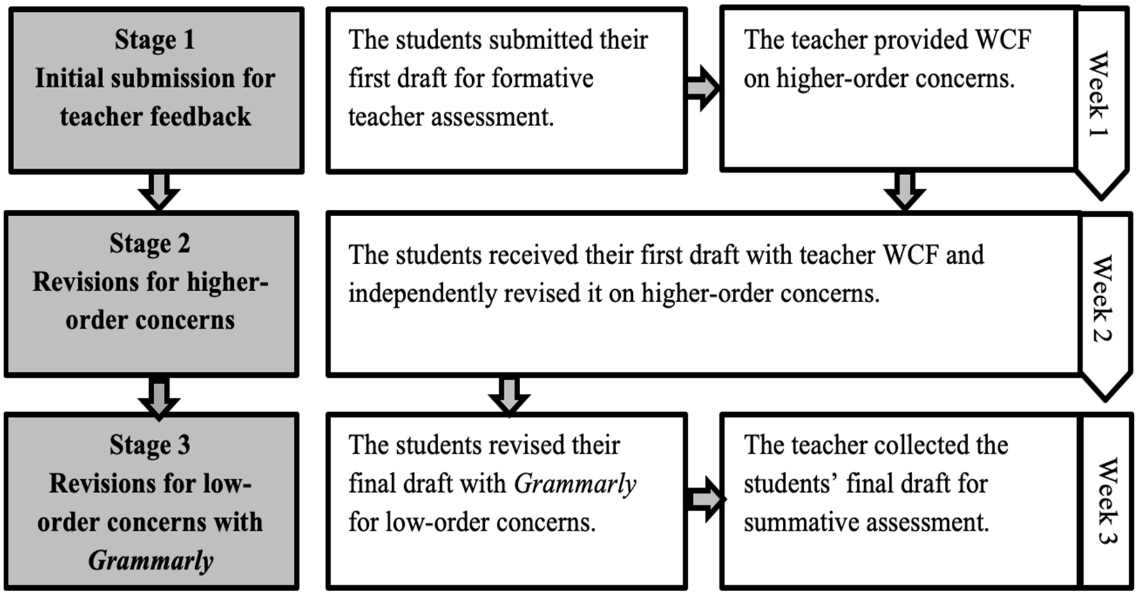
In this study, a free version of Grammarly (https://app.grammarly.com/) was utilized. The free version of Grammarly provides feedback on spelling, punctuation, grammar, and conventions, including spacing, capitalization, and dialect-specific spelling. Grammarly instantly provides feedback for improvement once a paper is uploaded online. The uploaded paper appears on the left side of the screen with errors underlined in red (i.e., indirect feedback) while direct feedback appears on the right side of the screen (Appendix K). Direct feedback contains the error type (e.g., grammar), possible error correction (e.g., Korean Peninsula -> the Korean Peninsula), and a suggestion (e.g., It appears that an article is missing before the word **Korean.** Consider adding the article.). Suggestions can be expanded for a comprehensive explanation of a grammar rule, i.e. a metalinguistic explanation (Appendix L).

Three-Stage Revision Process

For this study, the students worked on a literature review. The literature review was a multi-draft assignment for which the students were expected to use four peer-reviewed scholarly articles relevant to their research topic of interest. The information of what the literature review assignment is and the assignment rubric are shown in Appendix M and Appendix N, respectively. For this study, the literature review assignment involved three stages (Figure 7).

Figure 7.

The Three-stage Revision Process of the Literature Review Assignment


Stage 1 Initial submission for teacher feedback
The students submitted their first draft for formative teacher assessment.
The teacher provided WCF on higher-order concerns.
Week 1
Stage 2 Revisions for higher-order concerns
The students received their first draft with teacher WCF and independently revised it on higher-order concerns.
Week 2
Stage 3 Revisions for low-order concerns with *Grammarly*
The students revised their final draft with *Grammarly* for low-order concerns.
The teacher collected the students' final draft for summative assessment.
Week 3

Stage 1-Initial submission for teacher feedback. At the beginning of the first week, the students uploaded the first draft of their literature review to a Dropbox in the D2L Brightspace Learning Environment to receive formative teacher assessment. When writing their first draft, the students were asked to focus primarily on higher-order concerns.

Stage 2-Revisions for higher-order concerns. In the middle of the second week, the students received their first draft with teacher WCF on higher-order concerns. They were asked to independently revise their papers based on teacher WCF by the end of the second week. It is important to note that in Stage 1 and Stage 2, the students might have addressed low-order concerns. However, any revisions on low-order concerns that were completed at this time were not considered due to the study's emphasis on revision using Grammarly.

Stage 3- Revisions for low-order concerns with Grammarly. In the middle of the third week, the students worked in the computer classroom on revising their final draft with Grammarly. At the end of the class, the students submitted their final draft for summative teacher assessment. At this stage, formal data collection was employed.

Data Collection

Data triangulation was included in the research design. The research data were collected by means of screencasts, stimulated recall, and semi-structured interviews. While revising their final draft with Grammarly, the students were asked to record their revision process with QuickTime player. Before that, the students were trained on how to use the QuickTime player screencast function and Grammarly. This function allowed the students to record video of their computer screen and capture their revision process. The revision process lasted one class period. The students' screencasts were collected at the end of the class.

The students then had an individual stimulated recall session in which they watched their screencast video on a TV monitor and were asked to recall their thoughts at the time of correction of each error for which they received AWCF. According to Gass and Mackey (2000), a recall should occur as soon as possible after the event. In this study, the recall occurred within 48 hours of the students' error correction activity. The recall script and guiding questions can be seen in Appendix O. This introspective method was used to access students' thoughts as they were carrying out the revision activity (Gass & Mackey, 2000) to explore their cognitive, affective, and behavioral engagement with AWCF.

After the recall, the students had a semi-structured interview consisting of ten questions to explore their affective engagement with Grammarly and its feedback (Appendix P). Additional follow-up questions were asked to build up a more complete picture. Before the interview, a pilot test was conducted with a volunteer student from the other section of the same course to make necessary modifications to the questions. Some adjustments were made, including clarification of ambiguous questions and simplification of the language. The entire reflective exercise (stimulated recall and semi-structured interview) lasted roughly an hour. It was conducted in English and audio recorded.

Data Analysis

Analysis of screencasts

Behavioral engagement was explored through the analysis of the two students' screencasts. The analysis consisted of five phases. The first phase included the reclassification of error types detected by Grammarly. Although Grammarly automatically categorizes all errors into four error types (spelling, grammar, punctuation, and conventions), the error type *grammar* is too broad. Besides, its classification of error types raises doubts. For instance, Grammarly suggested a participant changing the word *educational* to *education* and classified this error as a spelling error. However, the error has to do with word form rather than spelling. Therefore, to address the ambiguity of error categorization in Grammarly, the errors for which the participants received AWCF were categorized according to Han and Hyland (2015)'s taxonomy of error categories (Appendix Q).

Additionally, because automated feedback is prone to be fallible (Chapelle et al., 2015), and this can ultimately affect student engagement with such feedback, the second phase involved diagnosis of the accuracy of AWCF the students received. For example, Grammarly suggested a student adding the indefinite article before *little* in the sentence: *Although there is many research that focus on the effect of insomnia on college students, only little literature investigated the relationship between insomnia and the college student's sleep hygiene in X University.* This AWCF was identified as inaccurate because adding *a* before *little* changes the intended meaning of the sentence, which was to show the gap by emphasizing almost no literature exists on the topic (Appendix R).

In the third phase, revision operations were identified. Revision operations were operationalized as any actions taken in response to AWCF. After repeatedly going through the screencasts, three categories of revision operations were determined: accept, reject, and substitute (Appendix S). If the correction suggested by Grammarly was clicked by the student to automatically fix an error, this was regarded as feedback being accepted. If the correction was left unclicked or dismissed and consequently no changes were made to the text, this was viewed as feedback being rejected. If the correction was substituted by the student's own correction to address the error, this was considered as feedback being substituted. To enhance the validity in data analysis, the inter-coder agreement was calculated after a second coder, a trained colleague, independently coded 50% of the data. The agreement rates for error categorization, accuracy/inaccuracy of AWCF, and revision operations were 96.1%, 100%, and 100%, respectively. The fourth phase involved the identification of revision strategies. Any strategies taken to enhance the quality of the draft in response to AWCF were identified as revision strategies such as consulting the Internet or dictionary to verify the accuracy of AWCF. The stimulated recall data was further analyzed to reaffirm any observed strategies.

In the fifth phase, the amount of time spent by the students on draft revision was determined by first counting how much time they spent on each error for which AWCF was given. Then, the total time of working on all errors in the draft was calculated for each student. Simple statistical calculations were made for all quantitative data.

Analysis of the stimulated recall and interview data

The students' recall and interview data were qualitatively analyzed to profile their behavioral, cognitive, and affective engagement with AWCF. Prior to coding, audio recordings of the recall and the interview were transcribed with Trint (https://trint.com) and then checked for accuracy against the original recordings. Then, all transcripts were organized by an individual participant. To further sort and reduce data, each participant's recall transcript was segmented into language related episodes (LREs), and the researcher's questions and the participant's responses were placed into two different columns in Excel: the researcher's questions in the first column preceded the participant's responses in the second column (Table 6). LRE was operationalized as any segment of the recall in which there is an explicit focus on a linguistic item (Swain & Lapkin, 1995), such as that one LRE corresponds to one error for which AWCF was given. The interrater agreement for identifying LREs was 100%.

Table 6

Example of Alex's Data Layout and Coding Sheet for Stimulated Recall

LRE 13	**Researcher**	**Alex**
AWCF: Heavily relied; Has heavily relied	So, when you were looking at this (AWCF) and then kind of decided to move on to the next one, why did you decide to move on to the next one?	Because I can't think of which one I should use. I'm not sure (**UNSURE***). So, therefore, I decided to put it aside and I reselected after this, maybe, easier ones (**FIXING EASY ERRORS***).

* Initial codes assigned to the participant's comments during the open coding stage.

Only the students' responses from the recall and the interview transcripts were coded following three steps: open coding, axial coding, and selective coding as described in Corbin and Strauss (2008). Initial codes were closely related to the original data in the recall and the interview. For example, a comment such as *I decided to put it aside and I reselected after this, maybe, easier ones* was assigned the initial code of 'fixing easy errors.' A memo was written next to the code: *the student recognized that some errors cause more difficulties than others. Thus, he first dealt with errors that were easy for him to fix*. Informed by previous studies that offered insight into the labeling and categorizing of student engagement (e.g., Han & Hyland, 2015; Zhang & Hyland, 2018), this code was further refined to 'planning for cognition' in the axial coding, which is one type of metacognitive operation. This code then was attributed to the 'cognitive engagement' category in selective coding. The codes were then compared across the two cases. After that, the case narratives were constructed. The same colleague again independently coded 50% of the data to enhance the validity. While the inter-coder agreement for behavioral and affective engagement reached 91.2 % and 93.4 %, respectively, the agreement rate for cognitive engagement was initially 67.4%. After extensive discussion and going through data multiple times, the agreement rate for cognitive engagement reached 82.7%.

Findings

Alex - Advanced L2 Writer

Alex's Behavioral Engagement with AWCF

Behavioral engagement with AWCF involved revision operations and strategies used to enhance the accuracy of the draft and time spent on revision. Table 7 illustrates Alex's revision operations in response to AWCF. The table shows Alex made 14 errors in his 914-word draft for which he received AWCF. Seven error types were identified in his draft, including errors on the use of verb form (1), word form (1), articles (5), punctuation (2), spelling (1), prepositions (3), and spacing (1). Of the seven error types, articles were the most frequent errors identified by Grammarly. Of 14 received AWCF, which were all accurate, Alex correctly accepted eight and incorrectly rejected six; thus fixing 57% of his total errors. This suggests Alex made moderate changes to his draft.

Table 7

Alex’s Revision Operations

Error type	**AWCF**	**Accurate AWCF**			**Inaccurate AWCF**		
		Accept	**Reject**	**Substitute**	**Accept**	**Reject**	**Substitute**
Verb form	**1**		1				
Word form	**1**	1					
Articles	**5**	2	3				
Punctuation	**2**		2				
Spelling	**1**	1					
Prepositions	**3**	3					
Miscellaneous*	**1**	1					
Total number of errors in a 914-word text	**14**	**8**	**6**				
Percentage	**100%**	**57%**	**43%**				

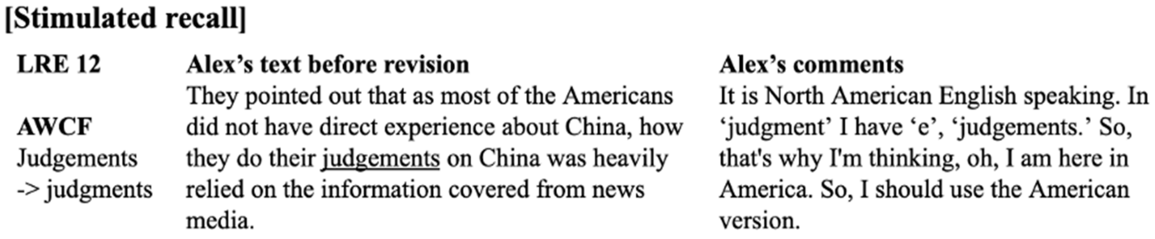
* Miscellaneous here refers to a spacing error.

Regarding revision strategies, Alex once sought extra assistance from the Internet to verify the accuracy of AWCF. To illustrate, Grammarly suggested Alex using the preposition *in* instead of *on* in the sentence *They showed that the media plays an important role on this international relationship.* The screencast video showed how Alex typed ‘play an important role’ in the Google search engine to see if this phrase often appears with *in* or *on.* Alex scrolled down to see examples, and once he realized the correct preposition is *in,* he accepted the feedback. As for time spent on making necessary revisions in his draft, Alex spent a little over five minutes (333 seconds) on this despite having a 50-minute class period.

