論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	An AI-based writing assistant ’ s impact on English language learners ’ writing proficiency
著者(和文)	GAYED JOHN MAURICE
Author(English)	John Maurice Gayed
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第12477号, 授与年月日:2023年3月26日, 学位の種別:課程博士, 審査員:CROSS JEFFREY SCOTT,野原 佳代子,山下 幸彦,佐藤 由利子,小泉 勇人,分山 達也
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第12477号, Conferred date:2023/3/26, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

An AI-based writing assistant’s impact on English language learners’ writing proficiency

by

John Maurice Gayed

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy, Global Engineering for Development, Environment, and Society

Department of Transdisciplinary Science and Engineering

School of Environment and Society

Tokyo Institute of Technology

March 2023

An AI-based writing assistant's impact on English language learners' writing proficiency

by

John Maurice Gayed
ORCID iD: 0000-0002-7028-3291

Department of Transdisciplinary Science and Engineering
Global Engineering for Development, Environment and Society
School of Environment and Society

Supervisor: Professor Jeffrey Scott Cross

TOKYO INSTITUTE OF TECHNOLOGY

March 2023

Dedication and Acknowledgements

> *"If you can do, or dream you can, begin it now, for boldness has genius, power and magic in it."*

- Robert Swan

While variations of this saying have often been attributed to Johann Wolfgang von Goethe, this version of the quotation comes from a famous TED speech by Robert Swan. His words come to mind whenever I come to a crossroads in my life and career. Keeping the idea of pushing through adversity is something I have tried to keep as a mantra in my life. Quite often the decisions we make don't result in the outcomes that were expected, but I'm certain that dithering and not moving forward with this doctoral research would have resulted in regrets somewhere down the line. My decision to start this doctoral degree came late in life and yet, somehow, the pieces seemed to have all fit together with minimal grey hairs gained over the past few years.

I would like to dedicate this doctoral dissertation first and foremost to my mom, Nabila. It was her passion and dedication to my education that set me down this path. She instilled in me a sense of hardheadedness and positivity that helped me push through when all the stresses of work and family became unbearable. Thank you, mom, my dear habibti. I hope I made you proud.

I am also indebted to my sister Nancy, who has always been and continues to be an inspiration to me. Thank you for being an amazing role model and a bedrock of calm and support.

This work is also dedicated to my family, Nao and Noah. You both put up with a husband and dad who worked constantly and came home late every day. Your smiles and hugs were the energy I needed to get this done.

Professor Cross has been a wonderful supervisor and I am indebted to him for his patient guidance all along the way. I had so many ideas when we first discussed my doctoral research, and his experience and expertise helped me stay on track from beginning to end.

All the members of Cross laboratory who gave me invaluable advice and pushed me to do better.

Lastly, I would like to thank Dr. May Kristine Jonson Carlon. My work could not have achieved this level of professionalism had it not been for your extensive knowledge and help. You have inspired me to become a better researcher and teacher and I count myself as blessed for having the opportunity to collaborate with you.

This work has also received the financial support of two Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) Grants, #22K00718 and #20H01719. These grants helped me expand this research and reach audiences that would not have otherwise been possible.

Table of Contents

Abstract

An AI-based online writing assistant with features that could potentially help second language users with their writing was developed from a wireframe concept into a usable web application. After the application was put into production form, it was tested on students in three experimental studies. The key features of the writing assistant are next word suggestions with confidence values using the publicly available GPT-2 language model and a reverse translation field that takes the users' English input and translates it back to them in their first language of choice. Later iterations of the writing assistant expanded its features to include metacognitive prompting and nudging. This research contains three major sections. One section is dedicated to three controlled empirical studies that employed the novel writing assistant and metacognitive training/prompting as a treatment condition. Another section focuses on a mixed-methods survey of educators to gain insight into their views on these AI based technologies being used in the classroom. Lastly, a brief overview of artificial intelligence trends and policies in the United States and Japan is detailed. A graphical abstract of the dissertation is represented in Figure 1.

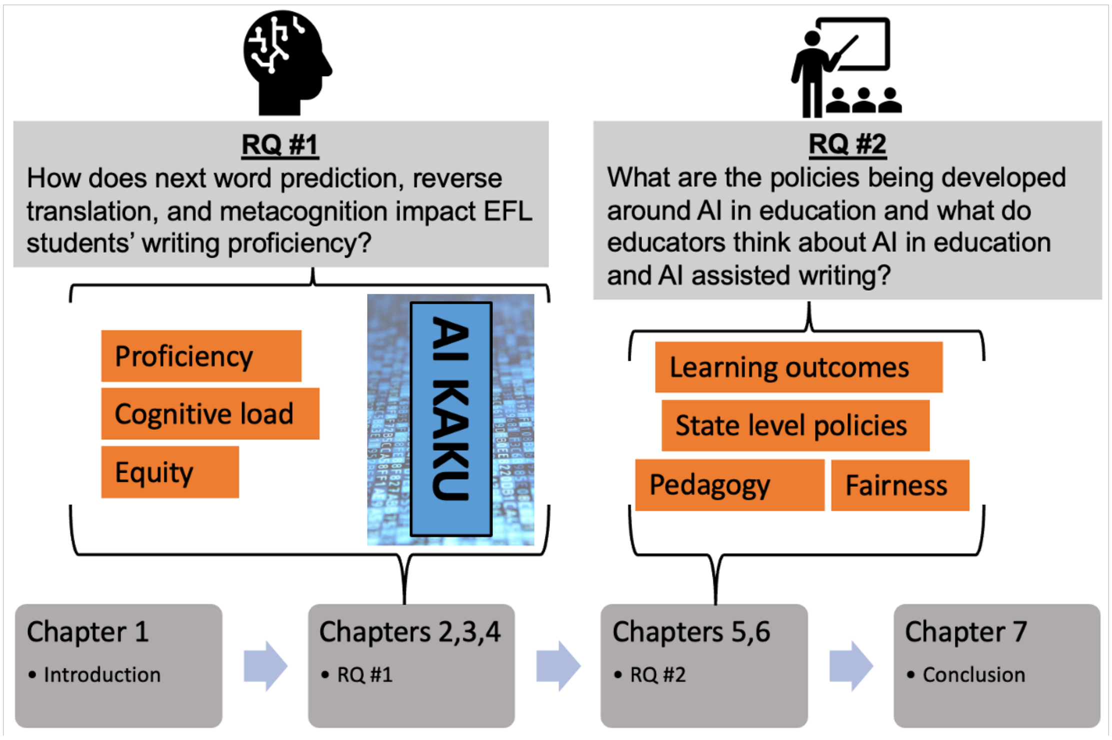

RQ #1
How does next word prediction, reverse translation, and metacognition impact EFL students' writing proficiency?
Proficiency
Cognitive load
Equity
AI KAKU
RQ #2
What are the policies being developed around AI in education and what do educators think about AI in education and AI assisted writing?
Learning outcomes
State level policies
Pedagogy
Fairness
Chapter 1
• Introduction
Chapters 2,3,4
• RQ #1
Chapters 5,6
• RQ #2
Chapter 7
• Conclusion

Figure 1. Graphical abstract

List of Tables

List of Figures

List of Acronyms

AI	Artificial Intelligence
AIED	Artificial Intelligence in Education
ANOVA	Analysis of Variance
APAC	Asia Pacific Regions
API	Application Programming Interface
AWE	Automated Word Evaluation
AWS	Automatic Word Suggestion
BNC	British National Corpus
CAI	Computer Assisted Instruction
CALICO	Computer-Assisted Language Instruction Consortium
CALL	Computer Assisted Language Learning
CEFR	Common European Framework of Reference
COCA	Corpus of Contemporary American English
EAP	English for Academic Purposes
EFL	English as a Foreign Language
EIKEN	Jitsuyo-Eigo Gino-Kentei
ELF	English as a Lingua Franca
ELL	English Language Learners
ESL	English as a Second Language
ETS	Educational Testing Service
GPT	Generative Pre-Trained Transformer
GSEC	Global Scientist and Engineers Course
ICALL	Intelligent Computer Assisted Language Learning
IELTS	International English Language Testing System
IQR	Interquartile Range

JSPS	Japan Society for the Promotion of Science
LD	Lexical Diversity
LFP	Lexical Frequency Profile
LLM-WSIME	Language Learners' Metacognitive Writing Strategies in Multimedia Environments
MEXT	The Ministry of Education, Culture, Sports, Science, and Technology
ML	Machine Learning
MT	Machine Translation
MLTunit	Mean Length of Minimally Terminable Unit
MOOC	Massive Open Online Course
MTLD	Measure of Textual Lexical Diversity
MWSQ	Metacognitive Writing Strategies Questionnaire
NDW	Number of Distinct Words
NITRD	The Networking and Information Technology Research and Development
NLP	Natural Language Processing
NNS	Non-native Speaker
OECD	Organization for Economic Co-operation and Development
SLA	Second Language Acquisition
STEM	Science, Technology, Engineering, and Mathematics
TESOL	Teaching English to Speakers of Other Languages
TOEFL	Test of English as a Foreign Language
TOT	Tip of Tongue
TPACK	Technological Pedagogical Content Knowledge
TTR	Type Token Ratio
WPM	Words per minute

List of Appendices

List of Publications

Peer-reviewed international journals:

1. Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing assistant's impact on English language learners. *Computers and Education: Artificial Intelligence, 3, 100055*. https://doi.org/10.1016/j.caeai.2022.100055

Peer-reviewed international post conference proceedings:

2. Gayed, J. M., Carlon, M. K. J., & Cross, J. S. (2022). The Matthew effect in CALL: Examining the equity of a novel intelligent writing assistant as English language support. In J. Colpaert, Y. Wang, & G. Stockwell (Eds.), *Proceedings of the XXIst International CALL Research Conference* (pp. 80–93). London: Castledown Publishers. https://doi.org/10.29140/9781914291050-12

3. Gayed, J. M., Carlon, M. K. J., & Cross, J. S. (2022). Impact on Second Language Writing via an Intelligent Writing Assistant and Metacognitive Training, *2022 IEEE Frontiers in Education Conference (FIE)*, 2022, pp. 1-9. https://doi.org/10.1109/FIE56618.2022.9962406

Chapter 1 Introduction

> *"Their trust in writing, produced by external characters which are no part of themselves, will discourage the use of their own memory within them. You have invented an elixir not of memory, but of reminding; and you offer your pupils the appearance of wisdom, not true wisdom, for they will read many things without instruction and will therefore seem to know many things, when they are for the most part ignorant and hard to get along with, since they are not wise, but only appear wise."*

- Plato, The Phaedrus 274d[1]

1.1 Background

The resistance to technological innovation and change in education can be traced back to the earliest examples of some cultures first developing systems to transfer knowledge. A famous example is seen in the above quotation from the Greek philosopher Plato. Here, the philosopher communicates his negative views of the modern invention of writing which he believed encouraged complacency. Plato felt that this "crutch" of writing would lead to a deficiency in the ability to gain "true knowledge". Certainly, his prediction was correct in the sense that a person's ability to store and recall information would be diminished when afforded the luxury of writing. However, time has shown us that we simply cannot ignore innovation even if it appears to be detrimental to cognitive function at first glance. How much human progress in the following centuries after Plato's dire declaration would have been stifled if the technological advancement of writing had not been embraced?

Research has shown that humans have the ability to redirect mental resources to maximize learning and knowledge transfer (Hecker et al., 2000). Examples of this can be seen in numerous examples of human development and this "distributed cognition" of mental resources is not just limited to academic endeavors. Specifically, this research focuses on innovation in educational technology which has seen numerous examples of disruptive technologies (word processors, video, the Internet, etc.) that were considered potentially

[1] Fowler et al., (1925). Plato (Vol. 5). Harvard University Press.

negative to human learning ability but, in fact, have been used successfully to further learning outcomes (Engelbrecht et al., 2020; Moranski & Henery, 2017). Even in cases where the more traditional paradigm exhibits some benefits over a newer (relative to the existing paradigm) technology, research has shown that the intersection of learning and technology is complex and deserves further attention (Hartley & Tynjälä, 2001).

1.2 Research Goals

"If I had asked people what they wanted, they would have said faster horses."[2]

The impetus behind this research came from a desire to improve the current digital writing paradigm for second language (L2) users. The researcher believes advances in natural language processing (NLP) had reached a point where they can be successfully applied in language education contexts. In addition to this, observing students struggling with their writing assignments in the classroom gave the researcher further inspiration to develop a digital application that can aid L2 students while they are writing in English.

Above the immediate need to improve the writing performance of students by applying newly developed technologies into the classroom, the researcher has seen little research into the impact of artificial intelligence (AI)-based technologies on the writing proficiency of L2 learners. Some AI-based technologies that have emerged range from Google's Smart Compose writing assistant which is mostly seen in their Gmail application (Gnacek et al., 2020), Grammarly which contains a suite of feedback options, as well as predictive text on mobile devices (Dizon & Gayed, 2021) or other dedicated applications such as Co-Writer Universal, which claims to offer writing assistance that integrates spell checking, grammar checking and word prediction under the framework of a word processor.

As such, predictive text seems to be a potentially useful tool for L2 learners as the technology offers real-time word-choice suggestions to writers based on the context of the words in each sentence from the first words typed. The researcher believes this might allow novice L2 writers the capability to reduce the cognitive load that is associated with the writing processes when students are asked to produce written output in English. However, research in this area, as Frankenberg-Garcia (2020) notes, is lacking in empirical analysis of the impact these tools have on students. This research then addresses this gap in the literature by

[2] Often attributed to Henry Ford, although the exact origins of this quotation are unclear.

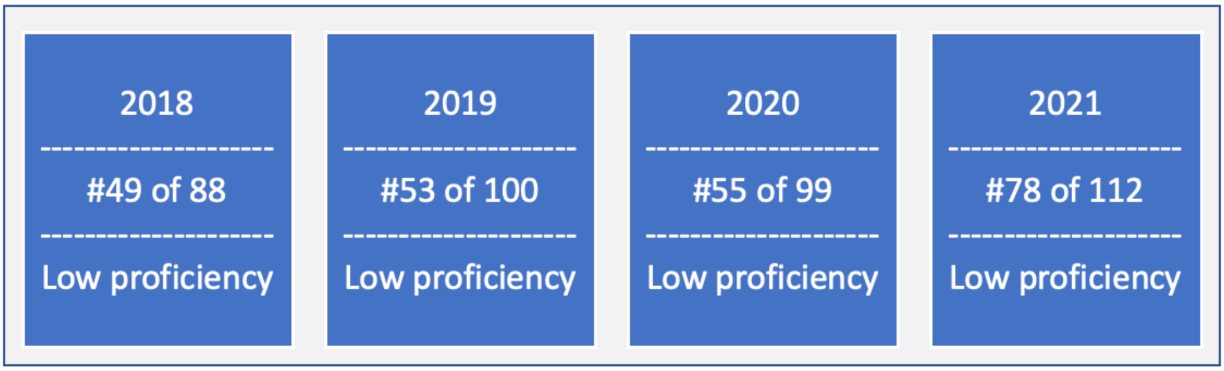
developing from the ground up a novel writing assistant specially designed for L2 users. After the application was wireframed, coded, and put into production form, the researcher assessed the effects the AI-based writing aid along with additional metacognitive tools that were tailored to improving L2 writing production.

The empirical studies in this research all focus on English as Foreign Language (EFL) learners studying at tertiary-level institutions or private language schools. The researcher believes the tools developed for this research will be able to reduce some of the cognitive load that is associated with the L2 writing process (Nawal, 2018), allowing EFL students the capability to produce richer (e.g., higher lexical frequency, less repetition), more complex writing when asked to produce written output in English. Gaining fluency in a second language can be broadly defined as "the ability to process language receptively and productively at a reasonable speed." (Nation, 2014, p. 11) and gains in writing fluency have often been cited as a goal of second language education (Alisaari & Heikkola, 2016) in addition to being the stated goal of many language learners.