Alex’s Cognitive Engagement with AWCF

Cognitive engagement concerned how deeply the students processed AWCF (noticing and understanding), their use of metacognitive operations to regulate their mental effort, and cognitive operations to process feedback and determine appropriate revisions. Regarding cognitive engagement with AWCF at the level of noticing, in the recall, Alex reported detecting all AWCF. He said feedback was conspicuous because errors were underlined in red and corrections were highlighted in green. Since Alex read Grammarly suggestions and because of AWCF's explicit and implicit nature, he said he relatively easily and quickly recognized the corrective intention of the majority of feedback.

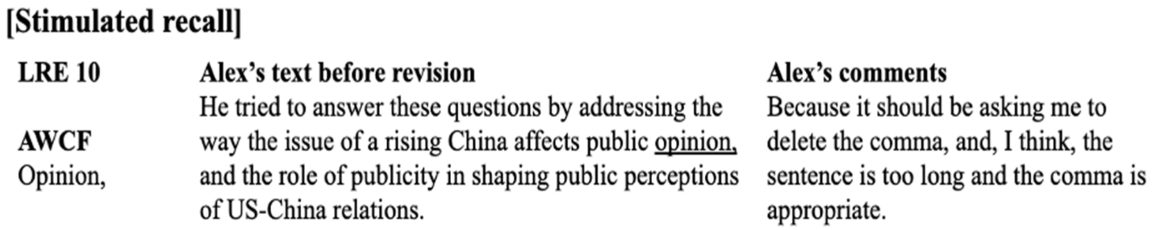
In terms of cognitive engagement with AWCF at the level of understanding, the recall revealed Alex understood the cause/nature of eight errors and how to correct those errors. Therefore, Alex accepted eight AWCF on those errors. The following example demonstrates Alex's understanding of the error:



[Stimulated recall]

LRE 12	**Alex's text before revision**	**Alex's comments**
AWCF Judgements -> judgments	They pointed out that as most of the Americans did not have direct experience about China, how they do their judgements on China was heavily relied on the information covered from news media.	It is North American English speaking. In 'judgment' I have 'e', 'judgements.' So, that's why I'm thinking, oh, I am here in America. So, I should use the American version.

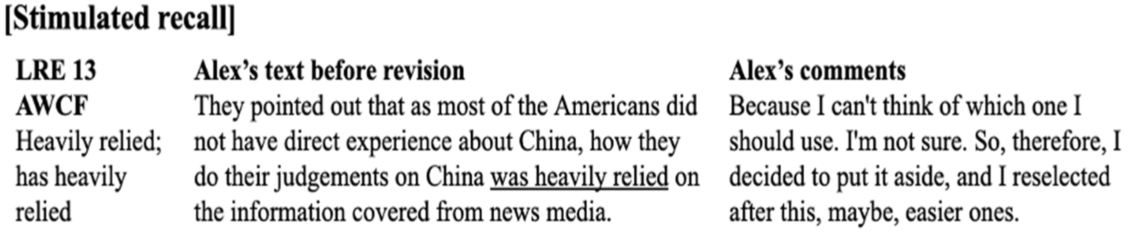
Alex appeared not to know the grammar rules of the six errors that were left uncorrected. For example, Alex incorrectly placed a comma in a compound object. In the recall, Alex reported reading Grammarly suggestion saying: *It appears that you have an unnecessary comma in a compound object. Consider removing it*. Despite the suggestion, Alex decided not to correct the error explaining this as follows:



[Stimulated recall]

LRE 10	Alex's text before revision	Alex's comments
AWCF Opinion,	He tried to answer these questions by addressing the way the issue of a rising China affects public opinion, and the role of publicity in shaping public perceptions of US-China relations.	Because it should be asking me to delete the comma, and, I think, the sentence is too long and the comma is appropriate.

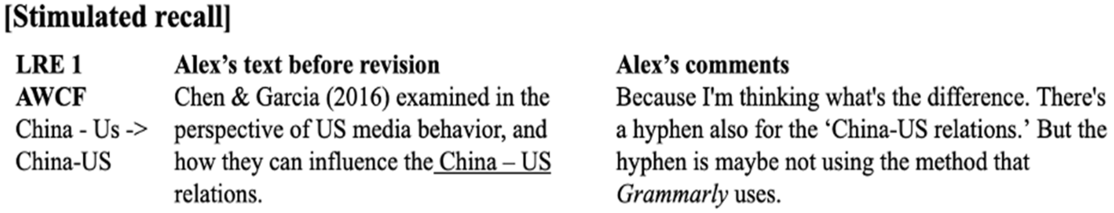
This example demonstrates Alex's lack of knowledge on the use of commas. If Alex had expended the suggestion for a metalinguistic explanation on comma usage, this could have facilitated understanding of the error he incorrectly rejected. Instead, he relied on his intuition, thus deploying the wrong revision operation. Overall, Alex's cognitive engagement was relatively extensive which was manifested in the use of several metacognitive and cognitive operations to regulate his mental processes. The metacognitive operation that helped Alex regulate his mental effort was planning for cognition:



[Stimulated recall]

LRE 13	Alex's text before revision	Alex's comments
AWCF Heavily relied; has heavily relied	They pointed out that as most of the Americans did not have direct experience about China, how they do their judgements on China was heavily relied on the information covered from news media.	Because I can't think of which one I should use. I'm not sure. So, therefore, I decided to put it aside, and I reselected after this, maybe, easier ones.

This operation demonstrates Alex was strategic. He first determined the difficulty level of an error. Then, he addressed easier errors, thus saving difficult errors for later. Another metacognitive operation Alex deployed was evaluating the outcome of the task by double- or triple-checking the revision operation he used. He often went back to the errors he accepted/rejected and read the entire sentence several times to ensure his decision was right. As for the use of cognitive operations, the recall revealed that to process feedback and figure out the appropriate revision operation, Alex reasoned:



[Stimulated recall]

LRE 1	Alex's text before revision	Alex's comments
AWCF China - Us -> China-US	Chen & Garcia (2016) examined in the perspective of US media behavior, and how they can influence the China – US relations.	Because I'm thinking what's the difference. There's a hyphen also for the 'China-US relations.' But the hyphen is maybe not using the method that *Grammarly* uses.

Additionally, Alex used context to determine the appropriate revision operation. He read the sentence where the error was detected to decide if the correction suggested by Grammarly worked in that sentence. Sometimes reading the sentence where the error was detected was not enough; thus, Alex needed more context to decide whether to accept or reject AWCF. In the recall, he stated, "I will first read a sentence with the error. If it is still not clear, I will read more sentences before the sentence with the error to check if this is correct or not."

Alex's Affective Engagement with AWCF

Affective engagement referred to the students' emotional reactions toward AWCF upon receiving it and their overall attitude toward AWCF. Alex's emotional reaction toward AWCF was often distrust. Alex questioned AWCF because he appeared not to find it as authoritative as his teacher's feedback. Although this was the first time Alex used Grammarly, he knew such computer-generated feedback can be occasionally inaccurate. In the interview, Alex reported, "[...], sometimes I was not 100% sure about what Grammarly recommended me. So, [...] I have to check it again and proofread whether feedback is correct or not."

This emotional reaction of doubt toward AWCF affected Alex's cognitive and behavioral engagement, which indicates that the three dimensions of engagement are linked to one another. Alex was quite extensively cognitively engaged with AWCF because he questioned its accuracy. However, instead of seeking extra assistance from other resources to verify the feedback's accuracy, Alex relied on his linguistic knowledge and intuition. His lack of knowledge of certain grammar and mechanics rules, however, resulted in rejecting six out of 14 accurate feedback.

Overall, Alex had a positive attitude toward AWCF and Grammarly. He found the tool to be helpful because, as he said in the interview, it helped him find mistakes he could not find on his own. According to Alex, Grammarly also "tells a little bit about the reason behind the error," and it is easy to operate. Alex said he would consider using Grammarly in the future for the following reason: "In the whole perspective, it is useful. At least, I can have more resources to proofread my paper" (the interview).

Kelsey - Low-intermediate L2 Writer

Kelsey's Behavioral Engagement with AWCF

Table 8 illustrates Kelsey's revision operations in response to AWCF. The table shows Kelsey made 26 errors in her 810-word draft for which she received AWCF. Ten error types were identified in her draft, including errors on the use of word choice (1), verb form (1), word form (3), articles (8), pronouns (1), punctuation (1), spelling (2), subject-verb agreement (4), prepositions (2), and spacing (3). Like in Alex's case, the most frequent errors identified by Grammarly were articles. Of 26 received AWCF of which 18 were accurate and eight were inaccurate, Kelsey correctly accepted 15, incorrectly rejected two, incorrectly substituted one, incorrectly accepted seven, and correctly rejected one; thus also fixing 57% of her total errors. This suggests Kelsey made moderate changes to her draft like Alex.

Table 8

Kelsey's Revision Operations

Error type	**AWCF**	**Accurate AWCF**			**Inaccurate AWCF**		
		Accept	**Reject**	**Substitute**	**Accept**	**Reject**	**Substitute**
Word choice	**1**		1				
Verb form	**1**	1					
Word form	**3**		1		1	1	
Articles	**8**	3		1	4		
Pronouns	**1**				1		
Punctuation	**1**	1					
Spelling	**2**	2					
Subject-verb agreement	**4**	3			1		
Prepositions	**2**	2					
Miscellaneous*	**3**	3					
Total number of errors in an 810-word text	**26**	15	2	1	7	1	
Percentage	**100%**	**57%**	**8%**	**4%**	**27%**	**4%**	

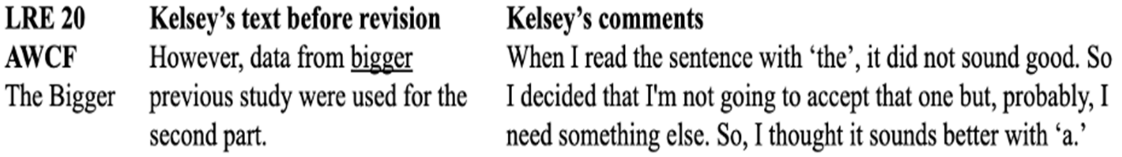
* Miscellaneous here refers to a spacing error.

Unlike Alex, Kelsey did not refer to any external resources to enhance the accuracy of her draft. Similar to Alex, Kelsey also spent a little over five minutes (323 seconds) on revision, despite having more errors in her draft than Alex.

Kelsey's Cognitive Engagement with AWCF

Although Kelsey managed to detect all AWCF and recognize their corrective intention by just looking at Grammarly corrections as she reported in the recall, she did not always understand the cause/nature of the majority of errors, especially errors on article usage. To illustrate, Grammarly suggested Kelsey deleting *the* before *gender* in the sentence: *The results were analyzed according to the gender*. Kelsey uncritically accepted Grammarly's accurate feedback. In the recall, she was unable to verbalize the underlying rule. Additionally, she admitted articles were her weakness and stated that when it comes to articles, “It's probably that I'm wrong, and Grammarly’s right.” It is not surprising Kelsey accepted all accurate and inaccurate AWCF on articles except one. She appeared to substantially rely on AWCF on articles due to the lack of control over that form.

Interestingly, when Grammarly suggested Kelsey adding the definite article, she incorrectly substituted it with the indefinite article. The fact that Kelsey neither accepted the feedback nor rejected it could indicate she had some doubts about AWCF, but still trusted enough to consider Grammarly correction. Thus, prompted by AWCF and drawing on her intuition, she substituted the suggested definite article with indefinite:



LRE 20	**Kelsey’s text before revision**	**Kelsey’s comments**
AWCF The Bigger	However, data from bigger previous study were used for the second part.	When I read the sentence with ‘the’, it did not sound good. So I decided that I'm not going to accept that one but, probably, I need something else. So, I thought it sounds better with ‘a.’

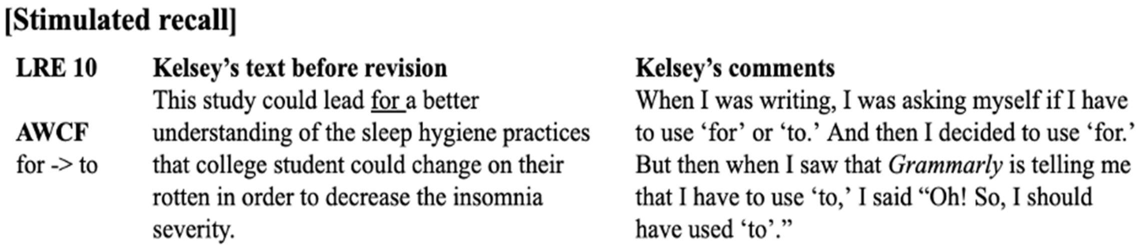
Unlike Alex, Kelsey had limited cognitive engagement with AWCF which was displayed by the infrequent use of metacognitive and cognitive operations. Like Alex, the metacognitive operation Kelsey deployed at the very beginning of the revision process was planning for cognition. The screencast showed how Kelsey scrolled down to see “if there are a lot of mistakes” (the recall). By looking at how much feedback she received, it seems that Kelsey wanted to be mentally prepared for revision. Kelsey also double-checked the revision operation she used to ensure her decision was right; however, that happened only once. Regarding the use of cognitive operations, Kelsey also used context to determine the appropriate revision operation. However, unlike Alex, Kelsey did not read Grammar*ly* suggestions nor she expanded them for a metalinguistic explanation, which indicates she did not take full advantage of Grammarly feedback to effectively respond to it.

Kelsey's minimal cognitive engagement was seen especially toward the end of the revision process as she uncritically started accepting AWCF. In the interview, she admitted: "At the beginning, I was thinking about every mistake. I used to read the sentence from the beginning, think about it a little bit but at the end, I just accepted."

Kelsey's Affective Engagement with AWCF

Kelsey's emotional reaction upon receiving AWCF was surprise. She did not expect to have 27 errors in her final draft because, according to her, she addressed low-order concerns before submitting her draft. Thus when Kelsey saw how much AWCF she received, she immediately questioned feedback. In the recall, she said, "I was thinking if all of these are really mistakes or some of them are just the computer thing that are mistakes but they're not."