1.3 Research in Context

English being used as a lingua franca is also stated as a reason for EFL leaners to improve their English proficiency. According to Beare (2020), the number of English learners around the world is only expected to grow; more specifically, the British Council's report "The English Effect" estimated in 2020, that two billion people will be using the English language. Japan, unfortunately, has historically ranked very poorly when assessed via standardized proficiency exams. Even more discouraging is the lack of improvement despite Japanese government efforts to reverse the poor proficiency exhibited by Japanese EFL leaners.

Based on data from Education First's (EF) annual assessment (EF bases their assessment on reading and listening skill tests) of English proficiency around the world (*EF EPI 2021 – EF English Proficiency Index – Japan*, n.d.), Figure 2 shows stagnating or decreasing proficiency among Japanese EFL learners between the years 2018-2021. This is despite the fact that the Japanese government is spending considerable capital to achieve stated goals such as, "…beginning in 2020 all high school graduates must achieve a level of English equivalent to B1 of the Common European Framework of Reference for Language (CEFR)." *The Japan Times* (2017).



2018	2019	2020	2021
#49 of 88	#53 of 100	#55 of 99	#78 of 112
Low proficiency	Low proficiency	Low proficiency	Low proficiency

Figure 2[3]*. Japan's English proficiency trend.*

In addition, English writing ability according to annual reports published by the English proficiency exam maker Educational Testing Service (ETS), (Figure 3) shows Japan as ranking in the bottom third when compared to other countries.

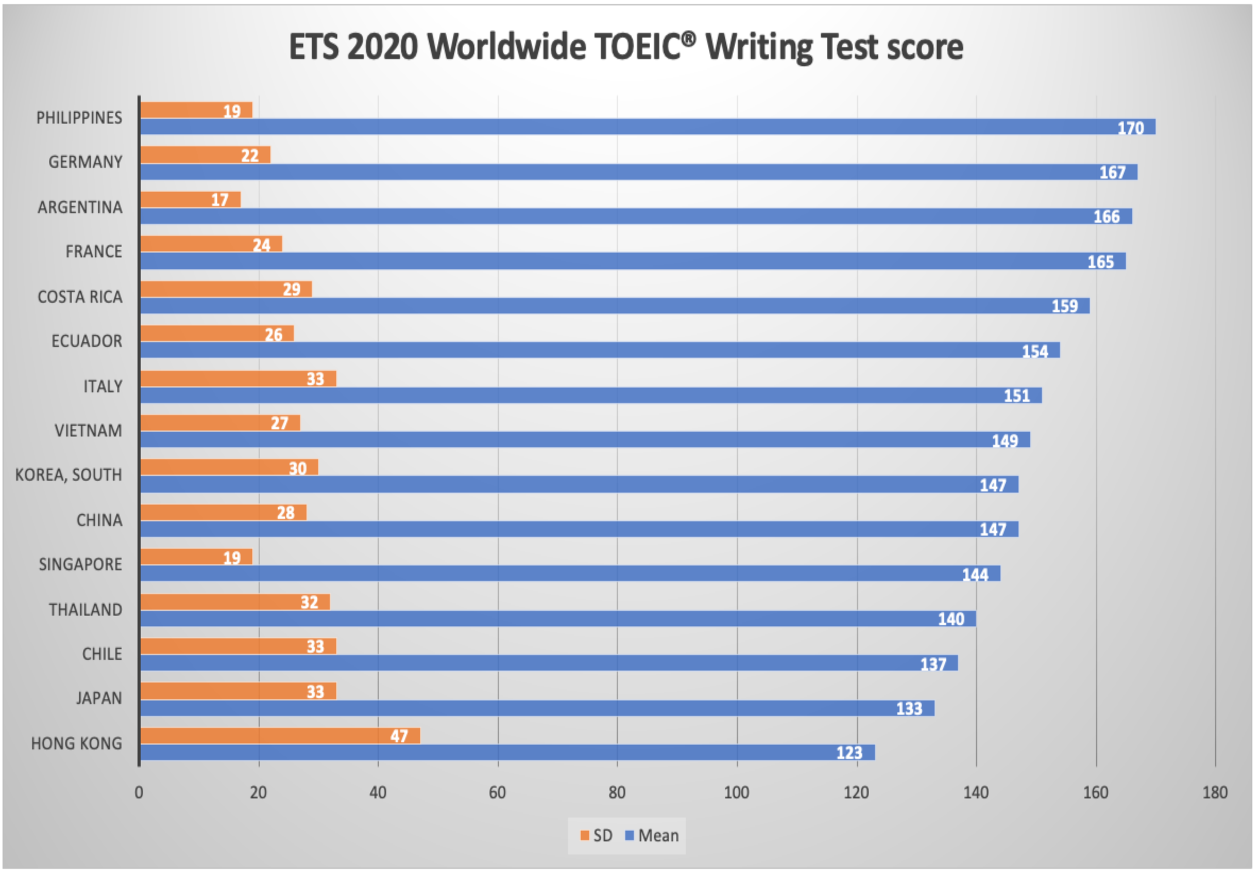


Figure 3[4]*. Worldwide TOEIC® Writing Scores*

("2020 Report on Test Takers Worldwide - TOEIC Speaking & Writing Tests," 2020) Certainly, the opportunity provided by technology can assist EFL students in Japan improve

[3] Adapted from Education First's data on English proficiency from 2018-2021. https://www.ef.com/wwen/epi/regions/asia/japan/

[4] Adapted from ETS TOEIC® 2020 Report on Test Takers Worldwide. https://www.iibc-global.org/library/default/toeic/official_data/pdf/Worldwide2020.pdf

their proficiency. This researcher aims to improve writing proficiency by providing EFL learners with a more supportive digital writing environment. By leveraging NLP technologies, EFL students can improve their writing output, ultimately improving their autonomy and agency as students participating on the global stage.

1.4 Framework

The researcher then began to conceptualize a digital writing assistant with features that could potentially help L2 users in their writing. After some internal discussion with laboratory members, the researcher decided to focus the writing assistant on two main features: next word suggestions with confidence values using the publicly available GPT-2 language model and a reverse translation field that takes the users' English input and translates it back to them in their first language of choice. Later iterations of the writing assistant expanded its features to include metacognitive prompting and nudging. The application was given the name "AI KAKU" and its user experience flow is visualized in Figure 4. The primary novelty of the digital writing assistant developed by the researcher is the next word prediction with the language model returning an associated confidence value for each prediction. This kind of writing assistance has not been employed in any known publicly available word processor and the researcher believes this to be a unique contribution in the field of Computer Assisted Language Learning (CALL).

The reverse translation function was added as a feature due to the prevalence of second language students using Machine Translation (MT) as part of their writing process (Ducar & Schocket, 2018). Traditionally, translation theory states that, "… translation, as a process, is always uni-directional: it is always performed in a given direction, 'from' a Source Language 'into' a Target Language" (Catford, 1965, p. 20). In digital writing, the typical use case of MT is users translating their first language into a second language (A → B) when the user needs word, phrase, sentence, and paragraph level assistance. AI KAKU's implementation, on the other hand, is a departure from this traditional paradigm and encourages the user to maintain writing in a second language by displaying their output back to them in their first language, in a sense an (A → A) loop. This is a form of digital scaffolding (Englert et al., 2005; Kang, 2018) that is also unique and has not been used in digital writing applications. More importantly, the researcher understands that the reverse translation displayed might introduce unnecessary "noise" to the cognitive processes involved in writing due to the translation engine's (Google Translate) mistranslation or producing translations that are not culturally appropriate. To that effect, AI KAKU has been updated to allow the user to turn off the reverse translation feature if the user finds it unhelpful.

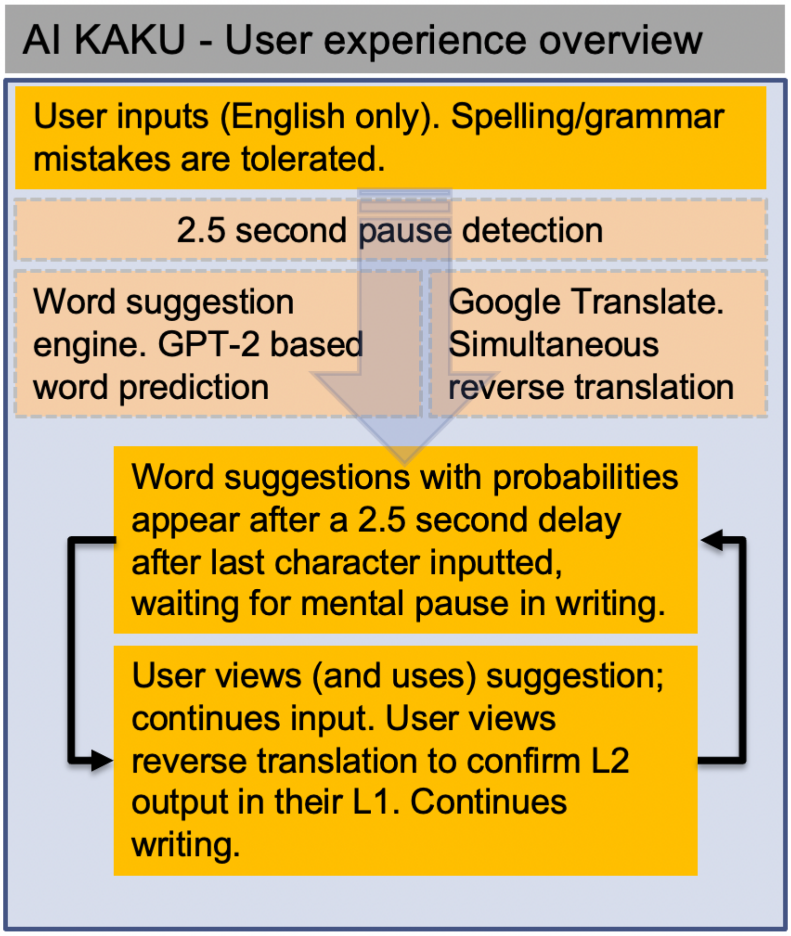
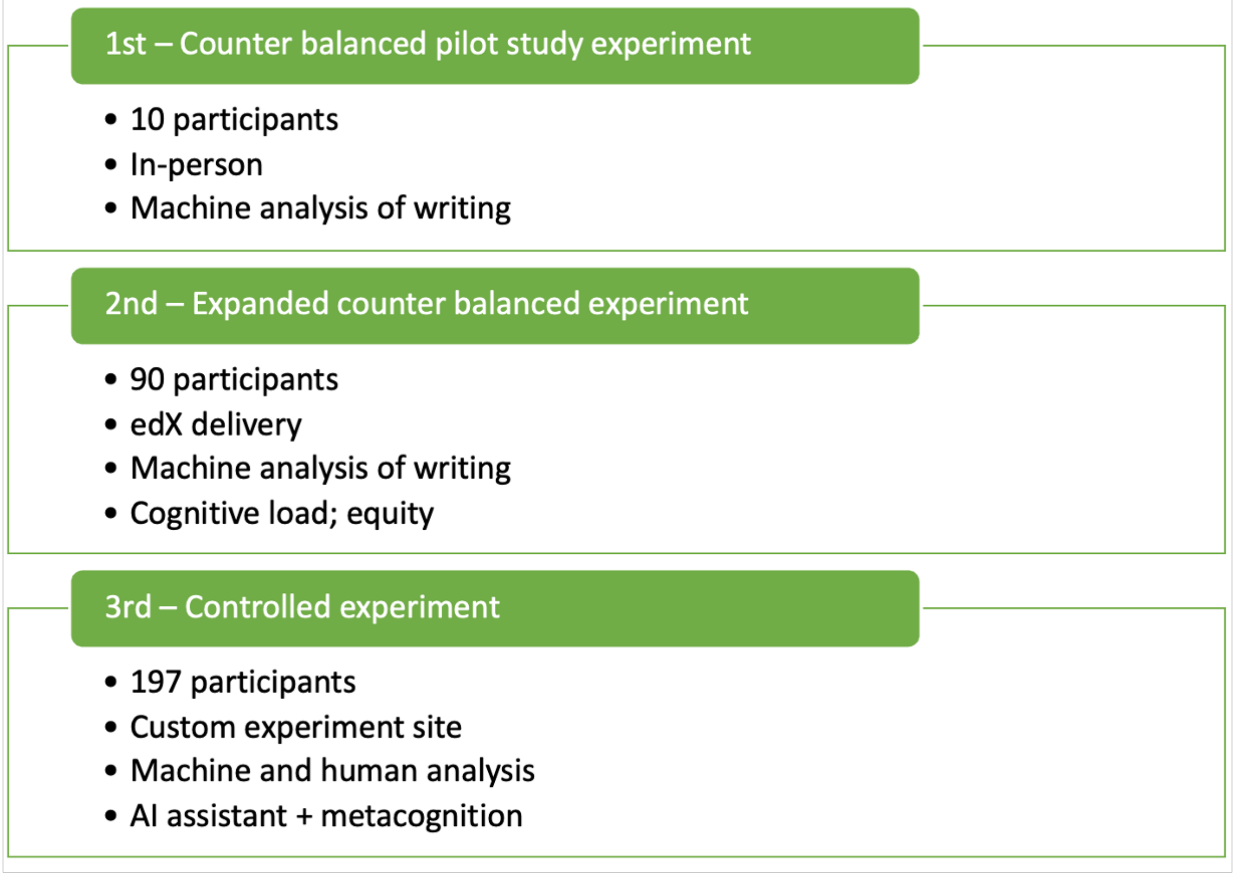

AI KAKU - User experience overview
User inputs (English only). Spelling/grammar mistakes are tolerated.
2.5 second pause detection
Word suggestion engine. GPT-2 based word prediction
Google Translate. Simultaneous reverse translation
Word suggestions with probabilities appear after a 2.5 second delay after last character inputted, waiting for mental pause in writing.
User views (and uses) suggestion; continues input. User views reverse translation to confirm L2 output in their L1. Continues writing.

Figure 4. Conceptional framework of AI KAKU.

1.5 Overview of Dissertation

The following chapters will expand upon and give detail to the themes that can be broken into three sections. One section is dedicated to three controlled empirical studies that employed the novel writing assistant and metacognitive training/prompting as treatment tools. Another section focuses on an overview of artificial intelligence trends and policies in the United States and Japan, and finally a mixed-methods survey of educators gives insight into educators' views on AI based technologies being used in the classroom and more specifically, their views on AI assisted writing in an EFL context.

Shortly after the conceptional framework of the writing assistant was created, a proof of concept was launched into production and tested in a pilot study. The pilot study is the basis of the first empirical study described in Chapter 2. The application was successfully launched on a production server and having some positive outcomes based on the data gained from the pilot study, the researcher decided to expand the research into a larger study that was served over the edX platform. Chapter 3 details this experiment which is similar in design to the pilot study but with a much larger number of participants. This experiment also investigates the cognitive load of participants while they are under control and treatment conditions in addition to examining writing quality factors. Another aspect that is investigated is how the treatment tool impacts participants of different English ability. Lastly, a more sophisticated (methodology) and in-depth (analysis) study is described in Chapter 4. This study is noticeably different from the first two studies as it employs a more traditional blind-controlled experimental design served to participants via a custom experiment site. In addition to employing the AI assistant in the treatment group, metacognitive training, prompting, and nudging were also included as part of the treatment factors. In this experiment, writing quality was analyzed with machine and human assessment. Writing samples produced under each experiment can be viewed in Appendix I (Writing Samples). Writing quality and cognitive load metrics are included with each sample. Finally, a visual representation of the different experiments is shown in Figure 5.



1st – Counter balanced pilot study experiment

- 10 participants
- In-person
- Machine analysis of writing

2nd – Expanded counter balanced experiment

- 90 participants
- edX delivery
- Machine analysis of writing
- Cognitive load; equity

3rd – Controlled experiment

- 197 participants
- Custom experiment site
- Machine and human analysis
- AI assistant + metacognition

Figure 5. Summary of experiments conducted.

Chapter 2 [5] Exploring an AI-based writing assistant's impact on English language learners: A pilot study.

Keywords: CALL, AI in education, L2 writing, cognitive load, AI agent, experimental research

2.1 Background

This chapter will describe the first of three experiments the researcher conducted for this doctoral research. Soon after the wireframe concept of the digital writing application was completed, coding was done so that the application could be installed on a production server and convenience sampling was used to recruit participants to join the pilot study. The researcher was still testing and fine-tuning the server environment and application code, therefore a pilot study with a limited number of participants was chosen as the first attempt to gain empirical data on the tool's impact on users. Prior to conducting the experiment, the researcher submitted a human subject research ethics application, and the researcher's host institution granted approval. The results of this study were published in a high-impact international journal and demonstrated that further research and development into the tool was warranted.

[5] Parts of this chapter have been published under: Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing assistant's impact on English language learners. *Computers and Education: Artificial Intelligence, 3, 100055*. https://doi.org/10.1016/j.caeai.2022.100055.

2.2 Introduction

In the last decade, research and development have certainly accelerated in the area of AI in Education (AIED) (Hwang et al., 2020). In particular, as Chen et al., (2020) identify natural language processing as a major area that is often applied to educational contexts. Yet, as they highlighted, more effort should be applied to the potential of applying AI in real classroom settings. The researcher of this study is proposing a novel application of AI that can be applied in the classroom. According to the British Council, it is estimated there are some 1.2 billion English language learners in the world (Sheehan, 2013). A common struggle for second language (L2) learners is the tip-of-the-tongue (TOT) state, a temporary mental state in language production where there is difficulty in retrieving an intended word (Abrams & Davis, 2016; Ecke & Hall, 2013; Stasenko & Gollan, 2019) EFL students who are tasked with producing written text in English often compose their ideas in their first language (L1) and then struggle mentally to translate those ideas into English while attempting to complete the writing task (Wolfersberger, 2003).

Regardless of the approach, writing in a second language (L2) involves considerable cognitive stress, such as translating from their L1 to L2 and engaging with digital mediating artifacts (online dictionaries, translation applications) to help them complete the writing task. This cognitive stress has been shown to hinder learners from focusing on higher-level writing tasks such as organization and revision (Kellogg, 2008) which are essential to developing writing proficiency and producing higher-level output. In addition, research has shown that language learners who have wider and richer lexical mastery achieve better comprehension and expression skills (Archibald & C. Jeffery, 2000; Liu, 2020; Saito & Akiyama, 2017). When framed in the bigger context of why writing skills are needed by EFL students, research by (Weissberg, 2006) has identified, the skill of writing is not gained in a vacuum but can be correlated to overall language acquisition as a language learner improves their ability.

When faced with a seemingly insurmountable task, L2 writers might turn to less scrupulous techniques to complete their work, such as wholesale machine translation of their L1 writing into the target language or using predictive text agents to produce whole blocks of text with little user input. For language learners, these activities do not contribute to their acquisition of the target language, nor does it contribute to an improvement in writing skills. Therefore, to assist EFL students in the writing process, the researcher's laboratory has designed and released an online-based writing application called **AI KAKU** (https://www.aikaku.app). The application's development stack is visualized in Figure 6 and the basic user experience concept behind AI KAKU can be seen in Figure 7.

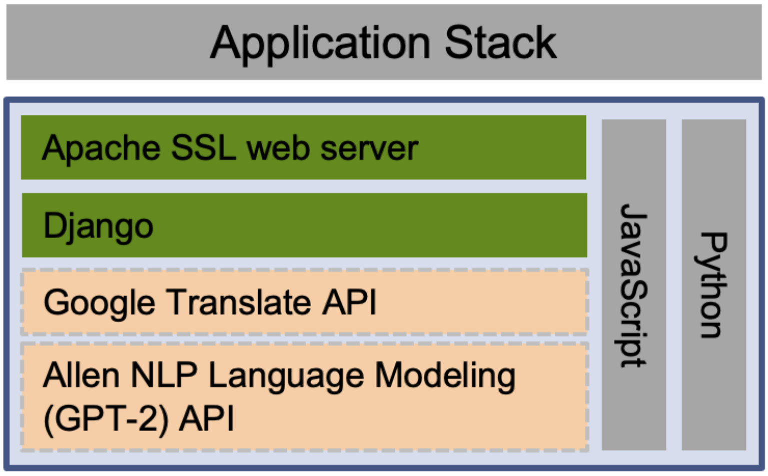


Figure 6. Application environment.

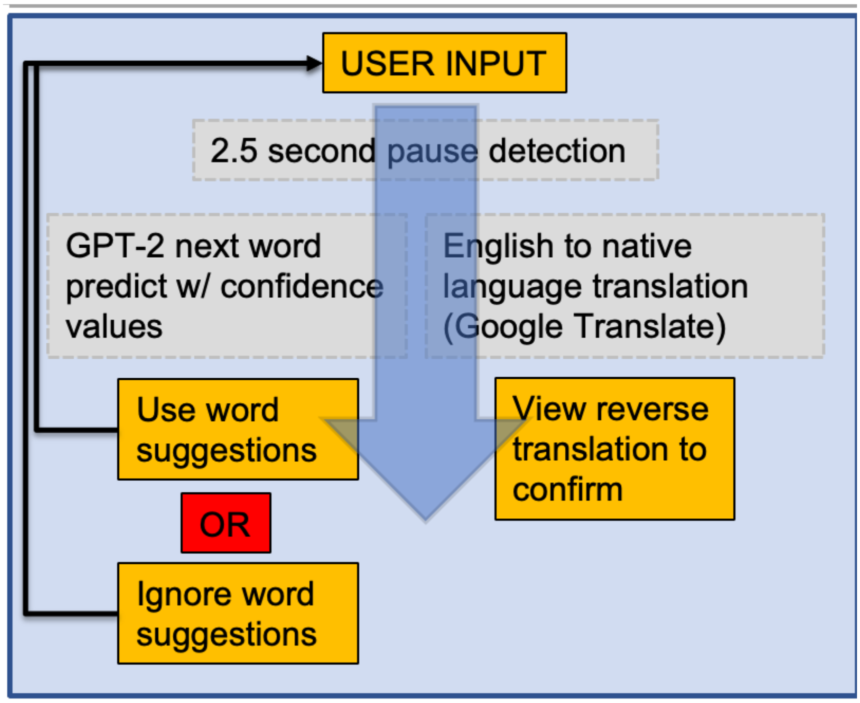

USER INPUT
2.5 second pause detection
GPT-2 next word predict w/ confidence values
English to native language translation (Google Translate)
Use word suggestions
OR
Ignore word suggestions
View reverse translation to confirm

Figure 7. User process flow in AI KAKU

The application has two distinct features that no known widely used writing application has implemented in production form. First, an AI-based word suggestion engine gives users word recommendations based on the user's input, similar to word suggestions that are commonly seen in text prediction applications on smart devices. The word suggestion engine that AI KAKU uses is based on an implementation of the GPT-2 language model (Radford et al., 2019) developed by the Allen Institute for AI (Gardner et al., 2018). The Allen Natural Language Processing (NLP) Next Token Language Model is published under the open-source Apache 2.0 license, allowing the researcher to modify and implement their own version for the purposes of this research. The second feature is a reverse translation function. This is being implemented via API access to the Google Translate application. As users write in English, a simultaneous translation in the user's L1 is displayed under their writing. The concept is to encourage the user to think in the L2 while giving L1 validation during the writing process. In other words, confirming for the user they are writing what they intended to write by showing them their input in their first language.

This study examines the participants' written text along two quality dimensions: lexical diversity (LD) and fluency. Lexical diversity is one measure of how "rich" a text is (Johansson, 2008); consequently, a writer that uses a wide variety of words in their text with little repetition commands a more sophisticated mastery of the language. The term fluency has been defined and quantified in several different ways in the literature. This study takes the approach of measuring fluency in two dimensions. One is the rate of production, or the amount of output produced in a given timeframe, a commonly seen method in the literature (Chenoweth & Hayes, 2001; Kawauchi & Kamimoto, 2000; Palviainen et al., 2012; Wolfe-Quintero et al., 1998). In addition to the production rate, the researcher measured the syntactic complexity or the number of clauses per t-unit in the participants' writing (Lu & Ai, 2015) to get a fuller understanding of written fluency.

AI advancements have led to more sophisticated intelligent writing assistants that offer synchronous feedback to the writer (such as Grammarly and Microsoft Editor). This user feedback is an expansion upon traditional word processing features such as spell and grammar checks that have existed since the earliest examples of digital writing appeared in the 1970s (Peterson, 1980). However, there has been little development in word processors aimed at EFL usage, and little research has been done into their potential impact on writing proficiency. This study attempts to analyze a technology that still has not seen widespread adoption. Even so, as Zheng and Warschauer, (2017) identified, technology is developing at a breathtaking speed and is fundamentally changing the way L2 students write. It is critical then to explore the potential effects these technologies have on second language production.

This research aims to address the following research questions:

1. How has AI KAKU impacted the lexical diversity and fluency of participant's writing?
2. What was the learner's impression of the utility of AI KAKU?

2.3 Related Work

2.3.1 Digital writing aids and considerations

Digital writing encompasses a range of skills (e.g., interacting with peers on social networking sites, blogging, online communication, and writing with word processors, among others) and has been examined as a research topic in applied linguistics and second language acquisition studies since it began to be widely used in the 1980s (Kirschenbaum, 2017). In a mixed-methods study by Moore, Rutherford and Craw (2019), the effects of digital writing tools on writing proficiency were examined with postsecondary students studying in Canada in an EFL context. The researchers found that while digital writing tools can improve writing proficiency, qualitative data showed it was important to have educators' guidance on a face-to-face level to complement the digital writing tool. Perry (2021) conducted a literature review of digital self-access resources (including writing resources) for L2 users and found evidence in the gathered data of strong efficacy when the tools were used in a well-structured program. However, the researcher found a gap in the literature for strong long-term acquisition improvements in the studies' participants.

From the perspective of digital literacy, Hamouma & Menezla (2019) highlight in a study of 80 EFL students a strong positive correlation between students having good digital literacy (including digital writing tool literacy) and students developing their English academic writing performance. Even among native-level English users, Purcell et al. (2013) elucidate in a survey of 2,462 educators the positive influence digital technologies have on student writing production. The mixed-methods study gathered survey and qualitative interview data to identify factors that influence student writing. One of the conclusions from the study is that newer digital writing platforms such as Google Docs are potentially transformative to the writing process due to their advanced capabilities.

Introducing digital aids, software applications, or any technology into the learning process must be examined for its possible negative influence on learning outcomes. Tight's

(2017) examination of learners of Spanish found that while the participants extensively used digital writing tools, low-level errors were still common in their output. The researcher argues for additional pedagogical involvement to improve the effectiveness of the tools being used. Kessler (2020) takes a qualitative investigative approach with two Chinese L2 English students to gain more insight into the participant's use of technology in the writing process. Using a case study design, data sources included screen recordings, interviews, stimulated recalls, and process logs. The researchers highlight the disparity between the participants' use of digital writing tools and the educator's knowledge of what their students are using.

As an example, the participants in the case study used tools that are not primarily intended for language support in their writing (e.g., Google Search). The deficit of useful language support features (e.g., checking collocations; predictive text guidance) raises an important issue surrounding educators' Technological Knowledge (TK) as a key component of the Technological Pedagogical Content Knowledge (TPACK) framework (Niess, 2011). Any subject educator implementing technology into their pedagogy should have a comprehensive understanding of applied technologies that exist in the field.

2.3.2 Text prediction as a writing assistant

Predictive text has gained some prominence on mobile devices to help users automatically complete their messages on physically restrictive mobile keyboards. More recently, predictive text is being used in email applications such as Google's "Smart Compose" which trains itself on users' email history and attempts to autocomplete sentences in the user's natural writing style (Chen et al., 2019). Increasing attention has been given to these technologies in applied research (Parisis, 2019) even in areas outside of Second Language Acquisition (SLA). Researchers have attempted to measure the impact of such as system on users' performance. Gnacek et al., (2020) used Smart Compose with college students in a within-subjects experimental design. Their experiment, however, failed to show significant improvements to either user performance or an improvement in the users' mental load while using Smart Compose. Dizon & Gayed (2021) used Grammarly in a counterbalanced 8-week study with 31 university students and found the intelligent agent reduced grammatical errors and improved the lexical variation of participants' writing. The researchers also found promise in the predictive text component of Grammarly as a way to support EFL writing. Mizumoto et al., (2017) developed a web-based support tool for research article writing called "AWSuM" or Academic Word Suggestion Machine. The tool is based on displaying frequency-based lexical bundles to users based on preselected genres. User feedback from the researchers' pilot study was largely positive and the researchers find the word suggestions to be an important feature that should be investigated further for their pedagogical implications. Evmenova et al., (2010) examined the potential of three word prediction programs (Co:Writer, WordQ and WriteAssist) for students with learning disabilities. The participants' learning disabilities negatively impacted their ability to spell during journal writing exercises. The repeated measures study showed gains in composition rate, total output, and spelling accuracy while participants were under the treatment condition. Social validity interviews also indicated that participants enjoyed using the word prediction programs.

2.4 Methods

This pilot study used convenience sampling (Patton, 2002) to enroll ten Japanese adult students who attend a language school that is known to the researcher. The volunteer participants take supplementary English lessons once a week at the language school. All had taken the "Jitsuyo-Eigo Gino-Kentei", better known by the abbreviation EIKEN, a widely used English proficiency test in Japan. The participants self-reported their levels at EIKEN grade 2 and pre-2. These levels are equivalent to the Common European Framework of Reference for Languages (CEFR) levels of B1/A2, or a Test of English as a Foreign Language Internet Based Test (TOEFL iBT) score of 45/20. Two groups were created, and participants were randomly assigned to either Group A or Group B. A counter-balanced, changing conditions research design was used in the study. One advantage of such an approach is the ability to conduct the research with fewer participants (Howitt & Cramer, 2020).

In addition, counterbalancing the experiment conditions helps control for context and carryover effects that are seen in within-subject research designs (Field, 2013). Specifically, participants in this study were not aware they were under a treatment condition when they were using the AI KAKU tool. In addition, a one-week break was given between conditions to counteract boredom/familiarity or other performance factors. The control condition had the participants using standard word processing software (Google Docs) in a timed (30 min) and goal limited (three hundred word) writing assignment. The treatment condition had participants using AI KAKU with the same parameters. All participants were observed by the researcher during the treatment and control conditions and were asked not to use any outside (Google Translate, etc.) assistance. Each participant was seated in front of a laptop computer with the writing instrument (Google Doc or AI KAKU) ready for use before the experiment started. To familiarize the participants with AI KAKU before the treatment condition started, a 5-min training session was conducted to introduce the participants to AI KAKU and its features. The research methodology is graphically represented in Figure 8 for the two groups A and B. Writing prompts were chosen from sample independent writing tasks of the TOEFL iBT, a commonly used test of English for students wishing to enter tertiary education in the United States. Each prompt under control and treatment conditions asked the participant to agree or disagree to a problem statement. The researcher used automated text analysis to detect differences between the two writing conditions, followed by quantitative analysis of post-activity survey data.