Once she started correcting errors, she experienced both positive and negative emotional reactions toward AWCF. For example, Kelsey liked AWCF when it confirmed her earlier doubts:



[Stimulated recall]

LRE 10	**Kelsey's text before revision**	**Kelsey's comments**
AWCF for -> to	This study could lead <u>for</u> a better understanding of the sleep hygiene practices that college student could change on their rotten in order to decrease the insomnia severity.	When I was writing, I was asking myself if I have to use 'for' or 'to.' And then I decided to use 'for.' But then when I saw that *Grammarly* is telling me that I have to use 'to,' I said "Oh! So, I should have used 'to'."

Kelsey did not like AWCF when she recognized it was inaccurate. In the interview, she said, "If I know that the tool is wrong, I will not accept the feedback." She also added, “If I am not sure, I will just accept feedback.” This demonstrates Kelsey’s dependence on AWCF when she was unsure about how to correct errors.

Regarding Kelsey’s attitudinal response toward AWCF, she found it to be useful. In the interview, she said she would use Grammarly in the future because it helped her correct some of her errors. She also noted, “[...] we all know technology sometimes makes mistakes. So, we should think about it before we make a decision.” This statement contradicts Kelsey’s actions. She hardly ever critically thought about AWCF as she felt the need to eliminate all errors because, as she stated in the interview, “[she doesn’t] want a teacher to look at [her] work and then immediately tell that [she’s] international.”

Discussion and Conclusion

Informed by the conceptual framework of student engagement with WCF and AWE feedback (Han & Hyland, 2015; Zhang, 2017; Zhang & Hyland, 2018; Zheng & Yu, 2018), this study focused on how two ESL college students behaviorally, cognitively, and affectively engaged with Grammarly feedback *(*AWCF) when revising their final draft. The findings offer insight into the complex process of student engagement with AWCF and provide implications for the use of automated tools for writing assessment in L2 classrooms.

The two students’ behavioral engagement with AWCF involved revision operations, revision strategies, and time spent on revision. Both students focused on eliminating Grammarly-detected errors and corrected 57% of their total errors, thus making moderate changes to their draft. This suggests the students did not effectively utilize Grammarly feedback to revise their final draft. This echoes previous studies’ findings that students tend not to make good use of automated feedback (Attali, 2004;

Chapelle et al., 2015; Warschauer & Grimes, 2008). Additionally, the two students barely used revision strategies to refine their draft, which indicates their behaviors remained at the surface level. Alex (advanced L2 writer) consulted the Internet once to verify the accuracy of AWCF. He primarily drew on his linguistic knowledge and intuition to correct errors. Because of his lack of knowledge about certain grammar and mechanics rules, however, he rejected accurate feedback. If Alex had used external resources or expanded Grammarly suggestion for a metalinguistic explanation, his behavioral engagement could have resulted in greater accuracy of the draft. Unlike Alex, Kelsey (low-intermediate L2 writer) did not utilize any external resources to enhance the accuracy of her draft nor did she read Grammarly suggestions and grammar rules. She appeared to substantially rely on AWCF and rarely thought critically about feedback. Nevertheless, this still resulted in moderate changes to her draft. This suggests that without cognitive engagement as in a multiple-choice examination where a student can receive a score by randomly guessing the answer, behavioral engagement with AWCF alone can result in successful revisions if accurate feedback is accepted. However, mere behavioral engagement cannot lead to true learning. Regarding time spent on revision, both students spent just over five minutes. This is in line with Warden (2000), who found that L2 students spent an average of six minutes on draft revision after AWE feedback which was attributable to fewer errors. This could be the case in this study too. This could also be due to limited cognitive engagement with AWCF or automatic application of Grammarly correction with a single click that could have sped up the revision process. It is noteworthy, however, that little revision time does not necessarily indicate minimal cognitive engagement with AWCF. Although both students spent five minutes revising, Alex had a more extensive cognitive engagement with AWCF than Kelsey despite having fewer errors than her.

From the cognitive perspective, which involved how deeply students processed AWCF (noticing and understanding) and their use of metacognitive and cognitive operations, both students noticed AWCF and recognized its corrective intention due to its implicit and explicit nature. However, unlike Alex who understood the cause/nature of the majority of errors and how to correct those errors, Kelsey had little awareness at the level of understanding. Her insufficient linguistic competence appeared to impact her ability to effectively process feedback and determine appropriate revisions. Similar findings have also been uncovered in Zheng and Yu (2018). Additionally, unlike Alex, Kelsey used fewer metacognitive and cognitive operations, which indicates her minimal cognitive engagement with AWCF. This consequently resulted in an overreliance on feedback. Conversely, Alex had an extensive cognitive engagement with AWCF which was manifested in the frequent use of metacognitive and cognitive operations. As a result, he could make independent judgments and selective incorporation of AWCF.

Regarding affective engagement with AWCF, which involved students' emotional reactions and attitudinal responses to feedback, both students experienced different emotional reactions. Alex questioned each AWCF he received but did little to verify its accuracy. He relied on his L2 knowledge and intuition to determine appropriate revisions which led to rejecting accurate AWCF. This suggests that questioning feedback is not enough; it is what happens after that. In contrast, Kelsey seemed to excessively depend on AWCF, especially when she lacked knowledge about target forms. Overall, however, the two students had a positive attitude toward Grammarly feedback which corresponds with findings of previous research that students generally tend to appreciate automated feedback (Dikli & Bleyle, 2014; Li et al., 2015; Li et al., 2017).

This study could extend previous research on student engagement with automated feedback in the following ways. Compared to Zhang (2017) who claimed automated feedback can have a positive impact on student writing if active behavioral, cognitive, and affective engagement are in place, the findings of this study revealed behavioral engagement with AWCF alone could potentially lead to successful revisions if accurate AWCF is accepted. However, behavioral engagement alone cannot lead to true learning. Additionally, the findings showed that while one participant's negative affective engagement with AWCF (questioning) positively impacted his cognitive engagement, the other's positive affective engagement with AWCF (trust) resulted in limited cognitive engagement. Students' language proficiency could be the cause of this. As Zheng and Yu (2018) claimed, limited linguistic knowledge could prevent students from fully processing feedback and making further revisions. Thus students may exhibit overreliance on automated feedback. The findings of this study also suggest that students with higher language proficiency are likely to question AWCF, spend more time processing it, and make selective incorporation of it.

The findings can provide several implications to enhance student engagement with automated feedback and better utilize Grammarly for assessment purposes in L2 writing classrooms. Based on the study's findings, students with low language proficiency may not be able to utilize Grammarly effectively as their lack of linguistic competence can prevent them from adequately understanding AWCF. Therefore, the use of Grammarly is recommended for students with more advanced English proficiency. To benefit from Grammarly feedback, students should receive proper guidance and training on how to effectively engage with it. First, for successful affective engagement with AWCF, students should be made aware of its strengths and weaknesses. Grammarly feedback contains a suggestion that could be expanded for a metalinguistic explanation that students should be encouraged to read to help them make appropriate revisions. Students should also be informed about the inaccuracies of AWCF to avoid excessive dependence on it. Second, for effective cognitive engagement with AWCF students should be guided to question and analyze it critically. Finally, for productive behavioral engagement with AWCF, students should be advised to confirm feedback with other sources or perhaps with other students or a teacher.

Grammarly and similar automated programs could serve as a useful resource for writing assessment in L2 classrooms if active engagement is in place. Teachers could incorporate them into the writing curriculum as a supplemental tool to facilitate low-order concerns of student writing development. To enhance L2 writers' independent and critical thinking, students' reflections on their use of automated feedback could become a writing assignment. This could prove useful for helping both students utilize automated feedback for self-assessment of their own writing more effectively and teachers estimate what is already working well and what still needs to be improved in terms of developing students' writing skills.

Some limitations in this study should be acknowledged. Caution should be made when generalizing the findings to different contexts and a wider student population as they were based on two students' engagement with AWCF. Additionally, only one draft was analyzed in the study, so development or changes in student engagement with

AWCF over time were not investigated. Future research may also consider including survey questions addressing students' self-confidence as this may also affect their engagement with AWCF.

CHAPTER V

CONCLUSION

The dissertation aimed at investigating the AWE system, Grammarly, from both system- and user-centric perspectives to find out its L2 writing pedagogical potentials. The results of the user-centric research that focused on Grammarly's error-detection/correction performance (Chapter II) revealed that the tool was highly accurate in detecting and correcting L2 errors; however, unlike human annotators, it skipped half of the L2 errors in the corpus. The findings of the first user-centric study that focused on six postsecondary, L2 writing teachers' use and perceptions of Grammarly as a complement to their feedback (Chapter III) revealed that despite using Grammarly to complement their feedback, the teachers provided feedback both on HOCs and LOCs and that there was no division of labor such as that Grammarly takes care of lower-order concerns, even though it is computationally adept at providing such feedback, and teachers take care of higher-order concerns. The study also revealed three factors that might have impacted teachers' feedback when they used Grammarly as a complement which were their use of Grammarly reports, perceptions of automated feedback, and beliefs about feedback and L2 writing course. Overall, despite Grammarly's limitations, four out of six teachers were positive about Grammarly and expressed their wish to use

Grammarly in their L2 writing classrooms. The findings of the second user-centric study that focused on two ESL university writing students' behavioral, cognitive, and affective engagement with Grammarly feedback (Chapter IV) demonstrated that the students had different levels of engagement with Grammarly's feedback. One student had greater cognitive engagement, which was manifested in his questioning of Grammarly's feedback. However, he did not verify his doubts by consulting sources and relied primarily on his intuition, which resulted in moderate changes to his draft. The other student had low cognitive engagement, which was manifested in her blindly accepting Grammarly feedback, thus showing her overreliance on the tool. Nonetheless, this also resulted in moderate changes to her draft. Generally, both students were favorable of Grammarly as an additional source of feedback.

The overall findings of this dissertation indicate that Grammarly has the potential to be used in L2 writing classrooms. Grammarly has satisfactory performance in terms of the accuracy of its feedback, which is in line with previous research (e.g., Ranalli and Yamashita, 2022; Koltovskaia, 2022). Grammarly is also perceived positively by teachers and students, despite having its limitations. This dissertation answers the question of "how this new technology can be used to achieve more desirable learning outcomes while avoiding potential harms that may result from limitations inherent in the technology" (Chen & Cheng, 2008, p. 95).

The following are general pedagogical recommendations for teachers to meaningfully use Grammarly in their L2 writing classrooms. When it comes to Grammarly's performance, teachers should inform their L2 learners that Grammarly may skip some of the errors committed by L2 learners as it is not designed with non-native speakers of English in mind. This may encourage students to look for the errors on their own. Teachers should also inform students that Grammarly’s feedback can be occasionally inaccurate. That is, it may flag non-errors and/or provide inaccurate correction for the error. Knowing this will enable students to be more critical about Grammarly’s feedback instead of looking at the tool as authority and relying on it.

When using Grammarly as a complement to their feedback, teachers should not offload all sentence-level errors to Grammarly because of its poor performance in detecting all L2 errors. To compensate for this limitation, teachers should provide their own feedback on local aspects of writing if necessary. Since Grammarly’s feedback on lower-order concerns focuses on treatable errors (e.g., subject-verb agreement, noun number) that are easy to fix because they are rule-governed, teachers could provide feedback on untreatable errors (e.g., sentence structure, word form, and word choice) as these errors are often harder for students to catch and repair as they cannot be explained by simple rules and require a nuanced understanding of the error. If teachers decide to provide feedback on untreatable errors, written corrective feedback research suggests that such feedback should be direct (e.g., Ferris, 2014); that is, a suggestion for how to correct the error should be given. Additionally, because Grammarly’s feedback is comprehensive and it can detect a lot of errors in student writing, such feedback may overwhelm and discourage students from revising their texts as students. To avoid such negative impact on students, teachers could use Grammarly to get an overview of what errors individual students and everyone else in class struggle with and provide focused feedback along with class activities on the most frequent errors.

Research on AWE has emphasized the importance of student training, particularly with respect to responding to automated feedback (e.g., Barrot, 2021; Link et al., 2014). In line with previous research, this dissertation also highlights a need for student training in using Grammarly and responding to its feedback. Since Grammarly automatically fixes an error with a single click, it is necessary to inform students to take time to process feedback. Students should be aware that blind acceptance of Grammarly's feedback may not lead to true learning and may even result in more errors in writing if Grammarly's feedback is inaccurate (Koltovskaia, 2020). Additionally, Grammarly should be recommended for students with higher-level of English proficiency as they may be able to understand computer-generated feedback better than students with a lack of linguistic competence. Finally, before implementing Grammarly into their L2 writing classrooms, teachers should consider L2 writing course objectives, students' needs, and their beliefs about written correction feedback. If teachers decide to introduce Grammarly to their students, Grammarly should not be the only source of feedback. Although AWE systems have been created to save teachers' time and ease the process of feedback provision, they cannot substitute a teacher (Chen & Cheng, 2008).

Apart from pedagogical implications, the dissertation also provides a theoretical implication. By using a theoretical framework of learner engagement to explore college students' engagement with automated feedback provided by Grammarly discussed in Chapter IV, the dissertation showed the benefits of grounding a study on AWE in a well-justified theoretical framework as this provides more rigor to the research. The study discussed in Chapter IV revealed a complex process of students' behavioral, cognitive, and affective engagement with automated feedback in ESL context, thus adding a missing piece in the literature that predominantly focused on students' engagement with automated feedback in EFL context. The study also showed that behavioral engagement with automated feedback alone can lead to positive impact on writing if accurate feedback is accepted, while previous research argued that for positive impact on writing, cognitive, affective, and behavioral engagement should all be in place (Zhang, 2017).