2.4.1 Treatment tool

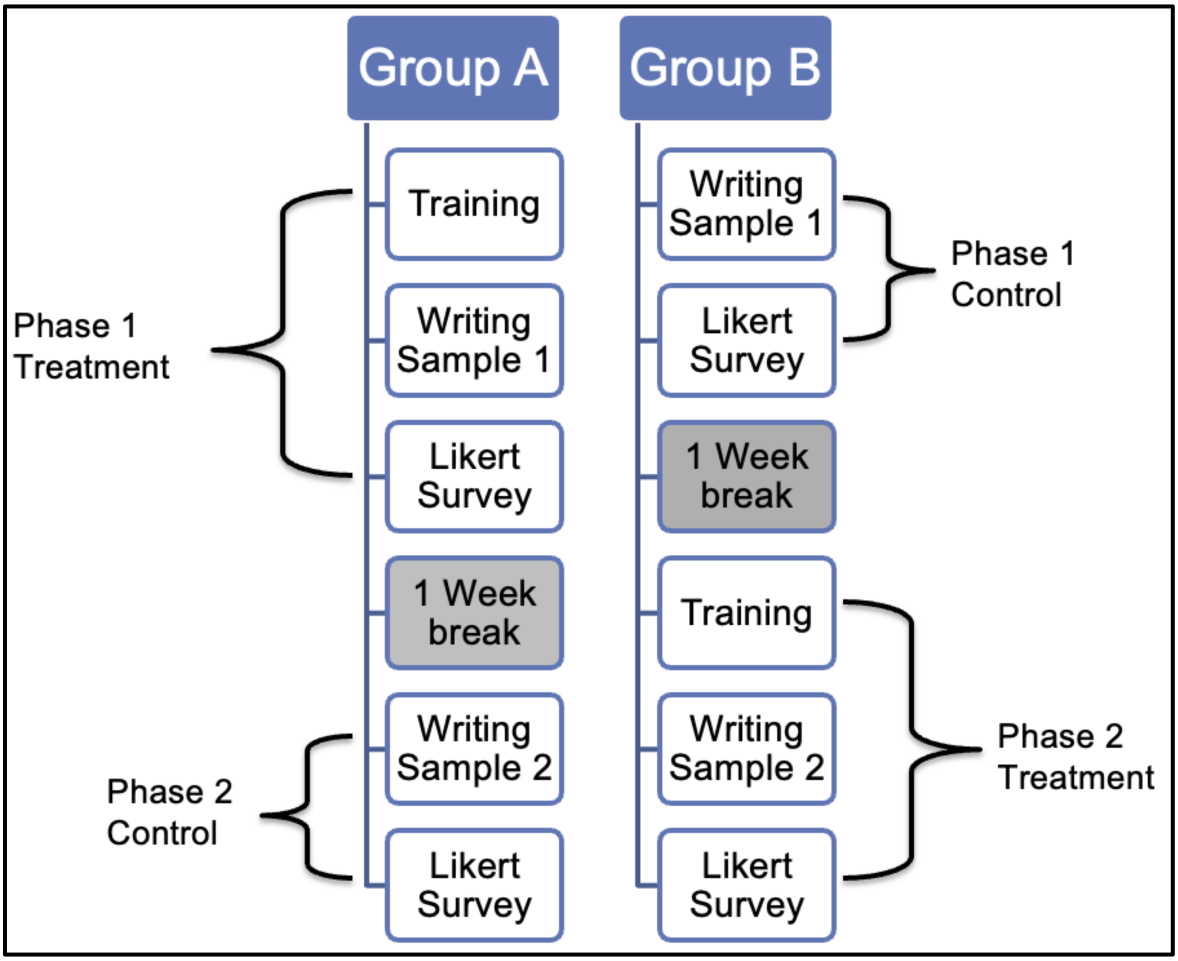
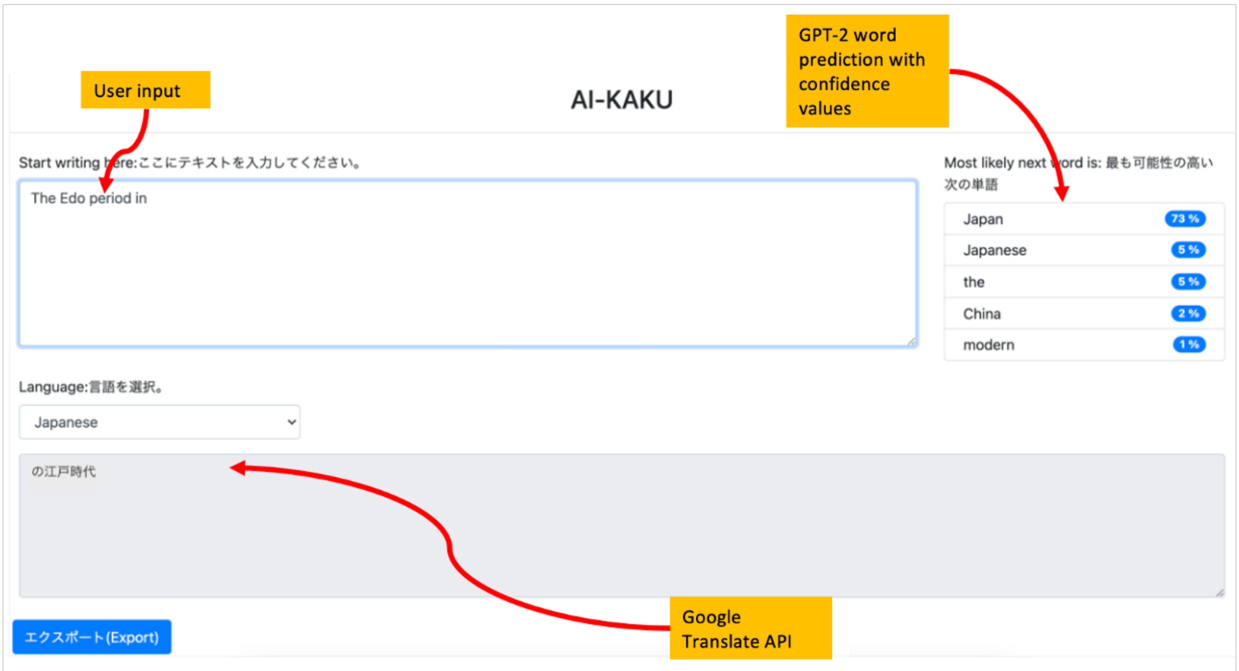


Figure 8. Research methodology used in the experiment.

AI KAKU is a web-accessible tool aims to reduce some of the cognitive load associated with the second language writing process (Nawal, 2018). The result would potentially allow EFL students to produce more and improved output than they would without assistance. AI KAKU's interface, as seen in Figure 9 is comprised of five main elements: an input field, a word suggestion engine with confidence scores, a language drop-down menu, a reverse-translate output field that translates the users' inputted English into their chosen first language, and a save/export button for users to be able to download their work. Importantly, while the word suggestions are based on user input and mimic the text prediction features seen on smart devices and other online-enabled services, the researcher has designed AI KAKU to be non-intrusive. While the user is in the writing process, they are not interrupted by automatic corrections or pop-ups in the input field. The word suggestions and updated reverse translation only appear after a 2.5s pause in typing. This ideally creates a space for user agency and discourages abuse of the text prediction feature. In addition, research has shown in translation tasks pauses of less that 2 seconds can be attributed to mechanical keyboarding issues (Muñoz Martín & Cardona Guerra, 2019). By delaying the appearance of AI KAKU's assistance to fall outside of the two-second window, psychomotor pauses in writing would not over-engage the user with assistance. As the user inputs tokens (words) into the input field, the predictive engine produces context-aware word suggestions and not simply high-frequency tokens.


AI-KAKU

Start writing here:ここにテキストを入力してください。

The Edo period in

Most likely next word is: 最も可能性の高い次の単語

Japan 73 %
Japanese 5 %
the 5 %
China 2 %
modern 1 %

Language:言語を選択。

Japanese

の江戸時代

Figure 9. AI KAKU's user interface.

The GPT-2 language model that AI KAKU uses for text prediction is based on a 345 million parameter transformer that is pretrained on a corpus of nearly 40 GB of data. This data is based on OpenAI's WebText corpus which is a curated database of web-based sources that have been filtered for quality via Reddit outbound link ratings (Radford et al., 2019). However, WebText is such a broad corpus that inherit biases/prejudices found in web-based sources manifest themselves in the model. If users input questionable or unsavory prompts, then the model does not distinguish fact from fiction, nor does it block unsavory words from appearing. In this sense, using AI KAKU in a classroom setting should be closely monitored by the instructor to prevent misuse and undesirable content being displayed to students.

One factor in AI KAKU's development that was identified as a potential distraction for the user is the machine translation engine's (Google Translate) difficulty in translating sentence fragments. The user process described in Figure 7 continually translates the inputted text into the users' first language of choice. By continuously translating for the user, it is possible that the machine translation will make obvious translation errors to the user as the translation engine cannot predict what the user *intends* to say and only translates what it has been given. The researcher suspects that users of higher L2 linguistic ability would be able to understand the mistakes the machine translation is producing and work through them. In that regard, AI KAKU has been updated to let the user turn-off the reverse translate feature if it becomes a distraction/disruption to the user. Regardless, this is a known limitation of current machine translation technologies, and the researcher would like to understand more about how lower-level users interact with this limitation in future studies.

2.4.2 Data collection

One measure of student writing ability is fluency which can be defined by different measures. This study takes the approach of defining fluency as the total words written in a given time and the syntactic complexity of that writing. This measure allows the researcher to observe any improvement or deterioration in participants' production rate and relative quality of that production in each condition. Another commonly used measure is Lexical Diversity (LD), or the range of different words used in a text. Texts with a lower LD range tend to use the same words repeatedly. LD is commonly used in second language research, and LD indices are suggestive of writing quality, vocabulary knowledge, and speaker competence (McCarthy & Jarvis, 2010). The researcher conducted text analysis of the participants' writing with the Text Inspector, a web-based text analysis tool (*Text Inspector*, 2020) to analyze the participant's writing samples. Text Inspector gives two measures of LD: vocd-D, a probability-based LD measure, and the Measure of Textual Lexical Diversity (MTLD). MTLD is especially pertinent as it is less sensitive to text length than other LD measures (McCarthy & Jarvis, 2010). In order to check the reliability of the Text Inspector tool, the researcher compared its output with the widely used koRpus package in R (Mizumoto & Plonsky, 2016) via the Langtest web interface (Mizumoto, 2015) and compared the output from both techniques. According to a Pearson correlation analysis, there was a strong positive correlation, $r\ (18) = 0.98$, $p < 0.00001$, indicating to the researcher that the Text Inspector tool is a reliable method to measure lexical diversity. A boxplot of the individual scores from both techniques are depicted in Figure 10.

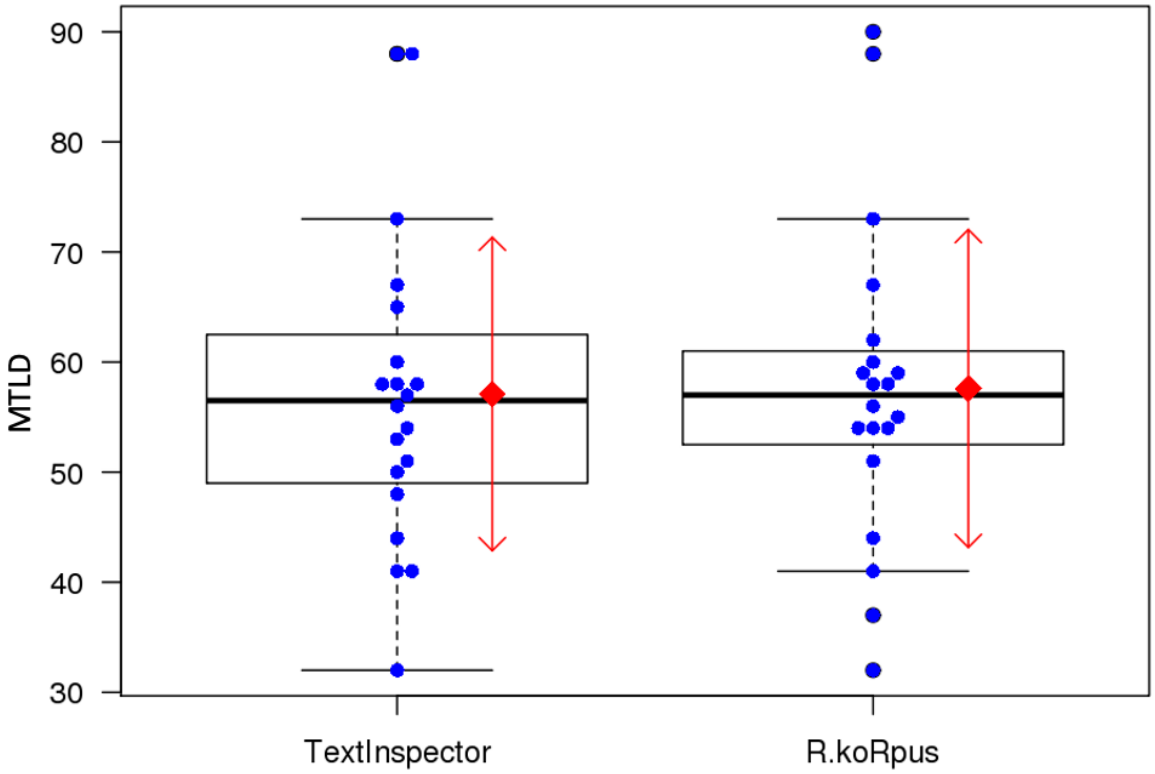


Figure 10. Scatterplot of Text Inspector MTLD scores checked against R.koRpus.

Note: Means and +/- SDs are displayed in red.

Following each writing condition, a ten-question, six-point Likert survey in both English and Japanese was given to the participants to gain some perspective on the users' perceived difficulty in completing the writing task and the participants' attitudes and perception of using AI KAKU. The answers were on a scale of 1 (Strongly disagree) to 6 (Strongly agree). Finally, the data gathered was summarized with descriptive and inferential statistics.

2.5 Results

As seen in Table 1, the mean values for vocd-D and MTLD show that the average lexical diversity was greater under the AI KAKU writing condition, perhaps indicating improved performance when under the treatment condition. However, a Mann-Whitney U test indicates that the difference was not statistically significant for neither the vocd-D analysis, U = 40, p = 0.47 nor MTLD, U = 46.5, p = 0.81. Thus, the researcher cannot claim that writing under the treatment condition (AI KAKU) leads to higher lexical diversity.

Table 1. Lexical diversity and fluency measures. N = 10.

Condition	vocd-D		MTLD		Fluency (output / 30 minutes)		Fluency (clauses / t-unit)	
	M	SD	M	SD	M	SD	M	SD
AI KAKU	63.43	17.2	57.9	13.7	144.9	44	1.64	0.42
Control	61.13	15.53	56.6	15.2	138.7	52.5	1.3	0.18

When the fluency measure is observed, we can see participants writing under the AI KAKU condition could output a greater number of words than the control condition. In addition, the results show a lower SD under the AI KAKU condition, indicating performance under the treatment condition is more clustered around the mean than the control condition. A Mann-Whitney U test for fluency production, however, shows that performance between each condition was not significant, U = 38.5, *p* = .38. On the contrary, fluency considering clauses produced per t-unit shows a positive significance with U = 21, *p* = .0315. Cohen's d calculation for effect size resulted in 1.05 which according to Plonsky & Oswald (2014), indicates a large effect size. This demonstrates to the researcher that the participants writing under the AI KAKU condition were able to produce sentences with more "sentence fluency". Text with a higher ratio of clauses per t-unit shows an ability to communicate complex ideas more fluently (Beers & Nagy, 2009), as texts with lower ratios are distinctly simpler in composition and content density.