Even though this dissertation evaluated Grammarly from both system- and user-centric perspectives, thus extending our understanding of the tool, it also opens up possible springboards for further research. There should be more system-centric research that focuses on Grammarly's error-detection/correction performance using the bigger corpus of texts written by ESL learners to be able to see its performance potential and generalize the results. Additionally, the system centric-research on Grammarly should focus not only on the accuracy and overall coverage of its automated formative feedback but also on the scoring ability of its feedback. Grammarly's overall performance score i.e. a holistic score that Grammarly automatically generates has also the potential to be used in L2 writing classrooms, for example, for diagnostic purposes, such as identifying students' writing weaknesses and initiating individualized development plans for students, which warrants more research. More user-centric studies are also needed that focus on teachers' and students' use of Grammarly in authentic contexts. Future studies on teachers' use of Grammarly to complement their feedback and students' engagement with Grammarly's feedback should be longitudinal and classroom-based as such research provides a better understanding of how to implement the tool more effectively. Additionally, since the bulk of user-centric research on AWE is often product-oriented and often lacks theoretical foundation, future studies should be informed by a theoretical framework, including learner engagement, as such research provides a more complete and comprehensive picture of the phenomenon under investigation, thus moving forward the filed.

Despite their technological limitations, AWE systems are here to stay. If they are here to stay and to be used for instructional purposes, it is imperative to minimize unwanted learning outcomes. By showing the performance of Grammarly and teachers' and students' use and perceptions of it, the dissertation hopes to increase L2 writing teachers' awareness of not only the limitations of AWE systems but also their strengths. This knowledge could help teachers make informed decisions about the implementation of AWE systems in their L2 writing classrooms.

REFERENCES

Anson, C. M. (2006). Can't touch this: Reflections on the servitude of computers as readers. In. P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 38-56). Utah State University Press. https://doi.org/10.2307/j.ctt4cgq0p

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

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

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology, 37(1), 67-81.*

Barrot, J. S. (2021). Using automated written corrective feedback in the writing classrooms: effects on L2 writing accuracy. *Computer Assisted Language Learning*, 1–24. https://doi.org/10.1080/09588221.2021.1936071

Bestgen, Y., & Granger, S. (2011). Categorizing spelling errors to assess L2 writing. *International journal of continuing engineering education and life-long learning, 21*, 235-252.

.

Brent, E., & Townsend, M. (2006). Automated essay grading in the sociology classroom: Finding common ground. In. P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 177-198). Utah State University Press. https://doi.org/10.2307/j.ctt4cgq0p

Burstein, J., & Chodorow, M. (1999). *Automated Essay Scoring for nonnative English speakers*. Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, College Park, MD.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essays evaluation: The Criterion online writing service. *AI Magazine, 25*(3), 27-36.

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

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using. automated writing evaluation. *Language Testing, 32*(3), 385-405.

Chen, C., & Cheng, W. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12*(2), 94-112.

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

Chodorow, M., Tetreault, J. R., & Han, N. (2007). Detection of grammatical errors involving prepositions. In F. Costello, J. Kelleher, & M. Volk (Eds.), *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions* (pp. 25–30). Prague: Association for Computational Linguistics.

Condon, W. (2006). Why less is not more: What we lose by letting a computer score writing samples. In. P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 211-220). Utah State University Press. https://doi.org/10.2307/j.ctt4cgq0p

Corbin, J., & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory (3rd ed.)*. Los Angeles, Calif: Sage Publications.

Cotos, E. (2018). Automated Writing Evaluation. In J. I. Liontas & M. DelliCarpini (Eds.), *The TESOL encyclopedia of English language teaching*. Wiley-Blackwell.

Cotos, E. (2014). *Genre-based automated writing evaluation for L2 research writing: From design to evaluation and enhancement.* New York, NY: Palgrave Macmillan.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks, California: SAGE Publications.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment, 5*(1). 1- 35.

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.

El-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, S. (2002). Intellimetric^{TM}: From here to validity. In M. D. Shermis, & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Ellis, R. (2010). Epilogue: A framework for investigating oral and written corrective feedback. *Studies in Second Language Acquisition, 32*, 335-349.

Ene, E., & Upton, T. A. (2014). Learner uptake of teacher electronic feedback in ESL composition. *System, 46*, 80-95.

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

Ericsson, P. F. (2006). The meaning of meaning: Is a paragraph more than an equation? In. P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 28-37). Utah State University Press. https://doi.org/10.2307/j.ctt4cgq0p

Ericsson, P. F., & Haswell R.H. (2006). *Machine scoring of student essays: Truth and consequences*. Utah State University Press.

ETS Criterion. (2019). *TOEFL iBT Test independent writing rubrics.* https://www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf

ETS Criterion. (2021). *Performance Descriptors for the TOEFL iBT Test.* *https://www.ets.org/s/toefl/pdf/pd-toefl-ibt.pdf*

ETS Criterion. (2022a, February 25). *Sign In to Your Criterion® Account.* https://criterion.ets.org/criterion/default.aspx

ETS Criterion. (2022b, April 4). *Converting rubric scores to scaled scores: Writing and* *speaking sections of the new TOEFL iBT test*. http://www.etweb.fju.edu.tw/elite/ETS%20-%20ibt%20TOEFL%20Converting_Rubric.pdf

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). (2nd ed.). New York, NY: Cambridge University Press

Ferris, D. R. (2014). *Treatment of errors in second language student writing (2nd ed.).* Ann Arbor: The University of Michigan Press.

Fredricks, J. A., Blumenfield, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59-109.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Grammarly. (2022, January 2). *Choose your plan*. https://www.grammarly.com/upgrade?utm_campaign=funnelPremiumAboveTheFold&utm_medium=internal&utm_source=funnel

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), 1-43.

Guo, G., Feng, R., & Hua, Y. (2021). How effectively can EFL students use automated written corrective feedback (AWCF) in research writing? *Computer Assisted Language Learning*, 1–20.

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

Han, Y., & Hyland, F. (2015). Exploring learner engagement with written corrective feedback in Chinese tertiary EFL classroom. *Journal of Second Language Writing, 30*, 31-44.

Herrington, A., & Moran, C. (2006). WritePlacer Plus in place: An exploratory case study. In. P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 114-129). Utah State University Press. https://doi.org/10.2307/j.ctt4cgq0p

Hyland, & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching*, *39*(2), 83–101. https://doi.org/10.1017/S0261444806003399

Jiang, Y., Yu, S., & Wang, C. (2020). Second language writing instructors' feedback practice in response to automated writing evaluation: A sociocultural perspective. *System, 93*, 1-11.

Jones, E. (2006). ACCUPLACER's essay-scoring technology: When reliability does not equal validity. In. P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 93-113). Utah State University Press. https://doi.org/10.2307/j.ctt4cgq0p

Koltovskaia, S. (2022, March 22). *Grammarly's error-detection/correction performance and L2 writing teachers' use of it to complement their formative feedback.* American Association for Applied Linguistics (AAAL), Pittsburgh, PA. https://www.xcdsystem.com/aaal/program/HZ57qut/index.cfm?pgid=145&SearchTerm=Koltovskaia

Koltovskaia, S.(2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing, 44*, 100450–. https://doi.org/10.1016/j.asw.2020.100450

Lai, Y. (2010). Which do students prefer to evaluate their essays: Peers or computer program. British *Journal of Educational Technology, 41*(3), 432–454.

Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology, 19*, 50–68.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. *Synthesis Lectures on Human Language Technologies, 3*, 1-134.

Lee, I. (2019). Teacher written corrective feedback: Less is more. *Language Teaching, 52*(4), 524-536.

Leko, M. M. (2014). The value of qualitative methods in social validity research. *Remedial and Special Education, 35*(5), 275–286. https://doi.org/10.1177/0741932514524002

Li, Z. (2021). Teachers in automated writing evaluation (AWE) system-supported ESL writing classes: Perception, implementation, and influence. *System*, *99*, 1-14. https://doi.org/10.1016/j.system.2021.102505

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.

Li, Z., Feng, H., & Saricaoglu, A. (2017). The short-term and long-term effects of AWE feedback on ESL students' development of grammatical accuracy. *CALICO*

Journal, 34(3), 355-375.

Liao, H. (2015). Using automated writing evaluation to reduce grammar errors in writing. *ELT Journal, 70*(3), 308-319.

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

Link, S., Mehrzad, M., & Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2020.1743323

Liu, S., & Kunnan, A. J. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of WriteToLearn. *CALICO Journal*, *33*(1), 71–91.

Maeng, U. (2010). The effect and teachers' perception of using an automated essay scoring system in L2 writing. *English Language and Linguistics, 16*(1), 247-245.

McGee, T. (2006). Taking a Spin on the Intelligent Essay Assessor. In. P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 79-92). Utah State University Press. https://doi.org/10.2307/j.ctt4cgq0p

O'Neill, R., & Russell, A. M. T. (2019). Stop! Grammar time: University students' perceptions of the automated feedback program Grammarly. *Australasian Journal of Educational Technology, 35*(1), 42-56.

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research, 26*(4), 407-425.

Quinlan, T., Higgins, D., & Wolff, S. (2009). Evaluating the construct-coverage of the e-rater scoring engine. *ETS Research Report Series, 1*, i–35.

Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning, 31*(7), 653-674.

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology, 37*(1), 8-25.

Ranalli, J., & Yamashita, T. (2022). Automated written corrective feedback: Error-correction performance and timing of delivery. *Language Learning & Technology*, *26*(1).

Rothermel, B. A. (2006). Automated writing instruction: Computer-assisted or computer-driven pedagogies? In. P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 199-210). Utah State University Press. https://doi.org/10.2307/j.ctt4cgq0p

Sheen, Y., Wright, D., & Moldawa, A. (2009). Differential effects of focused and unfocused written correction on the accurate use of grammatical forms by adult ESL learners. *System, 37*(4), 556-569.

Stevenson, M. (2016). A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition, 422,* 1-16.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19,* 51-65.

Stevenson, M., & Phakiti. A. (2019). Automated feedback and second language writing. In K. Hyland. *Electronic resources for feedback* (pp. 125-142).

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics, 16*(3), 371-391.

Tetreault, J., & Chodorow, M. (2008). The ups and downs of preposition errors detection in ESL writing. Proceedings of the 22nd International Conference on Computational Linguistics (COLING), Manchester, UK.

Thi, N. K., & Nikolov, M. (2021). How teacher and Grammarly feedback complement one another in Myanmar EFL students' writing. The Asia-Pacific Education Researcher. doi:10.1007/s40299-021-00625-2

Vantage Learning. (2003a). Assessing the accuracy of Intellimetric for scoring a district-wide writing assessment (RB-806). Newton, PA: Vantage Learning.

Vantage Learning. (2003b). How does Intellimetric score essay response? (RB-929). Newton, PA: Vantage Learning.

Vantage Learning. (2004, September 14). *MY Access!™ Online Writing Practice with Immediate Diagnostic Assessment, Constructive Feedback and Instructional Assessment*. https://www.vantage.com/pdfs/myaccess.pdf

Vantage Learning. (2006). Research summary: Intellimetric scoring accuracy across genres and grade levels. Retrieved from http://www.vantagelearning.com/docs/intellimetric/IM_ReseachSummary_InteliMetric_Accuracy_Across_Genre_and_Grade_Levels.pdf

Wang, Y., Shang, H., & 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), 1-24.

Warden, C. A. (2000). EFL business writing behaviors in differing feedback environments. *Language Learning 50*(4), 573-616.

Ware, P. (2011). Computer-generated feedback on student writing. *TESOL Quarterly, 45*(4), 769–774. doi:10.5054/tq.2011.272525

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*(1), 22-36.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 157-180.

Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.). *Handbook of automated essay evaluation: Current applications and new directions* (pp. 36-54). Routledge.

Whithaus, C. (2006). Always already: Automated essay scoring and grammar-checkers in college writing courses. In. P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 166-176). Utah State University Press. https://doi.org/10.2307/j.ctt4cgq0p

Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers and Education*, *168*, 1-11.

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education, 100*, 94-109. https://doi.org/10.1016/j.compedu.2016.05.004

Yin, R. K. (2009). *Case study research: Design and methods* (2 ed.). London: SAGE Publications.

Zhang, Z. (2017). Student Engagement with computer-generated feedback: A case study. *ELT Journal, 71*(3), 317-328.

Zhang, Z., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing, 36*, 90-102.

Zheng, Y., & Yu, S. (2018). Student engagement with teacher written corrective feedback in EFL writing: A case study of Chinese lower-proficiency students. *Assessing Writing, 37*, 13-24.

Qassemzadeh, A., & Soleimani, H. (2016). The impact of feedback provision by Grammarly software and teachers on learning passive structures by Iranian EFL learners. *Theory and Practice in Language Studies, 6*(9), 1884-1894.

APPENDICES

Appendix A. Argumentative Essay Prompt.