2.5.1 Survey data

Out of the ten questions used in the study's survey to participants, the researcher will highlight two of the most pertinent. Concerning perceived effort, "Q1: It took a lot of effort to complete the writing task" and "Q2: It was difficult to express my ideas in English" are two offline measures the researcher created to gain insight into the participants' perceived mental effort while writing under each condition. To measure for internal consistency with the Likert instruments used, Cronbach's Alpha was calculated to be α = 0.857, which according to Bougie and Sekaran (2019), an α > 0.8 is considered to be "good" consistency of the instruments used in this study. The descriptive results in Table 2 show less perceived effort to complete the writing and more ease in the ability of the participant to express themselves in English when writing under the treatment condition.

Table 2. Measures of perceived mental effort. N = 10.

Condition	Q1		Q2	
	M	SD	M	SD
AI KAKU	5.1	0.9	5.1	1.1
Control	5.6	0.7	5.6	1.4

However, in testing for significance in the survey results, a Mann-Whitney U test resulted in a U-value of 33.5 for Q1 and 37.5 for Q2. The critical value of U at $p < 0.05$ results in 23. Therefore, the results for both Q1 and Q2 are not significant. The researcher believes this could be due to the participants' lack of training with the treatment tool. Introducing an unfamiliar technology to a user/classroom should be done in a rigorous and careful manner (Johnson et al., 2016).

Further, more technology acceptance measures were analyzed, as seen in Table 3. When asked "Q3: How many times did you use the word suggestions?", the average response was 3–4 words on a scale of 1–10.

Table 3. Usage frequency and ease of use measurements. N = 10.

Condition	Usage frequency	Ease of use
M	3-4	5.1
SD	0.7	0.7

Concerning ease of use, participants were asked "Q4: Learning to use AI KAKU was easy for me.", the reaction from the participants about using AI KAKU was largely positive, with 95% of the participants indicating affirmative responses on the 6-point Likert scale.

2.6 Discussion

Syntactic complexity (in terms of clauses / t-unit) was the only factor that showed AI KAKU having a positive significance compared to the control condition based on the data collected from this study. While the researcher hypothesized AI KAKU would be able to show improvements on other measures of writing performance (LD, production rate), the results were not in line with the researcher's expectations. Several confounding factors (e.g., lack of training on AI KAKU, small sample size) could have influenced the non-significant outcome. In addition, this pilot study only examines student writing quality via machine assessment measures. The researcher believes human assessment of student writing can give a more holistic understanding of participant performance. That is not to say machine assessment is entirely lacking.

Research has shown evidence of positive correlations between human and machine assessment, as studies by Powers et al., (2000), Shermis et al., (2010), and Landauer (2003) have demonstrated. However, those studies also indicate areas where human assessment can fill the gaps that automated machine scoring leaves. Machine assessment can miss writing features that are not explicitly tested for (Powers et al., 2000) or tend to weigh surface errors (grammar, spelling) more than contextual mistakes (McCurry, 2010). Obvious organizational errors that human assessors easily identify can also remain undetected by machine algorithms (Patterson, 2007).

Several factors could have contributed to the insignificant LD and production rate results, including the potentially low usage of the word suggestion engine (average usage was 3.5 words/145 = 2% of total average words written), lack of experience and hands-on time with using AI KAKU, and possible unintended negative influence word suggestions might have on the participants. However, the students writing under the AI KAKU condition did produce more words with a lower variance between the participants, potentially indicating AI KAKU is helping lower-level participants more, equalizing their performance to higher-level writers in the study.

The offline measurements of perceived effort were also inconclusive. Again, the researcher sees the necessity of further training and practice with AI KAKU for the participants to maximize the system's benefits. We cannot claim using AI KAKU in this pilot study reduced the participants' perceived mental effort during their writing process. In addition, a confounding factor of participants' familiarity with typing on a keyboard in English was not considered in this study. While the researcher' casual observation did not detect significant typing speed differences between the participants, future studies about the potential effectiveness of AI KAKU should include typing words per minute (WPM) as one variable to consider. Further analysis from an expanded study is necessary to gain more insight into the usage of AI KAKU and its potential impact on L2 writing proficiency.

2.7 Conclusion

This pilot study attempted to measure the impact on L2 writing when using a newly developed writing assistant called AI KAKU. While inferential analyses results between AI KAKU and the control condition only show significance along one dimension (clauses / t-unit), the descriptive statistics point to the tool's potential usefulness. The researcher feels encouraged by the results of this pilot study and decided to conduct a larger scale study with more participants and additional assessment instruments applied. Predictive text and AI-based learning agents as a whole are a growing trend in education (Arnold et al., 2020; Fung, 2010; Waldron et al., 2017), and the researcher believes these warrant additional development and research into their application.

Future studies using AI KAKU or other predictive text digital writing tools should expand upon this pilot study to gauge the impact they might have on student writing. Some key considerations would be how the word suggestion and reverse translations features might introduce additional "noise" to EFL students while they are writing, potentially degrading their performance. In addition, further follow-up with participants on how they used the tool and if there are features they found useful or features they found to be counterintuitive.

As mentioned earlier in this chapter, the instructors' TPACK knowledge is a key factor in introducing new technology into the classroom. This research does not address this aspect as the tool being developed is still highly experimental. However, if a wider roll-out of a writing assistant such as AI KAKU is to be considered, educator/instructor training on the tool should be considered as an essential component to the introduction of the tool to the classroom. Teaching students how to appropriately use technology in the learning process is something that requires educators to be present with the students as they are engaging with the technology. Existing studies on other Computer Assisted Language Learning (CALL) technologies such as Google Translate have identified the importance of teacher training and knowledge (Benda, 2013; Urlaub & Dessein, 2022).

To build on the success of this pilot study, the researcher decided to conduct another study with a larger group of participants. Examining aspects of the participant's cognitive processes and how AI KAKU impacts participants of varying English proficiency were considered important aspects to get a stronger grasp of how the writing assistant interacts with EFL users. This pilot study was conducted live in person with students studying at an English language school. However, due to the COVID-19 pandemic, it was determined that the best way to recruit more participants was to hold the experiment on an online platform. edX was chosen as the online platform of choice as it allows for random group assignments and timed tasks for writing tasks. The next chapter describes the expanded experiment which is similar to the pilot study in research design but with expanded analysis on the treatment tool's impact on participants.

Chapter 3 [6] Examining the equity of a novel intelligent writing assistant as English language support via cognitive load and writing quality measures

Keywords: L2 writing, Matthew effect, computer-mediated feedback, CALL, CALL equity, AI agent, Cognitive load

3.1 Background

This chapter details the second experiment that expands upon the pilot study described in Chapter 2. Encouraged by some of the results from the pilot study, the researcher decided to conduct another experiment with more participants and delve deeper into analyzing the participants written output and their cognitive load while doing the task. While it uses a similar counter-balanced research design, the researcher expanded the number of writing tasks (from 2 in the pilot study to 4 in this study) to gain more data points on each participant when they are under both control and treatment conditions. In addition, the researcher wanted to investigate performance gaps between higher proficiency EFL students and EFL students with lower proficiency. A common issue with technologies that are introduced in the classroom is how they impact students of different proficiency levels. Lastly, the researcher employs the Ayres (2006) and Paas (1992) cognitive load instruments to measures the participants overall and intrinsic load under each writing condition. The results of this study were published in the post-conference proceedings of an international CALL conference.

[6] Parts of this chapter have been published under: Gayed, J.M., Carlon, M.K.J, & Cross, J.S. (2022). The Matthew effect in CALL: Examining the equity of a novel intelligent writing assistant as English language support. In J. Colpaert, Y. Wang, & G. Stockwell (Eds.), *Proceedings of the XXIst International CALL Research Conference* (pp. 80–93). London: Castledown Publishers. https://doi.org/10.29140/9781914291050-12.

3.2 Introduction

Recent advancements in natural language processing (NLP) research have brought new opportunities to apply these cutting-edge technologies to computer-assisted language learning (CALL). For instance, grammar and spellcheck applications have become mainstream tools for English as a Second Language (ESL) / English as a Foreign Language (EFL) educators (Chun et al., 2021; Park, 2019). Thanks to these recent advances in NLP, simple rule-based systems such as grammar checkers have added intelligent context-sensitive features to make the feedback they give users better reflect individual writing styles and intended output. This allows for greater user autonomy and the potential for improved output (Gayed, Carlon, Oriola, et al., 2022) creating an environment for better learning and learner agency

An issue that CALL practitioners should be aware of is the potential for the Matthew effect to influence the learning outcomes of their students. This effect, for example, can be seen when children fall into different reading levels—stronger readers develop faster and weaker readers fall further behind (Stanovich, 2009). The Matthew effect in language learning can be exacerbated when educational technologies are introduced. The edtech Matthew effect manifests when the more affluent learners benefit more from educational technologies due to differences in technology and human support access, making inequalities in education bigger (Reich, 2020). As such, CALL practitioners should be cognizant of which learners are receptive to their interventions, both technology and non-technology-related, to prevent disadvantageous positions from being compounded.

The researcher believes this is a rarely identified phenomenon in CALL and this researcher's focus on the effect to be novel. In a sense, the researcher is proposing the term **"CALL Matthew effect"**. As mentioned earlier, the accumulated advantaged effect has received wide attention in several fields, (educational psychology, educational technology, economics, political science, etc.,) yet the research area of computer-assisted language learning has not given the effect much attention. The researcher believes this to be a point that deserves more attention and research. As new educational technology tools are introduced into practice the accumulated advantage of the user should be a dimension to considered when measuring the tool's impact on the user.

This paper focuses on a digital writing assistant and its potential impact on EFL writing. Most current word processing platforms were not built with EFL users in mind and generally give feedback to the user (via grammar and spell-check) only after the user has entered some input into the system. The researcher has developed a digital writing assistant with a basic framework conceptualized around EFLs. Given that this newly developed writing aid has the potential to influence student writing, the researcher explored the equity of using the tool with students with different English skill levels.

Research Question

This paper examines the intersection of CALL and educational psychology by probing the CALL Matthew effect on the participants of this study. The research questions we are addressing include:

1. How much improvement can be detected from different level EFL participants while under experiment and control conditions?
2. How prevalent is the CALL Matthew effect among the participants?

Writing proficiency has often been cited as a goal of second language education (Alisaari & Heikkola, 2016) and certainly the goal of language learners themselves. This study introduces a novel digital writing assistant that can potentially aid EFL students in achieving that goal. It is worth noting that even though research has shown smart digital devices have to potential to harm a person's cognitive function (e.g., memory recall) (Tanil & Yong, 2020), we can find little argument for going back to life without smart devices. As such, the removal of smart agents from education is an unpractical approach, yet educators and developers should be more aware of the potential negative impacts smart agents may have on learners.

3.3 Related Works

3.3.1 EFL challenges

There has been much research on the topic of digital tools and their impact on writing. More so, from a CALL perspective, digital mediums have been studied for their possible influence on language learners' ability to write in a second language (L2). Research has shown that writing in a second language is more difficult than writing in one's first language (L1) (Javadi-Safa, 2018; Silva, 1993), and not having strong English writing skills can adversely affect academic performance (Tan, 2011).

A longitudinal study by Laufer (1994) examined the lexical development of advanced second language learners' writing. When the participants' lexical frequency and lexical variation were analyzed, the researcher found only marginal improvements to the former, no improvement in the latter, and no correlation between the two elements were identified. Alfaqiri's (2018) study on Saudi Arabian EFL students investigated the writing difficulties and challenges participants experienced. Data from 114 participants showed that metacognitive strategies were key to improved writing. Additionally, participants' struggle with grammar was identified as a major factor inhibiting higher-level writing production.

Thus, EFL challenges come from at least two fronts: having sufficient lexical and grammatical ability to execute. A common element that restricts L2 writing fluency is the inability to retrieve lexical elements (Schoonen et al., 2009) and having enough cognitive resources to make way for metacognitive strategies that can improve their writing. These two challenges present a feedback loop. For L2 writers, much of the cognitive load comes from translating L1 thoughts to L2 (Nawal, 2018). To be able to think directly in L2 as opposed to translating from L1 and thus optimize cognitive load, writers must have sufficient grammar knowledge and vocabulary to begin with. Retrieving somewhat familiar but not frequently used vocabulary can lead to the tip-of-the-tongue phenomenon which can be frustrating and impede production if not properly resolved (D'Angelo & Humphreys, 2015). To be able to succeed in highly cognitive tasks, one should be able to offload some of the cognitive efforts to the environment whenever practical (Hollan et al., 2000). For L2 writing, being able to produce is arguably more critical than being able to fix grammatical errors, thus these ancillary tasks are good candidates for tool support.

3.3.2 Automated Writing Evaluation

Automated writing evaluation (AWE) systems have gained prominence in digital writing as the sophistication of the feedback available has improved with the integration of NLP technologies. These can be built-in systems (e.g., Microsoft's Editor) or independent software packages (e.g., Grammarly) that can be integrated into existing word processors. AWEs are also slowly becoming popular as language learning support tools. Sevcikova's (2018) study of college-aged participants using AWEs for writing found that the systems can improve language learning. More importantly, students showed greater confidence and motivation while using an AWE. Looking into the accuracy of an AWE and how it is compared to human-based assessment, Dodigovic and Tovmasyan (2021) found that the AWE could largely reproduce the quality of human raters when it came to detecting and remediating errors. However, they found certain errors (e.g., coordination, subordination, and relative clauses) were often undetected by AWEs, leading the researchers to the conclusion that AWEs cannot be solely relied upon for evaluation and assessment. Additionally, Zhang's (2020) study on students' use of an AWE showed that engagement with AWEs differed based on the student's English level. Higher-level students were more cognizant of the revision stage of writing and were able to use the feedback they were given more effectively.

3.3.3 CALL Matthew effect

Confounding factors are commonly exposed and elucidated in second language acquisition research. However, one confounding factor that the researchers found to be less commonly highlighted in CALL literature is the presence and impact of the Matthew effect on learning outcomes (Lamb, 2011). This effect, as seen in Penno et al., (2002) study of children's vocabulary acquisition, was seen to be unavoidable across treatment conditions. In the study, treatment interventions were not enough to overcome the effect as higher-level students made greater vocabulary gains than lower-level students. Ngiam and See (2017) examined the link between e-learning CALL applications and music. In their research, the Matthew effect was identified as one negative factor where wealthier students, possessing more cultural capital, were able to perform better than poorer students who did not possess the same level of capital. The poorer students then found themselves in a downward negative spiral, with little awareness of how to improve.

Fortunately, the EFL Matthew effect can be mitigated. For instance, Messer and Nash (2018) were able to minimize the EFL Matthew effect in young English speakers by using visual mnemonics in a CALL study. The researchers found their computer-assisted intervention was effective in improving vocabulary acquisition in the participants. However, as previously mentioned, using the current state-of-the-art AWEs may not be conducive to minimizing the Matthew effect. Even without the usual culprits of the edtech Matthew effect (e.g., technology access and human support), introducing technology can increase the Matthew effect just because the learners do not have sufficient skill to make sense of the feedback they are given by the technology. We will be referring to the EFL Matthew effect magnified by technology as the **CALL Matthew effect**.