Due Dates	
Rough Draft	
Final Draft	

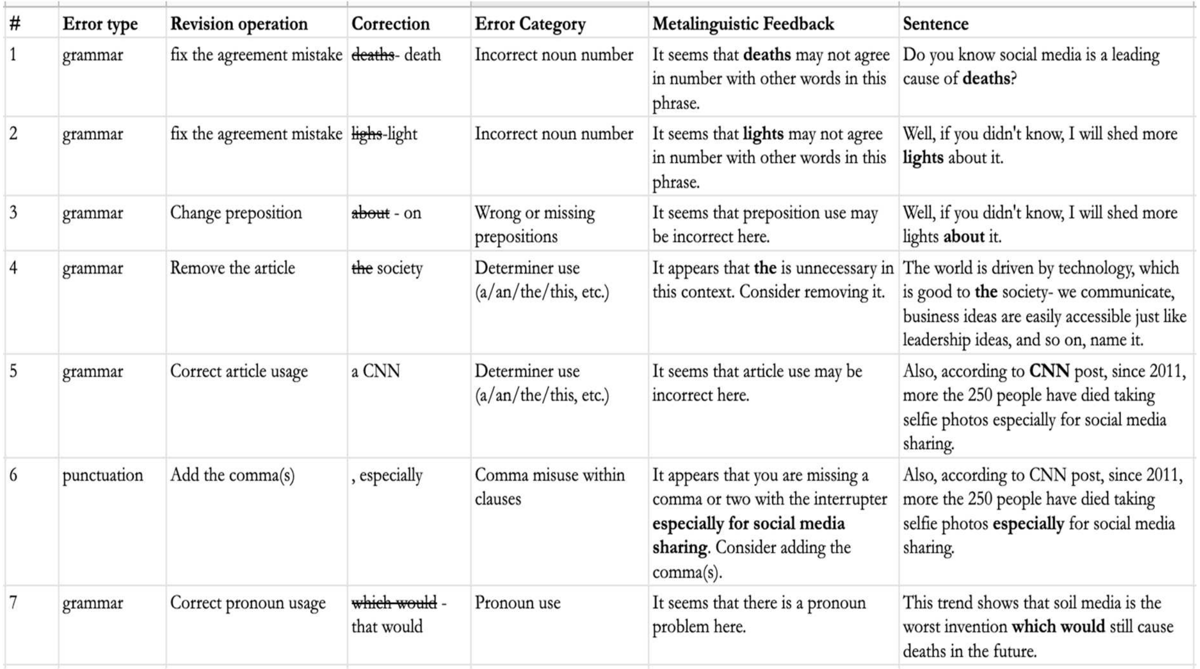
Topic:

New technologies are being invented and refined constantly in the world we live in today. While many of these technologies were created for the good of society, some have impacted the world in negative ways. *Discover Magazine* is creating a special issue that reflects on the changes in technology over the last 100 years. Write an article for this special issue **arguing** one invention the world would be better **without**.

Instructions:
Please follow the following guidelines to craft your essay:

1. Craft your own argument using sources as support.
2. No matter which topic you choose, you should consider **2-4 main points** for your argument.
3. Address **at least one counter-argument**.
4. Your essay must make use of at least **3 sources**, which you must cite using **in-text citation** and a **reference page** in **APA style**. Do not use Wikipedia or a source that is not in English as one of your 3 main sources. You may use websites from Google or journals from our library databases.
5. ANY outside information, either words or ideas, you MUST cite according to APA format. If you have any questions about this, please contact me or visit the Writing Center.
6. Your essay should be properly formatted according to APA guidelines, proofread, and **at least 1000 words in length**.

Appendix B. Snapshot of Google Spreadsheet Tab for Essay _001.



#	Error type	Revision operation	Correction	Error Category	Metalinguistic Feedback	Sentence
1	grammar	fix the agreement mistake	~~deaths~~- death	Incorrect noun number	It seems that **deaths** may not agree in number with other words in this phrase.	Do you know social media is a leading cause of **deaths**?
2	grammar	fix the agreement mistake	~~lighs~~-light	Incorrect noun number	It seems that **lights** may not agree in number with other words in this phrase.	Well, if you didn't know, I will shed more **lights** about it.
3	grammar	Change preposition	~~about~~ - on	Wrong or missing prepositions	It seems that preposition use may be incorrect here.	Well, if you didn't know, I will shed more lights **about** it.
4	grammar	Remove the article	~~the~~ society	Determiner use (a/an/the/this, etc.)	It appears that **the** is unnecessary in this context. Consider removing it.	The world is driven by technology, which is good to **the** society- we communicate, business ideas are easily accessible just like leadership ideas, and so on, name it.
5	grammar	Correct article usage	a CNN	Determiner use (a/an/the/this, etc.)	It seems that article use may be incorrect here.	Also, according to **CNN** post, since 2011, more the 250 people have died taking selfie photos especially for social media sharing.
6	punctuation	Add the comma(s)	, especially	Comma misuse within clauses	It appears that you are missing a comma or two with the interrupter **especially for social media sharing**. Consider adding the comma(s).	Also, according to CNN post, since 2011, more the 250 people have died taking selfie photos **especially** for social media sharing.
7	grammar	Correct pronoun usage	~~which would~~ - that would	Pronoun use	It seems that there is a pronoun problem here.	This trend shows that soil media is the worst invention **which would** still cause deaths in the future.

Appendix C. The Error Categorization Rubric Based on Grammarly.

Error type	Error category	Metalinguistic Feedback	Example from Data
	Conjunction use (Add the word(s)/Remove the conjunction)	The sentence is missing the part of the **not only ... but also** construction. Consider adding the missing word(s).	Still not only such, **the digital components can** be recycled to make a new one.
	Determiner use (a/an/the/this, etc.) (Add an article/Change the article/Change the determiner/Correct article usage/Correct determiner usage/Correct the article-noun agreement/Remove the article/Remove the determiner)	The noun phrase **classroom** seems to be missing a determiner before it. Consider adding an article.	Having technology in **classroom** not only help students in learning but it also help the learners to retain their skills in different disciplines as compared to the traditional classroom way of teaching and learning.

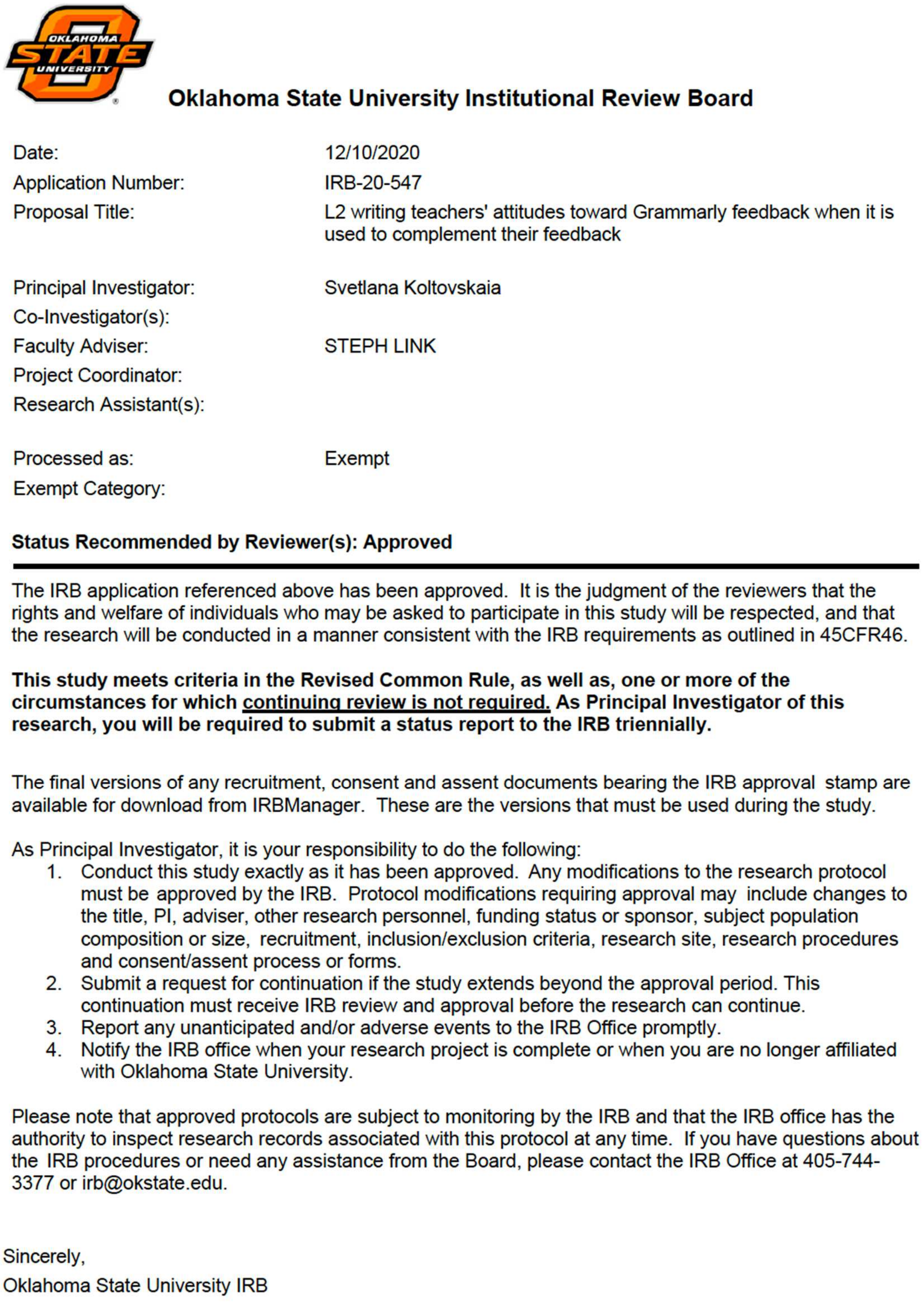
Grammar	**Incorrect noun number** (Change noun form/Change the noun form/Change to a genitive case/Change to a plural noun/Fix apostrophe usage/Fix the agreement mistake)	It seems that this noun form may be incorrect.	Another issue about **kids** mental health, the kids will take everything seriously they cannot handle a joke because of their anger.
	Faulty subject-verb agreement (Change the verb form/Correct subject-verb agreement)	The plural verb **play** does not appear to agree with the singular subject **a kid**. Consider changing the verb form for subject-verb agreement.	When a kid **play** with the PlayStation it will affect his mental and physical health in many ways.
	Incorrect verb forms (Add a missing verb/Add the auxiliary verb/Add the particle/Change the form of the verb/Change the verb form/Delete extra word/Fix the infinitive/Rewrite the sentence)	It appears that your sentence or clause uses an incorrect form of the verb **be.** Consider changing it.	Today, the animation industry has much more matured, and more and more new technologies are **be** used in the animation industry.
	Modal verbs (Change the verb form)	The word **to** is usually unnecessary after the modal verb **will.** Consider removing it.	In the rest of the essay, I will **to** tell you why reading digital text is better than reading printed text.
	Misuse of modifiers (Change to singular/Change the adjective/Change the wording/Change the adverb/Change the quantifier/Change the verb form/Change the	It appears that the number **billions** is modifying a noun and should be in the singular form. Consider changing it.	In fact, two **billions** plastic straws are being consumed each year in the world.

	word/Replace the words)		
	Misuse of quantifiers (Correct quantifier usage/Replace the quantifier)	It appears that the quantifier **many** does not fit the uncountable noun **software.** Consider changing it.	Additionally, there are **many** good software that have been developed and are used to supplement the class curriculum.
	Pronoun use (Add a pronoun/Change the pronoun/Correct pronoun usage/Delete the pronoun)	This sentence appears to be missing a pronoun. Consider adding the pronoun.	Video games are great to entertain ourselves in **free** time.
	Wrong or missing prepositions (Add the preposition/Change preposition/Change the preposition/Replace the preposition/Replace the word/Verify preposition usage)	It seems that preposition use may be incorrect here.	Well, if you didn't know, I will shed more lights **about** it.
Punctuation	**Closing punctuation** (Change the punctuation/ Replace the punctuation)	Consider using a question mark to signify that this sentence is a question.	If an A.I. makes a wrong assumption and does something wrong how can we be sure that it won't happen **again.**
	Comma misuse within clauses (Add a comma/Add the comma(s)/Remove the comma)	Your sentence contains a series of three or more words, phrases, or clauses. Consider inserting a comma to separate the elements.	Printed books are completed by itself meaning that there is no dependency on any electronic device including internet connection,

			internet network **and** service..
	Misuse of semicolons, quotation marks, etc. (Remove the colon)	It appears that the colon in your sentence is unnecessary. Consider removing it.	It was later upgraded to aid business transactions and management. Popular social networks **include:** WhatsApp, Snapchat, Twitter, and the most widely used Facebook.
	Punctuation in compound/complex sentences (Remove the comma)	It appears that you have an unnecessary comma before the dependent clause marker **because**. Consider removing the comma.	First, I think **printed** version is better than digital **text,** because it has more reliable sources.
Spelling	**Commonly confused word** (Replacing the word)	The word **effects** may be used incorrecly. Review the following notes to determine the appropriate usage for your context.	It also **effects** a person mentally where a person lacks concentration or focus easily because his or her mind is on the message alerts and not the surroundings.
	Confused words (Correct your spelling/Replace your word)	The word **think** doesn't seem to fit this context. Consider replacing it with a different one.	For instance, people in my country when they become teenager the first **think** they think about is what should I do after graduating from high school or how can I get the car I want.

	Misspelled words (Add a hyphen/Change the capitalization/ Correct your spelling)	It appears that the word **china** may be a proper noun in this context. Consider capitalizing the word.	Before planes the distance from Europe to **china** for example take months by a ship and may take years by horses.
	Unknown words (Change the spelling)	Our dictionary does not include the word **hanafuda**. You can add it to your personal dictionary to prevent future alerts.	At the beginning of Nintendo, it was just a small company which made toys and handmade **hanafuda** gaming cards in 1889.
Conventions	**Improper formatting** (Add a space/Remove a space/Remove the space)	It appears that you have an unnecessary space in **people' s**. Consider removing the space.	In a rapidly developing society, **people' s** level of education becomes increasingly significant.
	Mixed dialects of English (Change the spelling)	The spelling of **behaviour** is a non-American variant. For consistency, consider replacing it with the American English spelling.	Nonetheless, there are no regulations concerning the **behaviour** of men in these situations.
Conciseness	**Wordy sentences** (Change the wording/Remove redundancy/Remove the phrase/Remove the preposition/Remove wordiness/Replace the phrase)	The phrase **in order for** may be wordy. Consider changing the wording.	These inventions are made **in order to** make people's lives easier and better, however there are not only advantages but also lots of disadvantages of using the inventions.

Appendix D. IRB Approval



Oklahoma State University Institutional Review Board

Date:	12/10/2020
Application Number:	IRB-20-547
Proposal Title:	L2 writing teachers' attitudes toward Grammarly feedback when it is used to complement their feedback
Principal Investigator:	Svetlana Koltovskaia
Co-Investigator(s):	
Faculty Adviser:	STEPH LINK
Project Coordinator:	
Research Assistant(s):	
Processed as:	Exempt
Exempt Category:	

Status Recommended by Reviewer(s): Approved

The IRB application referenced above has been approved. It is the judgment of the reviewers that the rights and welfare of individuals who may be asked to participate in this study will be respected, and that the research will be conducted in a manner consistent with the IRB requirements as outlined in 45CFR46.