3.4 Methodology

3.4.1 Treatment tool – AI KAKU

Advancements in natural language processing and machine learning have led to the development of more sophisticated intelligent writing assistants which offer synchronous feedback to the writer compared to traditional text editors (Frankenberg-Garcia, 2020). In addition, there has been a large volume of research concerning the impact of those digital tools on the writing process (Ashton, 1999; Oh, 2020; O'Regan et al., 2010). AI-assisted writing technology is commonly seen in the form of next-word prediction on smart mobile devices and in some operating systems. Increasingly, next word prediction is becoming a feature available in commonly used word processors such as Google Docs and Microsoft Word. This nextgeneration type of writing assistance is presented to the user in addition to spelling and grammar correction that users have traditionally experienced. In addition, several applications give further feedback to the user in terms of word suggestions, style feedback, and formative assessment (e.g., Grammarly, Microsoft Editor).

Unfortunately, those tools are primarily aimed at L1 writers and were not intended to assist L2 users with their compositions. Market forces largely dictate software development and there is less demand for digital tools that are intended for the non-native level English user. This in turn translates to a paucity of literature about the effectiveness of said tools when EFL students are using them. This paper examines a digital writing assistant called "AI KAKU." The name is a take on the Japanese word "書く, kaku," which translates to "to write" in English.

The application was created to assist L2 writers as they are producing written text. The web-accessible artificial intelligence-based writing assistant tool aims to reduce some of the cognitive load that is associated with the second language writing process (Nawal, 2018), allowing users the capability to produce richer, more complex writing than they would without assistance. AI KAKU's interface, (see Figure 9 shown in Chapter 2) is comprised of five main elements: an input field, a word suggestion engine with confidence scores, a language drop-down menu, a reverse translate output field that translates the users' inputted English into their chosen first language, and a save/export icon for users to be able to download their work.

The framework behind AI KAKU outlined in the previous work of Gayed et al., (2022), will be briefly described here. The next-word prediction is implemented using AllenNLP application programming interface (API) based on Generative Pre-trained Transformer 2 (GPT-2) and the translation is powered by Google Translate API. Only English input is accepted to force thinking in the L2 and default browser grammar and spelling checkers are not blocked. To prevent tool abuse and possible distraction to the writing process, the translation and next-word predictions are only displayed after a 2.5-second delay.

3.4.2 Experimental design

The researcher utilized a counterbalanced research design with Japanese EFL participants (n = 90) who are studying English at private language schools. The potential effects on student writing while using the AI KAKU application are compared to a control condition without writing assistance. A counterbalanced design minimizes the confounding factors arising from treatment orders and allows all the participants in the study the opportunity to be under the treatment condition. Similar research designs have been employed in L2 research, as seen in Wang's (2019) study of vocabulary recall performance by Chinese students in a university setting or Dizon and Gayed's (2021) study examining Japanese university students using Grammarly as a treatment tool.

The participants were asked to self-report their Test in Practical English Proficiency (EIKEN) scores. The EIKEN test is the most widely used English testing program in Japan. The exam has a range of seven levels from Grade 5 to Grade 1. Grades 2 and 1 have subgrades (2.5 and 1.5). Grade 1 is the highest-level grade in the exam, being the equivalent of a TOEFL iBT score of 100/120 and Common European Framework of Reference for Languages (CEFR) level C1. Given that our participants are adult learners in optional professional development schools, their economic conditions and adeptness with technology may not be as varied as students in basic education. One way to analyze the equity of educational technology is to compare the performance of low-performing learners with that of high-performing learners (Doroudi & Brunskill, 2019). For this study, the participants were grouped into HIGH (EIKEN 1.5, 2) MIDDLE (EIKEN 2.5), and LOW (EIKEN 3, 4). No participant reported EIKEN level 1 or 5.

To make the experiment available and accessible to as many participants as possible, the researcher used the edX platform to host the experiment. After registration, the participants were randomly assigned to either start with the control condition or the treatment condition. The participants were not aware of which condition they were under to prevent any spillover effects from the counterbalanced research design. The flow of the experiment is shown in Figure 11 and a screenshot of the edX platform used is shown in Figure 12.

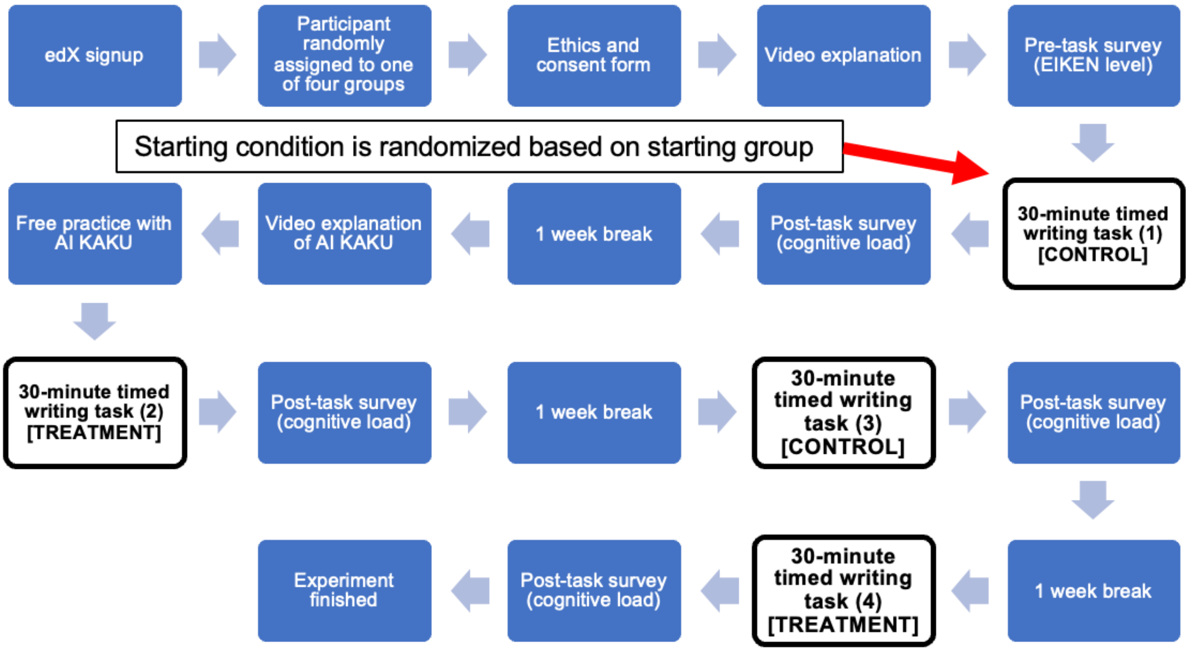
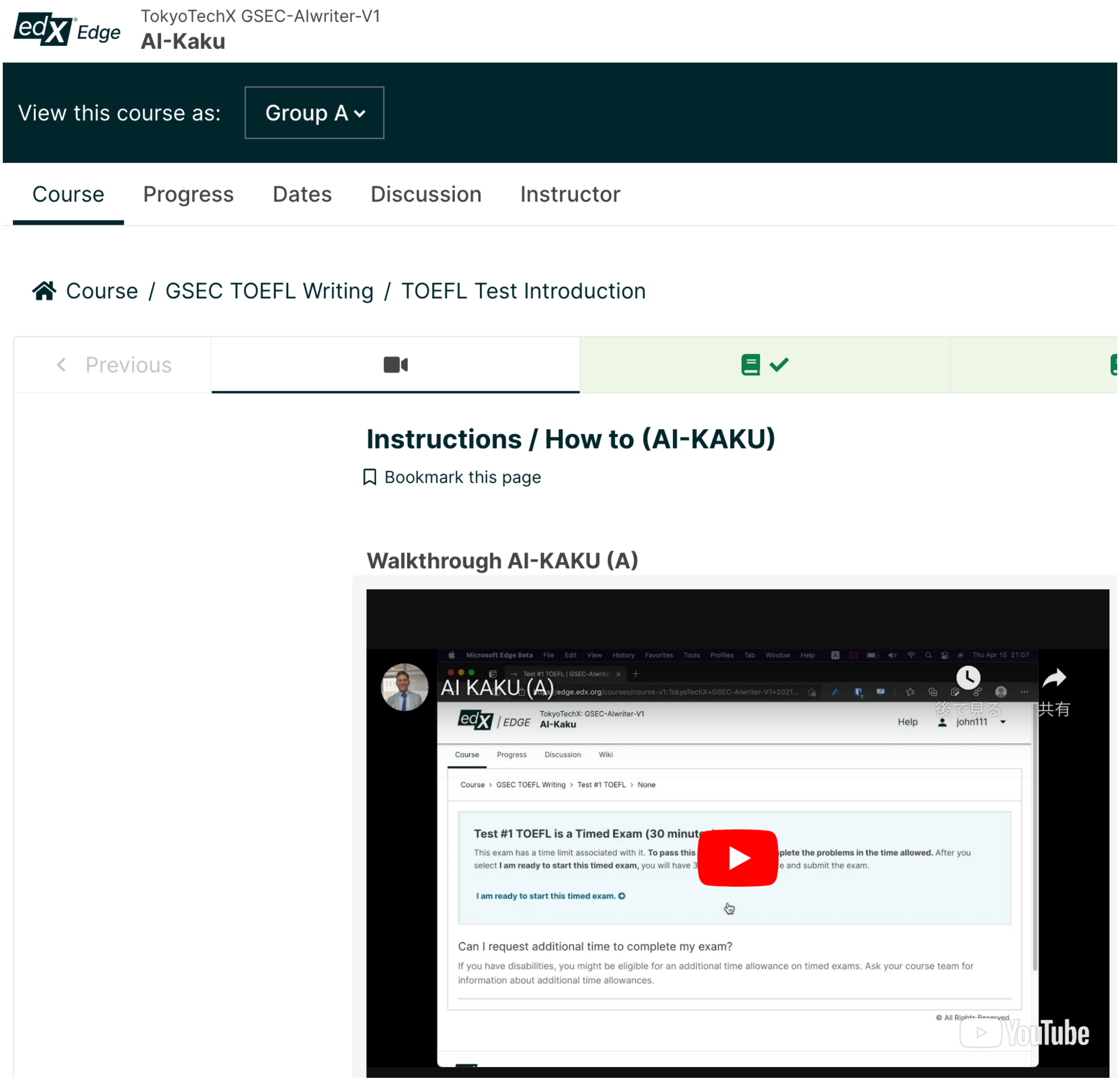

edX signup
Participant randomly assigned to one of four groups
Ethics and consent form
Video explanation
Pre-task survey (EIKEN level)
Starting condition is randomized based on starting group
30-minute timed writing task (1) [CONTROL]
Post-task survey (cognitive load)
1 week break
Video explanation of AI KAKU
Free practice with AI KAKU
30-minute timed writing task (2) [TREATMENT]
Post-task survey (cognitive load)
1 week break
30-minute timed writing task (3) [CONTROL]
Post-task survey (cognitive load)
1 week break
30-minute timed writing task (4) [TREATMENT]
Post-task survey (cognitive load)
Experiment finished

Figure 11. Flow of the experiment and writing conditions.



Instructions / How to (AI-KAKU)

Walkthrough AI-KAKU (A)

Test #1 TOEFL is a Timed Exam (30 minut

This exam has a time limit associated with it. **To pass this** **plete the problems in the time allowed.** After you select **I am ready to start this timed exam,** you will have 3 e and submit the exam.

I am ready to start this timed exam.

Can I request additional time to complete my exam?

If you have disabilities, you might be eligible for an additional time allowance on timed exams. Ask your course team for information about additional time allowances.

Figure 12. edX platform used for experiment.

After finishing the writing task, the participants were asked to complete a Likert survey that was displayed to the user in both English and Japanese. Perceived usefulness, cognitive load measures, and the number of times word suggestions were used during writing were some of the data points obtained through the survey responses. The participants were randomly assigned to one of four groups as seen in Table 4.

Table 4. Experiment's counterbalanced design.

GROUP A	GROUP B	GROUP C	GROUP D
Topic 1 (Treatment)	Topic 3 (Control)	Topic 3 (Treatment)	Topic 1 (Control)
Post exercise data collection via Likert questionnaire			
Topic 2 (Control)	Topic 4 (Treatment)	Topic 4 (Control)	Topic 2 (Treatment)
Post exercise data collection via Likert questionnaire			
Topic 3 (Treatment)	Topic 1 (Control)	Topic 1 (Treatment)	Topic 3 (Control)
Post exercise data collection via Likert questionnaire			
Topic 4 (Control)	Topic 2 (Treatment)	Topic 2 (Control)	Topic 4 (Treatment)

3.4.3 Lexical quality measurements

As for the writing topics the participants were prompted with, four were chosen from a publicly available database of the Test of English as a Foreign Language (TOEFL) administered by Educational Testing Service (ETS). TOEFL is a commonly used English language test administered to foreign students wishing to enter tertiary education in the United States. The researcher chose the "Independent Writing Task" from the TOEFL test, and all the questions chosen in the experiment asked the writer their opinion on commonly discussed social topics. By choosing a standardized test source for our writing prompts, the researcher could avoid weighted difficulty differences between writing prompts. In other words, all the prompts given to the participants have been validated to be of the same difficulty. The instructions asked participants to write at least three hundred words within the thirty-minute time limit they were given.

To gain objective measurements of writing quality, the researcher used machine assessment to measure three factors. Laufer and Nation's (1995) Lexical Frequency Profile (LFP) examines the word frequencies in a sample text. Less frequent words identified in the British National Corpus (BNC), or the Contemporary American English Corpus (COCA) are considered to be more "advanced" than high-frequency words. Specifically, the LFP measures the ratio of words written beyond the 2000-word frequency level. Lexical Diversity (LD) is another commonly used measure in second language research. LD identifies the range of different words used in a text. Texts with a lower range tend to use the same words repeatedly, indicating a lack of lexical development and sophistication. LD indices are suggestive of writing quality, vocabulary knowledge, and speaker competence (McCarthy & Jarvis, 2010). Finally, tokens are calculated to measure the rate of production. As an L2 writer progresses in proficiency, their linguistic retrieval speed improves and thusly their ability to turn ideas into written text also improves (Palviainen et al., 2012).

3.4.4 Cognitive load measurements

Cognitive load, or a person's working memory capacity, is often measured in educational research as a means to gain insight into learning efficiency and efficacy (Clark et al., 2011). This capacity is commonly measured by using offline measurements (e.g., Likert surveys), dual-task measurements (e.g., concurrent load while completing a task), and physiological measurements (e.g., heart rate). Furthermore, cognitive load can be separated into three sub-measurements: intrinsic load, or the relative difficulty of the task at hand; extraneous load, or external load (e.g., noise and distractions) that is caused by elements outside of the problem space; and germane load, or the load associated with the ability to bridge the problem space with existing knowledge.

This study employs offline measurements based on widely used cognitive load rating scales used in educational research. The Paas survey measures overall cognitive load via a nine-point Likert instrument (Paas, 1992). Responses range from 1 [very, very low mental effort] to 9 [very, very high mental effort]. To gain further insight into AI KAKU's potential influence on participants' writing proficiency, the intrinsic load was also measured via a nine-point Likert instrument (Ayres, 2006). Considering one of the researcher's goals while developing AI KAKU was to reduce the problem space for L2 writers, measuring intrinsic load gives the researcher a more granular look into the ability of AI KAKU to address that cognitive burden.

3.5 Results and Discussion

3.5.1 Overall effects

In total, 360 responses were obtained (180 under each writing condition) over the five weeks the study was conducted. After filtering for complete responses, data from 90 respondents were included in this study. Out of the 90 participants, 67 indicated their EIKEN level, data from these participants was used to investigate the CALL Matthew effect. Table 5 shows the breakdown of the respondents according to group assignment, gender, and reported EIKEN levels.

Table 5. Demographics of participants.

Variables	Levels	Values	Percentage
Group	A	26	28.88%
	B	21	23.33%
	C	20	22.22%
	D	23	25.55%
Gender	Male	34	37.77%
	Female	56	62.22%
EIKEN	1.5	2	2.99%
	2	22	32.84%
	2.5	29	43.28%
	3	13	19.40%
	4	1	1.49%

3.5.2 Lexical measures

A paired t-test was used to examine the difference between the control and treatment writing conditions. As seen in Table 6, the measures LFP and LD did not demonstrate statistical significance while the measure of Tokens is significant at *p* .004, *d* = 0.2 albeit according to Cohen's d measure, this is conventionally considered a "small" effect size.

To gain more insight into the significant result from the Tokens measure, a scatterplot was plotted, seen in Figure 13, showing the improvement participants demonstrated while under the treatment condition. While under the same writing constraints, the treatment condition allowed participants to produce longer texts, while the lexical diversity and lexical sophistication measures of their writing were largely the same.

Table 6. Lexical differences between writing conditions.

	Tokens	*t*-test	LFP	*t*-test	LD	*t*-test
Control	156.7 (52.3)	t = -2.8,	0.1 (0.04)	t = -0.19,	61.7 (18)	t = -0.37,
Treatment	167.8 (63.2)	df = 179,	0.1 (0.04)	df = 180,	62.2 (18.1)	df = 180,
		p = .004		p = .84		p = .7

Mean values. SD values are shown in ().
Note: Control is blue, and Treatment is red. Vertical lines show the mean. Higher values indicate more load.

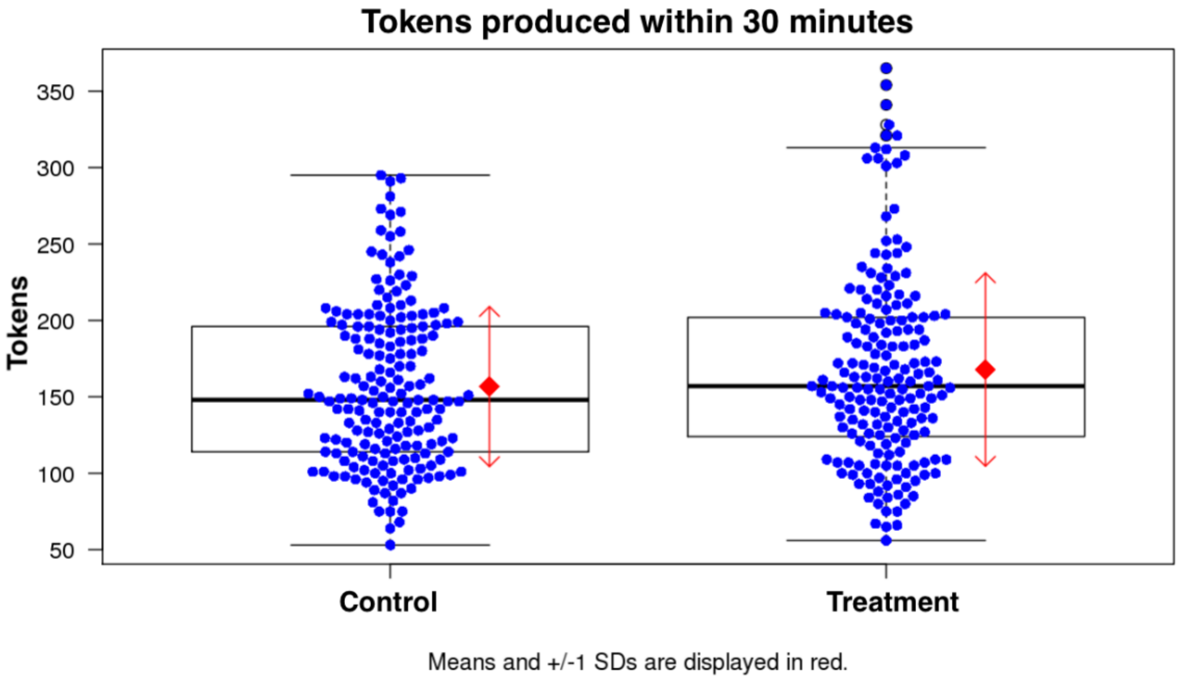


Figure 13. Scatterplot of token production under each condition.

3.5.3 Cognitive load measures

Since this study takes survey questions out of the Paas (1992) and Ayres (2006) inventory to measure cognitive and intrinsic cognitive load, the researcher needed to confirm the reliability of the questions used in this study. The value for Cronbach Alpha for the survey items was α = 0.57, which can be interpreted as “acceptable” according to Taber's (2018) meta-analysis of Alpha reliability measures. Results summarized in Table 7 show that while the difference in overall participant cognitive load did not show statistical significance, the intrinsic load was lower and significant at *p* .03, *d* = 0.13; a “small” effect size (Plonsky & Oswald, 2014). A histogram (see Figure 14) of the intrinsic load measure indicates that when participants were writing under the treatment condition, they experienced less perceived difficulty with the writing task at hand.

Table 7. Cognitive and intrinsic load differences.

	Cognitive load	*t*-test	Intrinsic load	*t*-test
Control	7.0 (1.4)	t = 0.7, df = 179,	6.3 (1.39)	t = -1.87, df = 179,
Treatment	6.9 (1.3)	p = .4	6.1 (1.48)	p = .03

Note: higher values indicate more load. Mean values. SD values are shown in ().

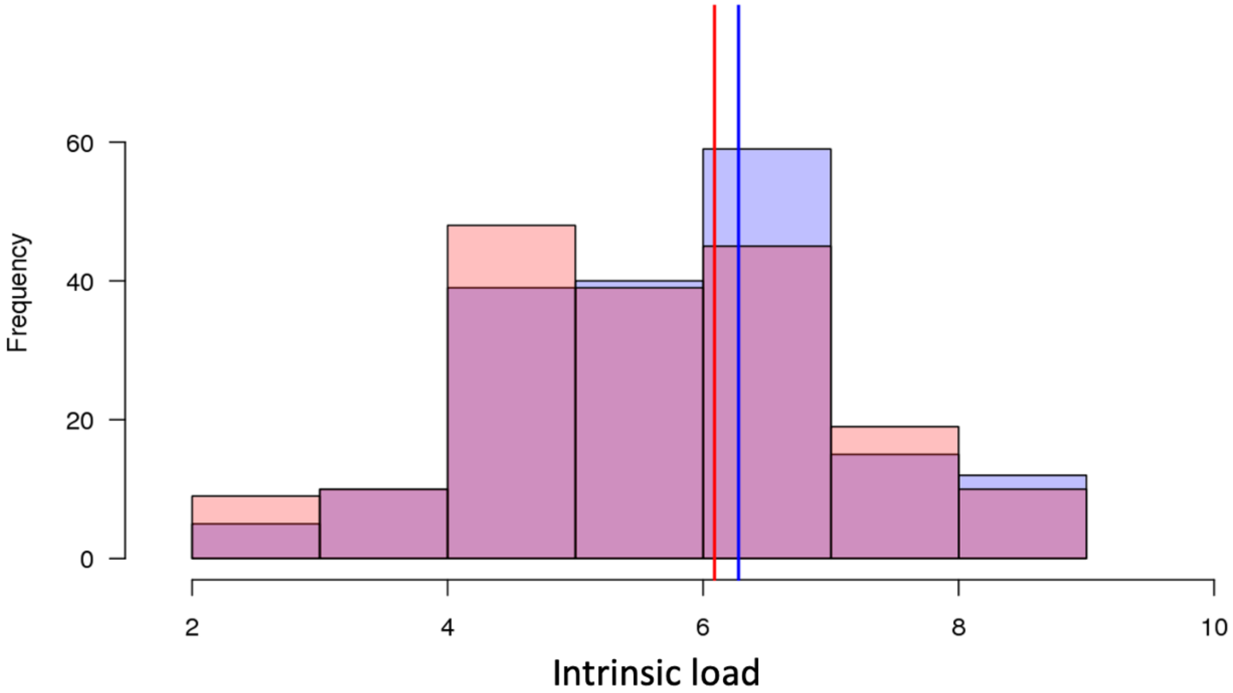


Figure 14. Impact of control and treatment on intrinsic load.

Note: Control is blue, and Treatment is red. Vertical lines show the mean. Higher values indicate more load.

Two significant outcomes from the experiment show us that participants were able to produce more tokens and felt the inherent difficulty of the writing task was less while they were using the writing assistant (AI KAKU). These results allow the researcher to approach the second research question regarding evidence of the Matthew effect and how the writing assistant impacted participants at different skill levels.

3.5.4 CALL Matthew effect

As mentioned earlier, participants were grouped into HIGH, MIDDLE, and LOW clusters (n = 67) based on their reported EIKEN levels. To investigate any evidence of the CALL Matthew effect between them, their writing performance and cognitive load measures were examined first across all the EIKEN levels and then across the three levels prescribed by the researcher. The box plots in Figure 15 show the distributions of cognitive load, intrinsic load, lexical frequency, lexical variation, and tokens for each of the assigned EIKEN clusters. The boxplot whiskers extend up to $1.5 * IQR\ /\ sqrt(n)$, where IQR is the interquartile range (the difference between the values at the first quartile and third quartile) and n is the data count. This convention was posited to represent data with a 95% confidence interval when comparing medians for most cases (McGill et al., 1978). Data beyond the whiskers are taken to be the outliers.

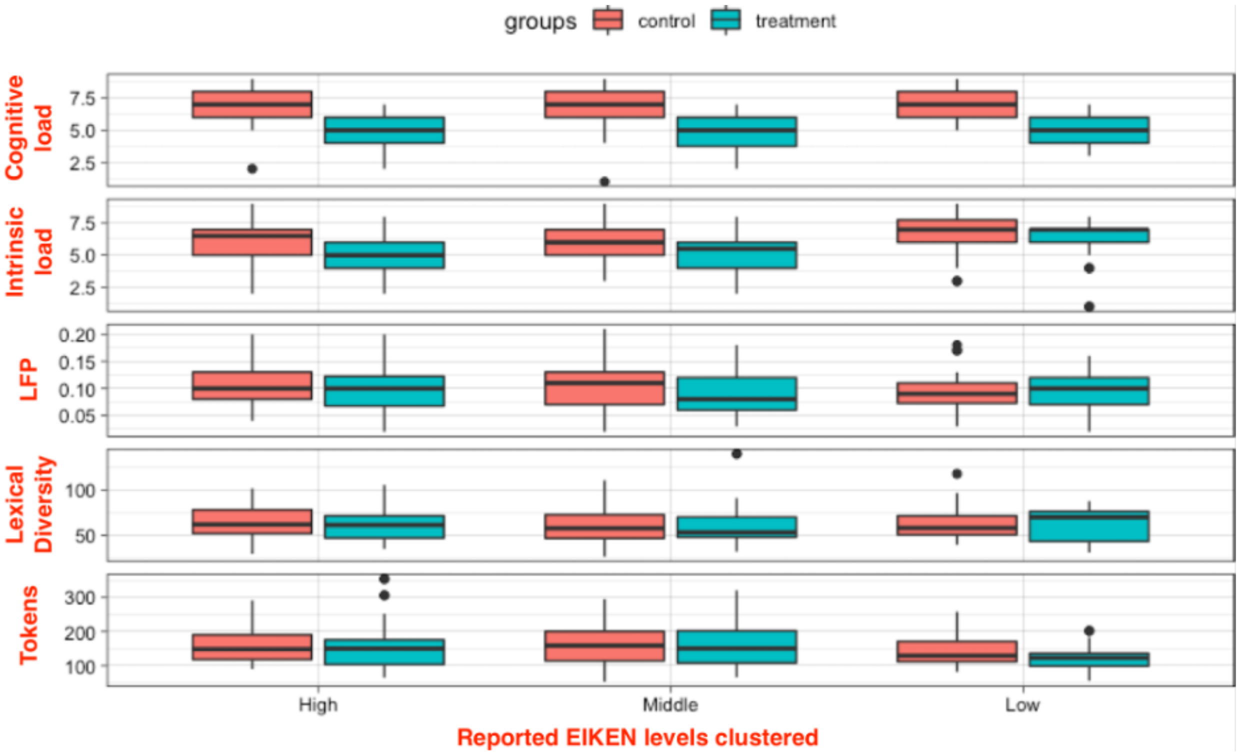


Figure 15. Participant performance when grouped into three clusters.

The figure shows cognitive load decreasing similarly across all three groups; intrinsic load, however, appears to decrease more for the HIGH and MIDDLE clusters, with the LOW cluster experiencing a similar load in both control and treatment conditions. Lexical frequency and lexical variation, interestingly, appear to be negatively influenced by the treatment condition. While the paired *t*-test showed no significance (see Table 6) between control and treatment conditions (EIKEN levels are disregarded here), the researcher feels the results from both lexical frequency and density warrant further investigation. It is possible the AI KAKU writing assistant is introducing additional noise to higher-level participants and somehow hindering or not positively influencing their writing performance. Alternatively, other forms of intervention may be considered to not just improve perceived load but also to affect writing performance more positively.

The researcher decided to split the clusters based on internal discussion and the descriptors of HIGH, MIDDLE and LOW have some flexibility in their definitions (i.e., EIKEN level 2.5 can arguably be considered a “high” level depending on what is being compared). To remove researcher bias in the analysis, a more detailed breakdown of performance per level without clustering can be seen in Figure 16. When broken out of the prescribed clusters, the data suggests higher-level participants are benefitting more from the AI assistant (AI KAKU) than lower-lower participants, suggesting evidence of the Matthew effect. The lexical frequency and diversity for the highest level (EIKEN 1.5) participants clearly show improvement that is not evident at the lower levels.

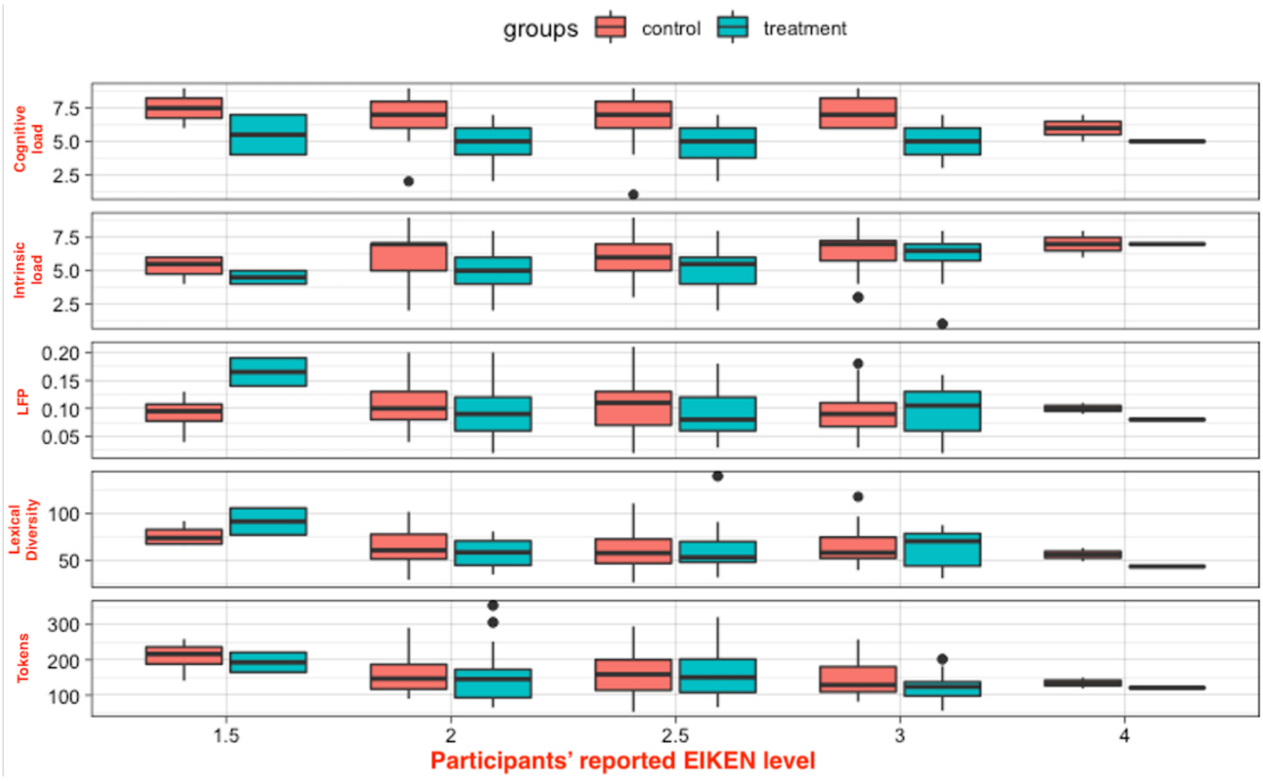


Figure 16. Performance across all EIKEN levels.