This study meets criteria in the Revised Common Rule, as well as, one or more of the circumstances for which <u>continuing review is not required.</u> As Principal Investigator of this research, you will be required to submit a status report to the IRB triennially.

The final versions of any recruitment, consent and assent documents bearing the IRB approval stamp are available for download from IRBManager. These are the versions that must be used during the study.

As Principal Investigator, it is your responsibility to do the following:

1. Conduct this study exactly as it has been approved. Any modifications to the research protocol must be approved by the IRB. Protocol modifications requiring approval may include changes to the title, PI, adviser, other research personnel, funding status or sponsor, subject population composition or size, recruitment, inclusion/exclusion criteria, research site, research procedures and consent/assent process or forms.
2. Submit a request for continuation if the study extends beyond the approval period. This continuation must receive IRB review and approval before the research can continue.
3. Report any unanticipated and/or adverse events to the IRB Office promptly.
4. Notify the IRB office when your research project is complete or when you are no longer affiliated with Oklahoma State University.

Please note that approved protocols are subject to monitoring by the IRB and that the IRB office has the authority to inspect research records associated with this protocol at any time. If you have questions about the IRB procedures or need any assistance from the Board, please contact the IRB Office at 405-744-3377 or irb@okstate.edu.

Sincerely,
Oklahoma State University IRB

Appendix E. A Hypothetical Scenario.

Imagine that you are teaching the L2 writing course (ENGL 1123) this semester with ten students in your class. Your students have just submitted rough drafts of their argumentative essay (the first major writing assignment) to you for formative feedback. You have 2-4 days to provide feedback. It is noteworthy that this semester, as part of the curriculum, you are using Grammarly to complement your feedback.

1. Provide formative feedback as you normally would. Formative feedback means feedback for learning. You need to focus on things you want your students to improve before they submit their essays for summative assessment (grades).

2. You can upload students' essays to Grammarly. But you need to register first. However, this is not necessary. You were given ten Grammarly reports for ten essays that you can look at to supplement your feedback.

3. Try to provide feedback within 2-4 days. This is usually how much it takes for teachers to provide feedback.

4. Once you finish providing feedback, upload essays with your feedback to the Google folder. Contact me for a follow-up interview. The interview should take no more than an hour.

Appendix F. Argumentative Essay Rubric.

Area	Points	Description
Content	26-30	EXCELLENT TO VERY GOOD: knowledgeable, substantive, thorough development of thesis, relevant to assigned topic
	21-25	GOOD TO AVERAGE: some knowledge of subject, adequate range, limited development of thesis, mostly relevant to topic, but lacks detail
	16-20	FAIR TO POOR: limited knowledge of subject, little substance, inadequate development of topic
	0-15	VERY POOR: does not show knowledge of subject, non-substantive, not pertinent, not enough to evaluate
Organization	26-30	EXCELLENT TO VERY GOOD: Clear thesis with controlling idea, organized, unified, coherent, has an introduction, body, conclusion, has transitions/topic sentences, wrap-and-tie sentences.
	21-25	GOOD TO AVERAGE: somewhat choppy, loosely organized but main ideas stand out, has some topic and wrap-and-tie sentences.
	16-20	FAIR TO POOR: ideas disconnected, lacks logical sequencing, limited topic and wrap-and-tie sentences.
	0-15	VERY POOR: does not communicate, no organization, not enough to evaluate, no topic or wrap-and-tie sentences.
Language Use	16-20	EXCELLENT TO VERY GOOD: effective complex constructions, few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions
	11-15	GOOD TO AVERAGE: effective but simple constructions, minor problems in complex constructions, several errors of agreement,

		tense, number, word order/function, articles, pronouns, prepositions, but meaning seldom obscured
	6-10	FAIR TO POOR: major problems in simple/complex constructions, frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions, and/or fragments, run-ons, deletions, meaning confused or obscured
	0-5	VERY POOR: virtually no mastery of sentence construction rules, dominated by errors, does not communicate, or not enough to evaluate
Documentation style	8-10	EXCELLENT TO VERY GOOD: includes properly formatted in-text citations and references page, follows APA page format (title page, page number, Times New Roman, 12 points, double-spaced, etc)
	5-7	GOOD TO AVERAGE: occasional errors of in-text citations and references page, follows APA page format with occasional errors (title page, page number, Times New Roman, 12 points, double-spaced, etc).
	2-4	FAIR TO POOR: frequent errors of in-text citations and references page, follows APA page format with frequent errors (title page, page number, Times New Roman, 12 points, double-spaced, etc)
	0-1	VERY POOR: no mastery of in-text citations and references page; does not follow APA page format
Mechanics	4	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions, few errors of spelling, punctuation, capitalization, paragraphing
	3	GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing, but meaning no obscured
	2	FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing, meaning confused or obscured
	0-1	VERY POOR: no mastery of conventions, dominated by errors of spelling, punctuation, capitalization, paragraphing, or not enough to evaluate
TOTAL POINTS	100	

Appendix G. Semi-structured Interview Questions.

I. *Demographic information*

1. What’s your preferred nickname?
2. Where are you from?
3. What’s your first language?
4. What’s your degree?
5. Do you have L2 writing teaching experience? If yes, how long have you been teaching L2 writing?
6. What’s your overall English language teaching experience?

II. *Questions about prior experience with AWE*

7. What do you know about automated writing evaluation (AWE) systems and similar tools? Have you ever used one before?
8. What is your general attitude toward using AWE in L2 writing classrooms?

III. *Questions about teachers’ perceptions of Grammarly to supplement their feedback*

9. Have you ever used Grammarly before? If yes, how long? for what?
10. (Reminds the scenario) How did you feel about using Grammarly to supplement your formative feedback in this scenario?
11. When using Grammarly to complement your feedback, how have your ways of feedback provision changed as to feedback amount and feedback type?
12. Can you tell me about the process of using Grammarly to supplement your feedback?
13. Did you find Grammarly useful? If so, please describe how did it help you?
14. Do you think Grammarly feedback is beneficial to students?
15. Would you use Grammarly as a complement to your feedback?
16. What would you recommend to other teachers in regard to using Grammarly to supplement their feedback?
17. Is there anything you want to add?

Appendix H. Error Categories Used to Code Teachers’ Feedback.

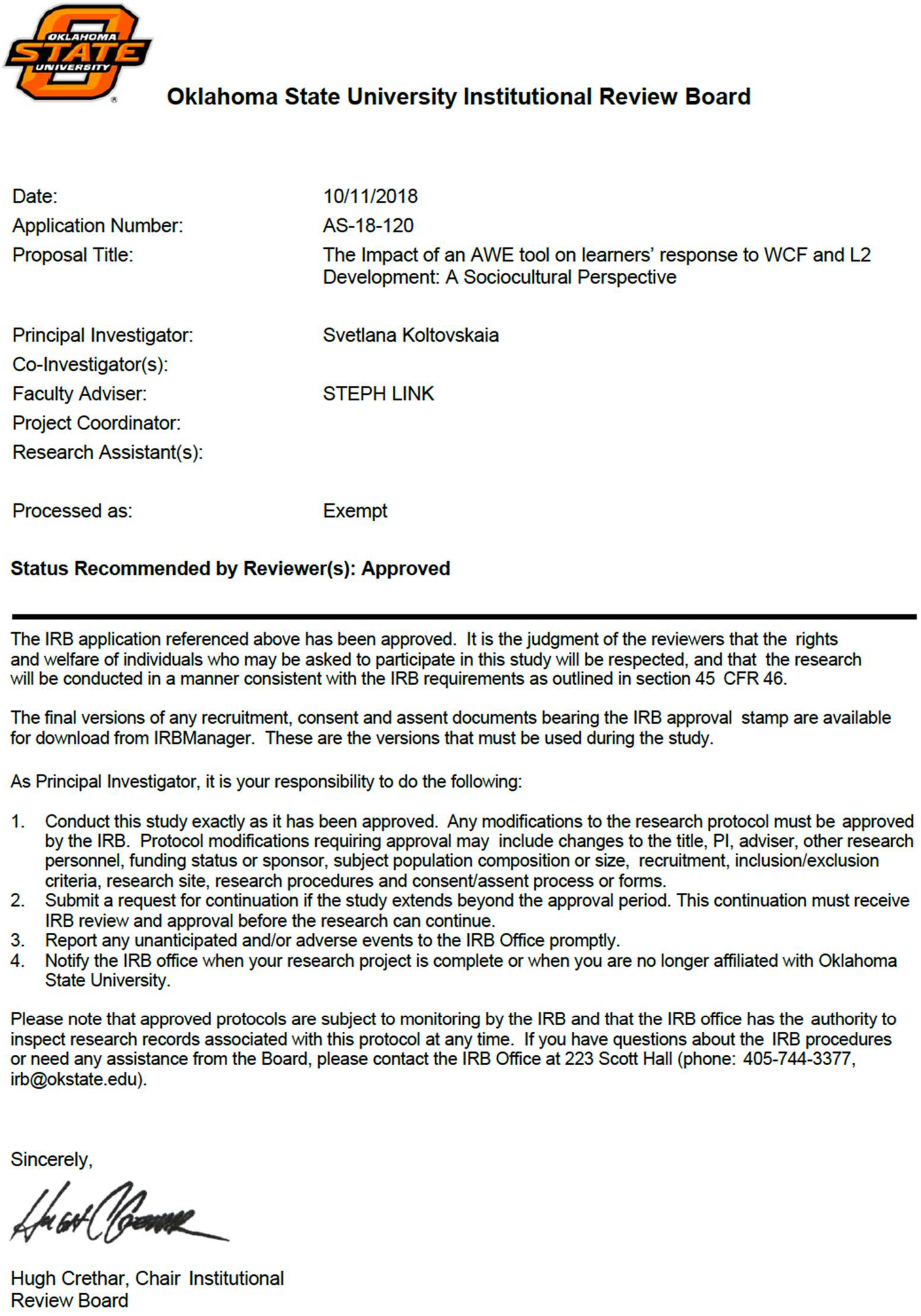
Grammarly feedback evaluation	Notes on Grammarly’s feedback in teacher’s comments	
Positive	Praise for achievement or encouragement about performance	
General	Comments on the overall quality of an essay in both HOCs and LOCs areas	
Higher-order concerns (HOCs) Discourse Level	Content	Clarity or understandability
		Development or lack of development
		Accuracy of information, truth value of claim, accuracy of interpretation
	Organization, coherence, cohesion	Transitions
		Thesis statement
		Topic Sentence

		Coherence, cohesion
		Idea placement
		Paragraph order
Lower-level concerns (LOCs) Form Level	Vocabulary	Word choice, collocations, phrasing
	Grammar/Syntax and morphology	Sentence structure
		Word choice
		Verb Tense
		Noun endings (singular/plural)
		Verb form
		Word form
		Articles/determiners
		Pronouns
		Preposition
		Conjunctions
		Subject-verb agreement
		Fragments
		Missing word
		Extra word, redundancy, or repetition
		Overall quality of grammar
	Mechanics	Punctuation
		Spelling
		Documentation or attribution
		Formatting and style
Adapted from Ene & Upton (2014) and Ferris (2006)		

Appendix I. Teachers' Feedback for Ten Essays.

			Mik %	Mei %	Maria %	Rob %	Jackson %	Heaven %
Grammarly feedback evaluation	Comments on Grammarly's feedback		-	-	-	-	-	1.7
Positive feedback	Praise for achievement or encouragement about performance		7.9	-	-	14.3	3.9	22.0
General	The overall quality of an essay in all its aspects often coupled with positive feedback		-	1.7	-	-	-	10.2
Higher-order concerns (HOCs) Discourse Level	Content	Clarity or understandability	10.5	1.3	3.2	-	9.3	5.1
		Development or lack of development	23.7	1.3	23.8	8.6	19.4	13.6
		Accuracy of information, truth value of claim, accuracy of interpretation	-	0.4	-	2.9	7.0	3.4
	Organization, coherence, cohesion	Transitions	5.3	-	-	11.4	-	-
		Thesis statement	2.6	0.4	-	-	1.6	1.7
		Topic Sentence	2.6	1.3	4.8	2.9	0.8	1.7
		Coherence, cohesion	7.9	0.4	1.6	14.3	-	-
		Idea placement	2.6	1.3	-	5.7	2.3	-
		Paragraph order	2.6	-	-	-	-	-
Lower-level concerns (LOCs) Form Level	Vocabulary	Word choice, collocations, phrasing	2.6	14.9	12.7	11.4	8.5	3.4
	Grammar/Syntax and morphology	Sentence structure	2.6	10.2	11.1	5.7	2.3	10.2
		Word choice	-	1.7	-	-	-	-
		Verb Tense	-	0.4	-	-	-	1.7
		Noun endings (singular/plural)	-	4.7	-	2.9	-	-
		Verb form	-	3.8	-	-	-	1.7
		Word form	-	4.3	-	-	-	5.1
		Articles/determines	-	3.8	-	-	-	1.7
		Pronouns	2.6	0.9	-	-	-	-
		Prepositions	-	3.4	-	-	0.8	1.7
		Conjunctions	-	1.7	-	-	-	-
		Subject-verb agreement	-	2.1	1.6	-	-	-
		Fragments	2.6	0.4	1.6	2.9	-	-
		Missing word	-	6.4	-	-	-	1.7
		Extra word, redundancy, or repetition	-	6.4	1.6	-	2.3	-
		Overall quality of grammar	-	0.4	-	14.3	-	-
	Mechanics	Punctuation	-	11.1	6.3	-	1.6	1.7
		Spelling	5.3	7.7	4.8	-	13.2	8.5
		Spacing	-	0.4	-	-	0.8	-
		Documentation or attribution	18.4	4.7	27.0	-	20.2	3.4
		Formatting and style	-	2.6	-	2.9	6.2	-

Appendix J. IRB Approval.



Oklahoma State University Institutional Review Board

Date:	10/11/2018
Application Number:	AS-18-120
Proposal Title:	The Impact of an AWE tool on learners' response to WCF and L2 Development: A Sociocultural Perspective
Principal Investigator:	Svetlana Koltovskaia
Co-Investigator(s):	
Faculty Adviser:	STEPH LINK
Project Coordinator:	
Research Assistant(s):	
Processed as:	Exempt

Status Recommended by Reviewer(s): Approved

The IRB application referenced above has been approved. It is the judgment of the reviewers that the rights and welfare of individuals who may be asked to participate in this study will be respected, and that the research will be conducted in a manner consistent with the IRB requirements as outlined in section 45 CFR 46.

The final versions of any recruitment, consent and assent documents bearing the IRB approval stamp are available for download from IRBManager. These are the versions that must be used during the study.

As Principal Investigator, it is your responsibility to do the following:

1. Conduct this study exactly as it has been approved. Any modifications to the research protocol must be approved by the IRB. Protocol modifications requiring approval may include changes to the title, PI, adviser, other research personnel, funding status or sponsor, subject population composition or size, recruitment, inclusion/exclusion criteria, research site, research procedures and consent/assent process or forms.
2. Submit a request for continuation if the study extends beyond the approval period. This continuation must receive IRB review and approval before the research can continue.
3. Report any unanticipated and/or adverse events to the IRB Office promptly.
4. Notify the IRB office when your research project is complete or when you are no longer affiliated with Oklahoma State University.

Please note that approved protocols are subject to monitoring by the IRB and that the IRB office has the authority to inspect research records associated with this protocol at any time. If you have questions about the IRB procedures or need any assistance from the Board, please contact the IRB Office at 223 Scott Hall (phone: 405-744-3377, irb@okstate.edu).

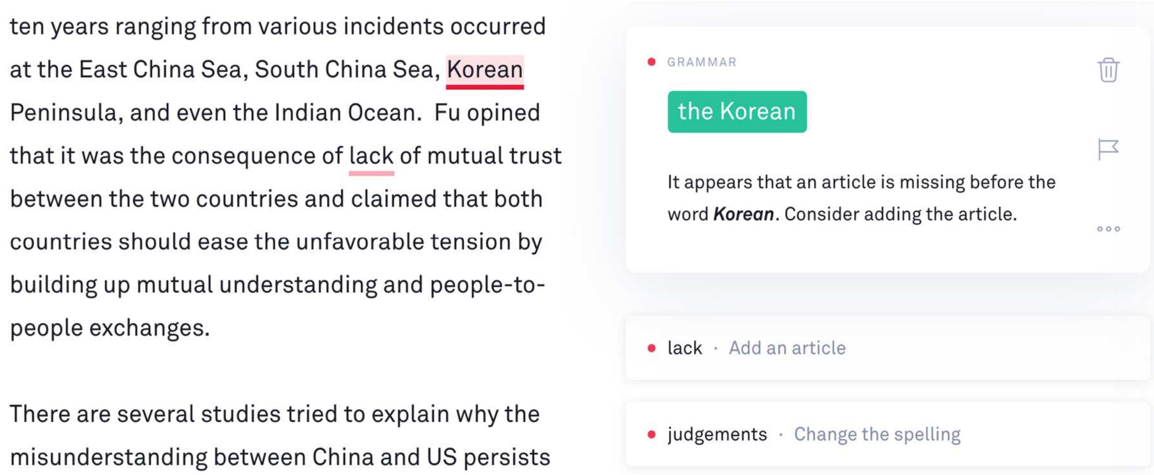
Sincerely,

Hugh Crethar, Chair Institutional
Review Board

Appendix K. Diagnostic Assessment Prompt.

For this diagnostic assessment, you will be asked to read the article "How to get a more sustainable style[2]" and write a 250-350 word summary. You should present the main points of the article. Use APA in-text citations if necessary to avoid plagiarism. You will have 45 minutes to read the article and write your summary. Save your file as diagnostic.myname. Upload your summary to Dropbox 5 minutes before the class ends.

Appendix J. Grammarly's Interface Showing Text Editor and Feedback on Grammar.



ten years ranging from various incidents occurred at the East China Sea, South China Sea, Korean Peninsula, and even the Indian Ocean. Fu opined that it was the consequence of lack of mutual trust between the two countries and claimed that both countries should ease the unfavorable tension by building up mutual understanding and people-to-people exchanges.

There are several studies tried to explain why the misunderstanding between China and US persists

GRAMMAR

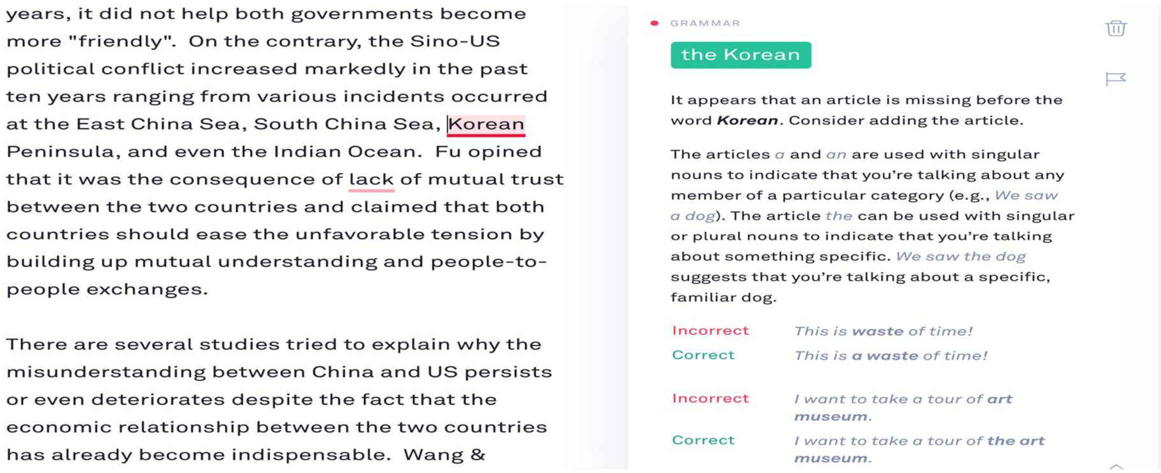
the Korean

It appears that an article is missing before the word ***Korean***. Consider adding the article.

lack · Add an article

judgements · Change the spelling

Appendix L. Grammarly's Interface Showing an Expended Suggestion on a Grammar Point.



years, it did not help both governments become more "friendly". On the contrary, the Sino-US political conflict increased markedly in the past ten years ranging from various incidents occurred at the East China Sea, South China Sea, Korean Peninsula, and even the Indian Ocean. Fu opined that it was the consequence of lack of mutual trust between the two countries and claimed that both countries should ease the unfavorable tension by building up mutual understanding and people-to-people exchanges.

There are several studies tried to explain why the misunderstanding between China and US persists or even deteriorates despite the fact that the economic relationship between the two countries has already become indispensable. Wang &

GRAMMAR

the Korean

It appears that an article is missing before the word ***Korean***. Consider adding the article.

The articles *a* and *an* are used with singular nouns to indicate that you're talking about any member of a particular category (e.g., *We saw a dog*). The article *the* can be used with singular or plural nouns to indicate that you're talking about something specific. *We saw the dog* suggests that you're talking about a specific, familiar dog.

Incorrect *This is **waste** of time!*

Correct *This is **a waste** of time!*

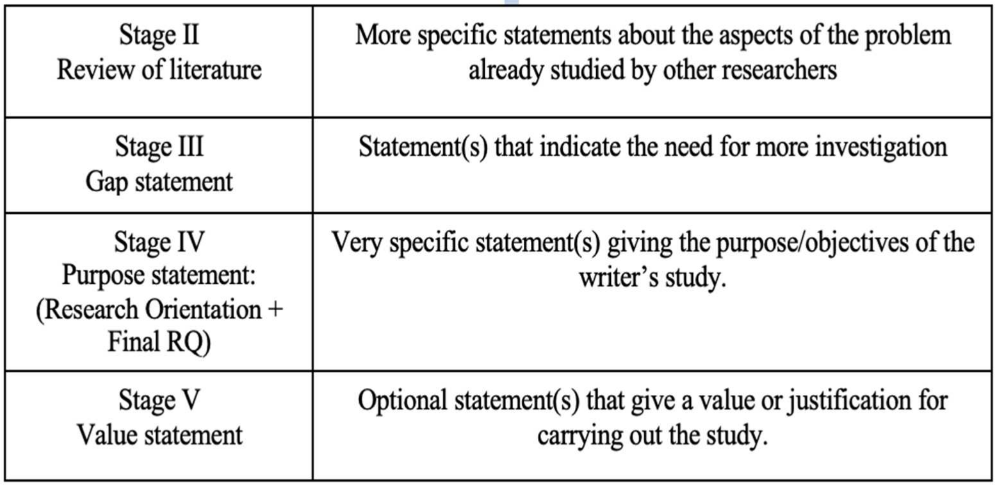
Incorrect *I want to take a tour of **art museum**.*

Correct *I want to take a tour of **the art museum**.*

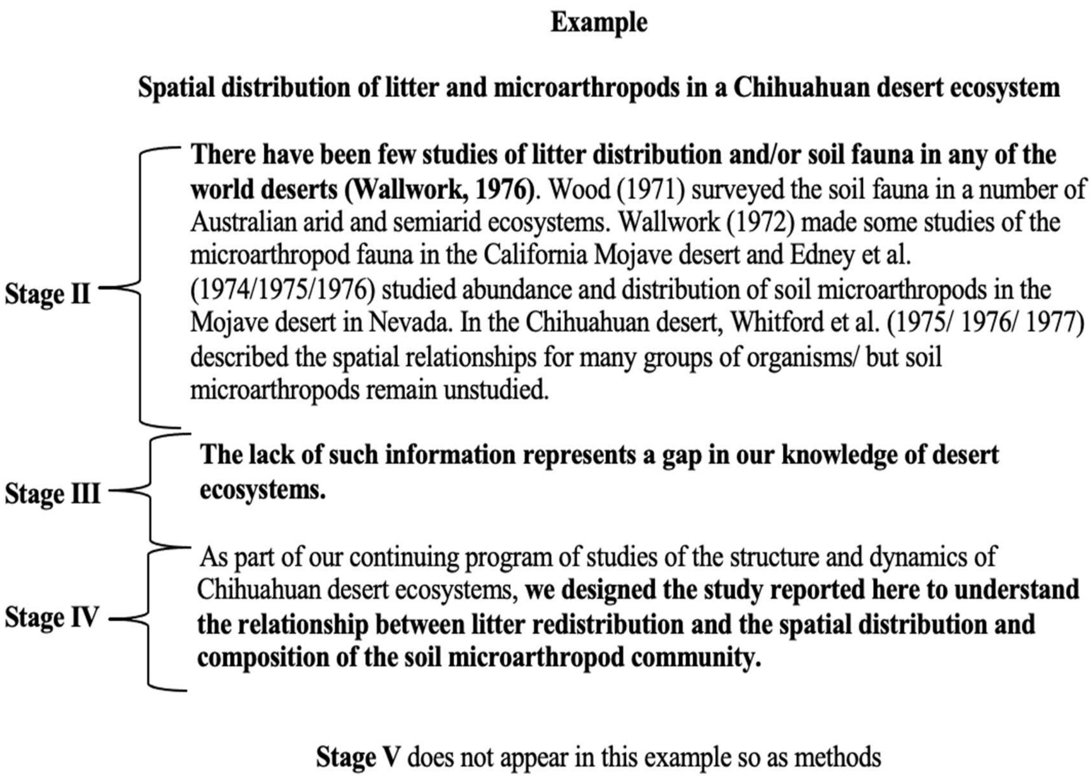
[2] The article can be found on this webpage: http://www.greenstrategy.se/how-to-get-a-more-sustainable-style-2/

Appendix M. The Literature Review Example.

Literature Review[3]



Stage II Review of literature	More specific statements about the aspects of the problem already studied by other researchers
Stage III Gap statement	Statement(s) that indicate the need for more investigation
Stage IV Purpose statement: (Research Orientation + Final RQ)	Very specific statement(s) giving the purpose/objectives of the writer's study.
Stage V Value statement	Optional statement(s) that give a value or justification for carrying out the study.



Example

Spatial distribution of litter and microarthropods in a Chihuahuan desert ecosystem

Stage II

There have been few studies of litter distribution and/or soil fauna in any of the world deserts (Wallwork, 1976). Wood (1971) surveyed the soil fauna in a number of Australian arid and semiarid ecosystems. Wallwork (1972) made some studies of the microarthropod fauna in the California Mojave desert and Edney et al. (1974/1975/1976) studied abundance and distribution of soil microarthropods in the Mojave desert in Nevada. In the Chihuahuan desert, Whitford et al. (1975/ 1976/ 1977) described the spatial relationships for many groups of organisms/ but soil microarthropods remain unstudied.

Stage III

The lack of such information represents a gap in our knowledge of desert ecosystems.

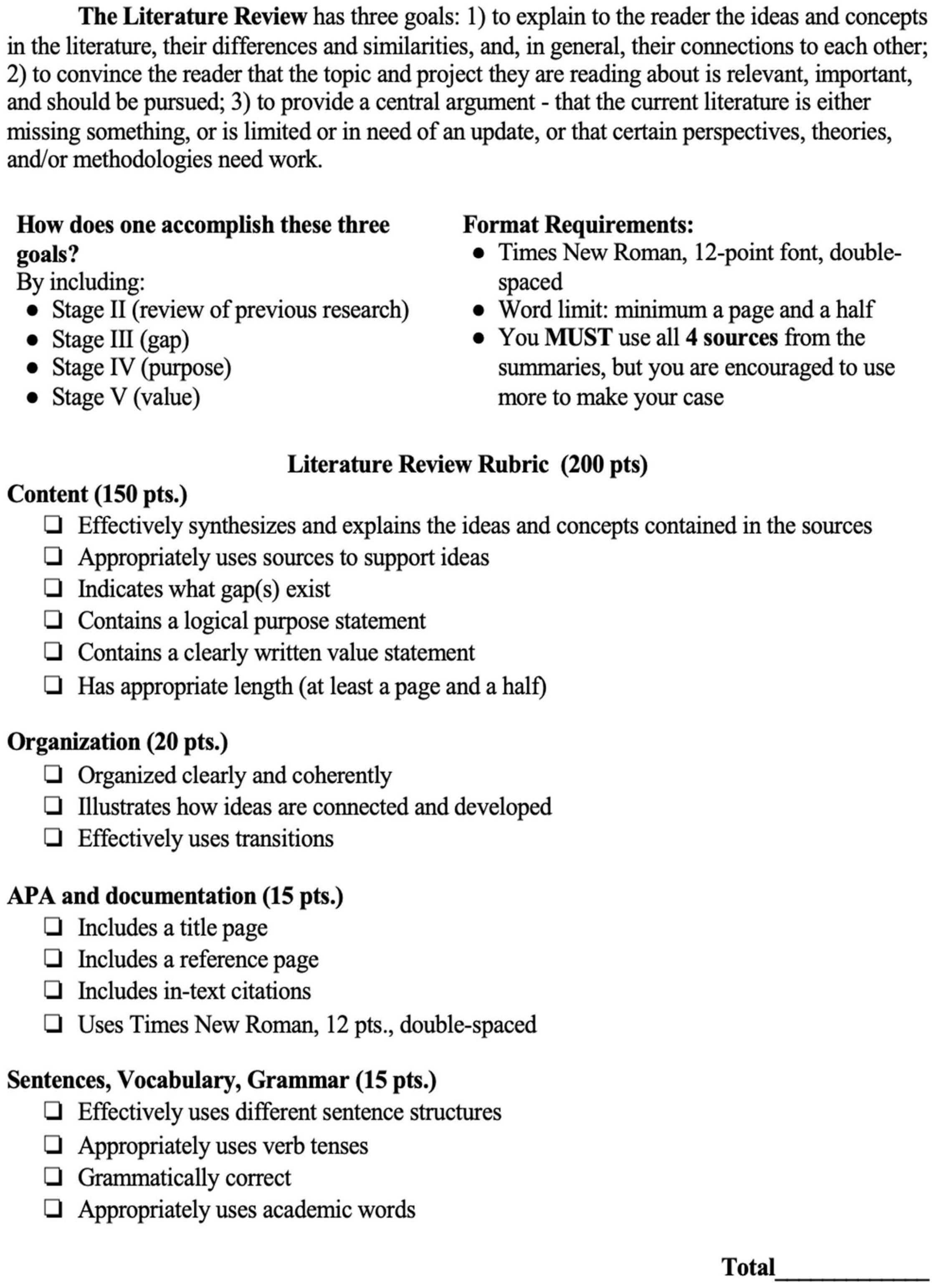
Stage IV

As part of our continuing program of studies of the structure and dynamics of Chihuahuan desert ecosystems, **we designed the study reported here to understand the relationship between litter redistribution and the spatial distribution and composition of the soil microarthropod community.**

Stage V does not appear in this example so as methods

[3] Adapted from Weissberg and Buker (1990).

Appendix N. The Literature Review Assignment Guidelines and Rubric.



The Literature Review has three goals: 1) to explain to the reader the ideas and concepts in the literature, their differences and similarities, and, in general, their connections to each other; 2) to convince the reader that the topic and project they are reading about is relevant, important, and should be pursued; 3) to provide a central argument - that the current literature is either missing something, or is limited or in need of an update, or that certain perspectives, theories, and/or methodologies need work.

How does one accomplish these three goals?

By including:

- Stage II (review of previous research)
- Stage III (gap)
- Stage IV (purpose)
- Stage V (value)

Format Requirements:

- Times New Roman, 12-point font, double-spaced
- Word limit: minimum a page and a half
- You **MUST** use all **4 sources** from the summaries, but you are encouraged to use more to make your case

Literature Review Rubric (200 pts)

Content (150 pts.)

□ Effectively synthesizes and explains the ideas and concepts contained in the sources
□ Appropriately uses sources to support ideas
□ Indicates what gap(s) exist
□ Contains a logical purpose statement
□ Contains a clearly written value statement
□ Has appropriate length (at least a page and a half)

Organization (20 pts.)

□ Organized clearly and coherently
□ Illustrates how ideas are connected and developed
□ Effectively uses transitions

APA and documentation (15 pts.)

□ Includes a title page
□ Includes a reference page
□ Includes in-text citations
□ Uses Times New Roman, 12 pts., double-spaced

Sentences, Vocabulary, Grammar (15 pts.)

□ Effectively uses different sentence structures
□ Appropriately uses verb tenses
□ Grammatically correct
□ Appropriately uses academic words

Total_____________

Appendix O. Stimulated Recall Script and Questions.

O.1. Stimulated Recall Script.[4]

We are going to watch the video of your error correction process (editing process with *Grammarly*). As we watch the video, I will be asking you questions about what you were thinking. As you watch your error correction process with *Grammarly*, try to recall what you were thinking at the time of error correction. Try to put your mind back into the task. Anytime you remember something, say it, interrupt me, ask me to stop the video if you want.

I am interested in finding out what you were thinking when you were correcting each error identified by *Grammarly* and why you accepted/rejected/ignored *Grammarly* feedback. It does not matter at all to me if those thoughts were silly or profound. I will audio-record our conversation so I do not have to divide my attention by taking notes. At the end of our stimulated recall, I will ask you a few questions about your opinion regarding *Grammarly*.

I am going to put the computer mouse on the table here and you can pause the video any time you want. So, if you want to tell me something about what you were thinking, you can click on the mouse to pause the video. If I have a question about what you were thinking, then I will click on the mouse to pause and ask you to talk about that part of the video. Is everything clear? Are you ready? Let's get started!

O.2. Stimulated Recall Guiding Questions.

(1) What were you thinking when you saw this number of alerts/ this many highlights?
(2) What were you thinking right then when you were reading the feedback/ when you paused after reading the feedback/ when you were correcting your error?
(3) Why did you reject/accept/ignore feedback provided by *Grammarly*?
(4) What did you think of the feedback provided by *Grammarly*?
(5) How did you arrive at accepting/rejecting/ignoring the feedback?
(6) Did you always understand the feedback provided by *Grammarly*? Why or Why not?

Appendix P. Semi-Structured Retrospective Interview Questions.

(1) Is this your first time using *Grammarly*? If yes, what is your overall impression of *Grammarly*? If not, how long have you been using Grammarly? What do you like about it? What do you dislike about it?
(2) In general, what do you think of *Grammarly's* feedback on the errors you made?
(3) Were you satisfied with the feedback provided? Why or why not?
(4) How do you think *Grammarly* helped you produce text with fewer errors?
(5) To what extent did *Grammarly* help you understand why you made errors?
(6) Do you think *Grammarly's* feedback is similar to human's feedback? Why or why not?

[4] Adapted from Gass and Mackey (2000)

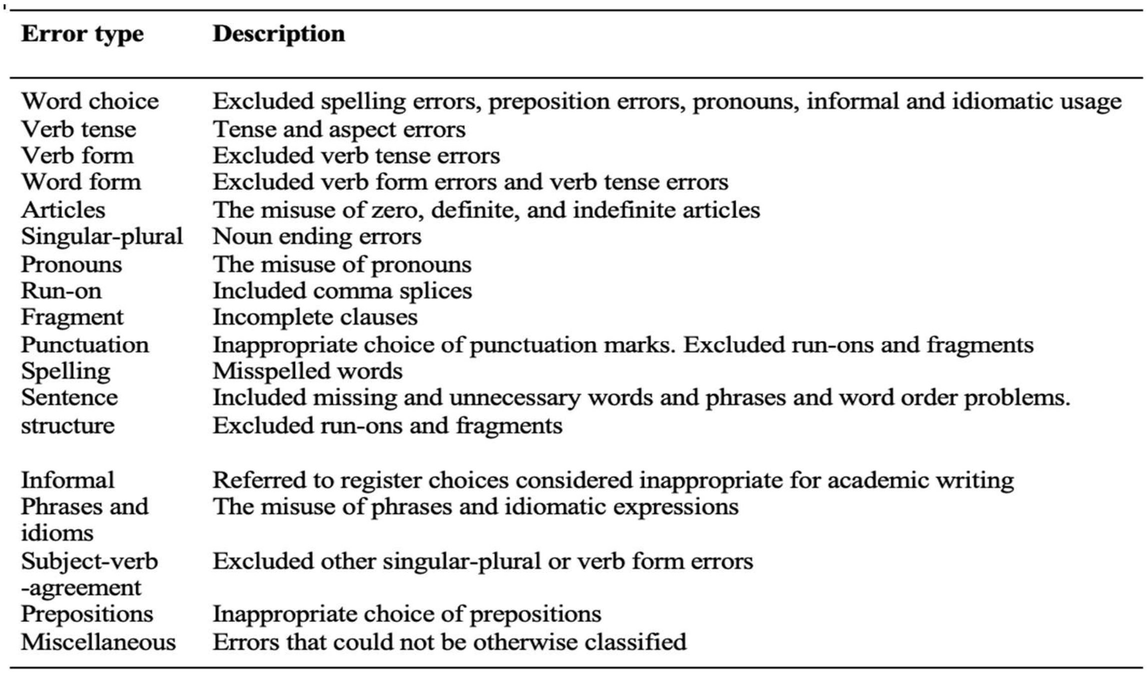
(7) Can you tell me a little bit about your proofreading/editing strategies? Did your strategy change when you used *Grammarly*? How much time do you usually spend on proofreading your paper? Did this time change with *Grammarly?* Why or why not?

(8) Will you consider using *Grammarly* in the future? Why or why not?

(9) What do you think of the usability of *Grammarly*? Did you encounter any problems when using *Grammarly*? Can you identify the strengths and weaknesses of *Grammarly* and its feedback?

(10) Is there anything else you have noticed about *Grammarly* that you would like to say?

Appendix Q. Han And Hyland's (2015) Taxonomy of Error Categories Adapted From Ferris (2006).



Error type	Description
Word choice	Excluded spelling errors, preposition errors, pronouns, informal and idiomatic usage
Verb tense	Tense and aspect errors
Verb form	Excluded verb tense errors
Word form	Excluded verb form errors and verb tense errors
Articles	The misuse of zero, definite, and indefinite articles
Singular-plural	Noun ending errors
Pronouns	The misuse of pronouns
Run-on	Included comma splices
Fragment	Incomplete clauses
Punctuation	Inappropriate choice of punctuation marks. Excluded run-ons and fragments
Spelling	Misspelled words
Sentence structure	Included missing and unnecessary words and phrases and word order problems. Excluded run-ons and fragments
Informal	Referred to register choices considered inappropriate for academic writing
Phrases and idioms	The misuse of phrases and idiomatic expressions
Subject-verb -agreement	Excluded other singular-plural or verb form errors
Prepositions	Inappropriate choice of prepositions
Miscellaneous	Errors that could not be otherwise classified

Appendix R. Accuracy of AWCF.

AWCF	Suggestions	Examples from the data	Accurate AWCF	Inaccurate AWCF
The Korean	It appears that an article is missing before the word **Korea.** Consider adding the article.	On the contrary, the Sino-US political conflict increased markedly in the past ten years ranging from various incidents occurred at the East China Sea, South China Sea, Korean Peninsula, and even the Indian Ocean.	✓	
A little	It appears that the phrase **only little** does not contain the correct article usage. Consider making a change.	Although there is many research that focus on the effect of insomnia on college students, only little literature investigated the relationship between insomnia and the college student's sleep hygiene in X University.		✓

Appendix S. Students' Editing Operations In Response To AWCF.

Editing operations	Description	AWCF	Suggestion	Examples from the data (after correction)
Accept	The error was corrected as *Grammarly* intended.	On -> in	It appears that the preposition **on** may be incorrect in this context. Consider changing it.	E.g., They showed that the media plays an important role in this international relationship, however, they did not further justify whether the media has played their roles properly or objectively.
Reject	The error was left uncorrected.	Persist,	It appears that you have an unnecessary comma in a compound object. Consider removing it.	E.g., The literatures reviewed above explained why the misunderstanding between China and US persists, and found that the media could exert much influence on International Relations by imposing an unreal image to a foreign country.
Substitute	The correction suggested by *Grammarly* was substituted by the student's own correction to address the error.	The bigger	The noun phrase **bigger previous study** seems to be missing a determiner before it. Consider adding an article.	E.g., However, data from a bigger previous study were used for the second part.

VITA

Svetlana Koltovskaia

Candidate for the Degree of

Doctor of Philosophy

Dissertation: AUTOMATED WRITING EVALUATION FOR FORMATIVE SECOND LANGUAGE ASSESSMENT: EXPLORING PERFORMANCE, TEACHER USE, AND STUDENT ENGAGEMENT

Major Field: English

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in English at Oklahoma State University, Stillwater, Oklahoma in July, 2022.

Completed the requirements for the Master of Arts in TESOL at Central Michigan University, Mount Pleasant, Michigan in 2015.

Completed the requirements for the Bachelor of Arts in Foreign Languages and Literature at Mirny Polyethnic Institute, Mirny, the Republic of Sakha, Russia, in 2009.

Experience:

Worked as a Graduate Teaching/Research Associate at Oklahoma State University in August 2016 – August 2022.

Worked as a Graduate Teaching Associate at Central Michigan University in August 2013 – May 2015.

Professional Memberships:
TESOL (Teaching English to Speakers of Other Language)
AAAL (American Association for Applied Linguistics)
CALICO (Computer-Assisted Language Instruction Consortium)