3.6 Conclusion and Future Work

The data gathered shows evidence that AI KAKU had some positive impact on the L2 writers who participated in this study. The participants produced more words and perceived less mental difficulty when answering the writing prompt with AI KAKU versus without it. While lexical diversity (LD) and lexical sophistication (LFP) did not show any improvement, the researcher believes longer exposure and training with the treatment tool would allow the participants to become more accustomed to the word suggestions and reverse translation provided by AI KAKU. Regardless, the results from this study are promising and further research into AI KAKU is warranted.

Regarding the second research question of evidence of the Matthew effect and how new technology such as AI KAKU impacts users of different skill levels, the researcher could see some effects regarding the cognitive load, lexical frequency, and lexical density. Lower-level user's intrinsic cognitive load remained high despite the assistance AI KAKU gave them during the writing process. On the other hand, higher-level users demonstrate reduced load and improved writing performance while under the treatment condition. Evidence of the CALL Matthew effect in the data supports the argument that higher-level users are benefitting more from the introduced technology than lower-level users. It is to be noted, however, that the distribution of EIKEN levels was heavily skewed to the middle/high levels of 2.5 and 2 and only 3% of the participants reported an EIKEN level of 1.5. Further investigation with a greater number of participants at each EIKEN level is needed to investigate if the effects found in this study can be replicated.

AI KAKU was developed to reduce the cognitive load during the writing process for EFL users. By reducing the problem space and guiding them to think directly in the L2 as opposed to translating their thoughts composed in their L1, learners can hopefully use their cognitive resources on higher-level writing aspects such as organization and revision. An unwanted effect of introducing technology in the learning process, such as in the case of AI KAKU use in English writing, is the widening educational achievement gap or Matthew effect. The researcher recommends instructional designers, CALL developers, and in-service educators be more aware of this potentially negative effect of CALL and develop strategies to mitigate the phenomenon.

Further research is needed into these mitigating strategies to reduce the confounding factor of the CALL Matthew effect. The results from this study are in contrast to a similar study by (Chon et al., 2021) that used machine translation (Google Translate) as a mediating agent. The researchers in that study found machine translation assisted the lower-level participants at a greater rate, bringing their performance closer to the higher-level participants. Chon et al's (2021) study does not address the Matthew effect and did not use an explicit mitigating strategy to reduce its effects. A pertinent question is then what are the factors that may exacerbate the Matthew effect among participants.

In addition, further investigation into AI KAKU’s impact on the writing process with a wider range of writing quality dimensions, including human assessment of participant writing is warranted. To the same extent that computer-assisted spelling and grammar-check have permeated writing in the modern age, AI-based digital agents will presumably be as commonplace as those older forms of digital assistance. Aspects of their potential should be studied further to ensure equitable access and benefit.

The results for the second experiment were promising. The researcher was able to recruit a larger base of participants (n = 90) by successfully conducting the experiment over the edX online platform. Results showed that participants writing under the treatment condition demonstrated improved fluency (total number of tokens produced) while also indicating lower intrinsic load. However, the researcher felt that AI KAKU’s impact on student writing could be improved by adding metacognitive support to the writing process. Metacognitive writing strategies use is a factor the researcher believed could extend the benefits of using AI KAKU. Metacognitive training in addition to pre-task planning, prompting and nudging were employed as techniques that would further support EFL writing.

Importantly, the two experiments thus far only examined participants writing samples via machine analysis. While machine analysis gives researchers objective data on writing samples, it does not give researchers any insight into more holistic writing quality aspects such as task completion (e.g., was the response relevant to the prompt?) or cohesion (e.g., were the ideas presented in a logical manner?). To address this gap in this research, the researcher expanded the number of writing quality factors to include human assessment in the final experiment detailed in the next chapter.

Chapter 4 [7] Impact on second language writing via an intelligent writing assistant and metacognitive training

Keywords: L2 writing, Digital writing, Metacognition, Educational software, experimental research, CALL

4.1 Background

The researcher’s final experiment deviates from the first two studies by employing a blind controlled research design. Due to this research design, participants’ baseline English ability was a confounding factor that was controlled for and discussed in the chapter. In addition, metacognition was considered as a treatment factor, which the first two studies did not investigate. The rationale of providing such metacognitive support and the results of that support is detailed in this chapter. Finally, as mentioned earlier, human assessment of participants’ writing was considering as a key dimension that should be examined to determine the impact the treatment tools were having on the participants. The results of this experiment were published in the proceedings of an international peer-reviewed conference.

[7] Parts of this chapter have been published under Gayed, J. M., Carlon, M. K. J., & Cross, J. S. (2022). "Impact on Second Language Writing via an Intelligent Writing Assistant and Metacognitive Training," 2022 IEEE Frontiers in Education Conference (FIE), 2022, pp. 1-9, https://doi.org/10.1109/FIE56618.2022.9962406

4.2 Introduction

The importance of academic English writing ability of Science, Technology, Engineering, and Mathematics (STEM) students has increased with many tertiary level institutions emphasizing the ability to contribute to knowledge transfer and exchange on a global level (Maringe & Foskett, 2012). Notably, Zhu (2004) notes that business and engineering-related programs are in high demand for international students looking to advance their academic careers. The task of writing itself is seen as an essential element in engineering education (Wheeler & McDonald, 2000) with "writing to learn" pedagogy supplementing engineering departments' goal of enabling graduates' communicative ability.

Post-graduate students are often tasked with writing thesis or research proposals. The same is true for non-native level English language learners using English in EFL (English as a foreign language) or ESL (English as a second language) environments. Researchers have identified some difficulties second language (L2) students face when tasked with writing, such as the ability to communicate research results (Dong, 1998), idea organization, and appropriate vocabulary use (Bitchener & Basturkmen, 2006). Examples such as Xiao and Chen's (2015) study of Chinese engineering students in an EFL context also identifies similar factors such as planning and organization strategies and language formulation as significant barriers that face engineering L2 writers.

Engineering students themselves recognize the need for domain-specific English skills to support their future professions (Koenig et al., 2020). This aligns well with actual industry practice: the International Organization for Standardization has 27 published standards relating to technical product documentation as of March 2022, not to mention five more under development and 23 standards already withdrawn, attesting to the inherent difficulty of documenting technical works for those to be usable by the public (*ISO - 01.110 - Technical Product Documentation*, n.d.) Through these standards, the need for complete, clear, concise, and consistent writing is emphasized.

Modern tools such as automated writing evaluation, grammar/style checkers, and next-word prediction algorithms can be used to overcome low-level thinking difficulties such as sentence formulation and lexical deficiencies. This can lead to higher-level thinking tasks such as idea development, organization, and revision. However, over-reliance on tools might impede independent writing skills development, returning to the original unwanted situation. In addition, metacognitive thinking can lead writers to make more efficient use of avail- able tools and eventually have higher writing output quality. Both digital writing aid use and metacognition enhanced writing skills have been investigated in other studies, but systematic research investigating both in action is limited.

4.3 Theoretical Framework

Writing ability is commonly seen as an indicator of language acquisition progression and proficiency (Aydoğan & Akbarov, 2014; Krashen, 2003; Llach, 2011). Gaining writing proficiency has numerous benefits. As Basturkmen and Lewis (2002) indicate, EFL learners who develop their writing skills also develop their ability of self-expression, and their confidence and enjoyment of written communication also expand. In addition, writing is not only a communication skill that is taught and assessed in academic settings but is also seen as an essential skill for professional success (Tardy & Matsuda, 2009).

EFL learners often struggle in the writing process (Nunan & Carter, 2001) and lack some of the formulating strategies (Ceylan, 2019) needed to become better writers. EFL learners may find comfort in writing in their L1 language then translate their writing to English and improve their English writing skills by comparing their translations with machine translations (Tsai, 2019). The machine translation use, however, can end up being crutches instead of scaffolds for learning, making it harder for the learners to become independent writers. On the other hand, those who choose to write in English directly may have vocabulary deficiency, making them susceptible to tip-of-the-tongue phenomenon. Studies have shown that prolonged struggle during this phenomenon makes a person focus more on the struggle than retrieving the needed word (Baker et al., 2010); thus, it is more likely for the person to struggle again when they need to use the same word in the future (Abrams & Davis, 2016). This is an unproductive cognitive load for EFL learners writing in English. Aside from these inherent difficulties, translation and vocabulary recall tap into the Remember and Apply levels of the revised Bloom's Taxonomy; these are lower levels whereas English writing itself requires the higher levels: Analyze, Evaluate, and Create. For EFL learners to improve their English writing, they must be supported in going beyond the lower-level thinking skills by removing associated barriers.

Metacognition, more colloquially known as thinking about thinking, has been shown to improve learning outcomes regardless of age or intelligence (Ohtani & Hisasaka, 2018). In the case of writing in the L2, metacognition could include understanding the source of writing difficulty and seeking help to address deficiencies, may it be through teacher support or the use of tools such as dictionaries. Consequently, highly metacognitive learners can reasonably be expected to persevere and not easily give up on cognitive challenges (Gama, 2004) and thus not succumb to over-reliance on assistance.

The digital writing assistant (AI KAKU) developed for this study (Gayed, Carlon, Oriola, et al., 2022) was created with a framework to support EFL learners in the writing process. Current word processing platforms (Microsoft Word, Google Docs) have features that primarily help the first language (L1) user but do little to assist L2 users struggling with language production. The researcher, therefore, attempts to measure the combined effects of using an intelligent writing agent and metacognitive training and prompting. This intersection is a unique approach to digital writing that the researcher predicts will become more prevalent in the future. As software tools become more sophisticated and assistive, learners will be required to focus on higher-level critical thinking skills to perform at the required levels.

This study aims to answer the following research questions:

1. To what extent do metacognitive training/prompting and the use of AI KAKU impact the writing proficiency of L2 participants?
2. Do participants improve their metacognitive awareness / pre-task planning after receiving metacognitive strategy training?

4.4 Literature Review

4.4.1 Intelligent writing assistants

Intelligent writing agents within the field of Computer Assisted Language Learning (CALL) are not new (see Bowerman's (1992) work in the early 1990s) but have seen increased research interest due to the sophistication of newer Artificial Intelligence/Machine Learning (AI/ML)-based technologies that have come to market. This development and research have led to Natural Language Processing (NLP) applications tailored to L2 learners and users. Gamper & Knapp's (2002) scoping review of intelligent CALL applications identified 19 systems that aim to support L2 writers. The systems identified in the researchers' review can be categorized as grammatical and semantic support (5 applications), collocation and sentence level support (7 applications), higher-level communication skills and user awareness (5 applications), and lastly, composition and schema support (2 applications).

Some applications reported in the literature include Dai et al., (2014) work on an NLP-based writing assistant for Chinese input that provides word and sentence-level suggestions to users. They cite the struggle writers may experience when thinking of the most appropriate word or phrase during the writing process as reasoning to develop their application. Chen et al., (2012) use NLP techniques to develop an application called "FLOW," which is intended to assist English as a second language (ESL) writers in composing and revising in English. Their initial testing with Chinese students indicates that word and phrase suggestions are beneficial to the user and recommend further development and testing of similar frameworks. The Automated Writing Evaluation (AWE) application Grammarly has received some attention (Dizon & Gayed, 2021) on its impact in the L2 classroom. The application was not developed for L2 users per se but contains several features that support L2 writing such as automatic writing feedback, text prediction, and post-writing evaluation, all supportive technologies for the L2 user.

4.4.2 Metacognition in engineering and language acquisition

Metacognition is concretely seen as the knowledge and regulation of ones' cognitive abilities (Flavell, 1979). In many ways, metacognition is considered to be domain-independent (Azevedo, 2020); that is, skills gained from metacognitive instruction done for a particular subject matter may transfer to a different learning domain. For instance, developing computational thinking skills alongside metacognition is anticipated to have a positive feedback loop as both reinforce problem solving skills (Yadav et al., 2022). Not only are both computational thinking and problem-solving skills important in engineering education, but metacognition itself is critical as most, if not all, engineering professions require lifelong learning (Marra et al., 2017). Metacognition is a powerful tool for lifelong learning as it enables an individual to use their life experiences to inform their learning through reflection.

Metacognitive techniques have also been well studied in second language acquisition studies. Dabarera, Renandya and Zhang (2014) use the Metacognitive Awareness of Reading Strategies Inventory (MARSI) in an experimental study with English as a second language (ESL) students in Singapore. The researchers found a significant (albeit small) gain in reading comprehension in the participants who were given metacognitive strategy instruction. Knospe, (2018) builds on the concept of "cognitive regulation" from Dimmitt and McCormick's (2012) work and investigates the potential metacognitive strategies have on foreign language writing. The researcher takes a case study approach with a single participant in a secondary school setting. Keystroke logging/screen capture software and stimulated recall interviews were used to gain insight into the participant's metacognitive knowledge while writing in a foreign language. The researcher highlights the importance of metalinguistic awareness, metacognitive knowledge of self, and metacognitive knowledge of the task as factors that the participant in the study engaged in to complete a writing task. Notably, the researcher states that these metacognitive strategies were transferable to contexts outside of L2 writing, such as L1 writing.

Metacognition in a computer-assisted learning environment is explored by Zhang & Qin (2018), who developed the Language Learners' Metacognitive Writing Strategies in Multimedia Environments (LLM-WSIME) questionnaire. Data from 400 Chinese EFL participants showed that metacognitive evaluating strategies as significant features in the participants. Importantly, Scardamalia & Bereiter (1987) who proposed the Knowledge Telling model of writing (text production is largely guided by fixed schemas), indicate that "good" writing can be produced when the writer has content organized in their working memory before they start writing. This further supports the researcher's aim of providing metacognitive support during the writing process.

4.5 Methodology

4.5.1 Research design

Figure 17 shows the experimental design used for this research. This study employed a pre-test/post-test for the metacognitive writing strategies measures and a repeated measures experimental research design with control and treatment groups for all the writing quality measurements. The experiment was conducted over a month from November to December 2021. After going through the research description and consent form approved by the institute's ethical research review board, the participants were randomly assigned to the control or the treatment group. The control condition consisted of a pre-test writing task, pre-test metacognitive writing strategies questionnaire, short training videos, two unassisted writing tasks, and a post-test metacognitive writing strategies questionnaire and survey. The experimental condition consisted of a pre-test writing task, pre-test metacognitive writing strategies questionnaire, short training videos, metacognitive prompts and nudges, two assisted writing tasks, and a post-test metacognitive writing strategies questionnaire and survey. All writing prompts in the experiment were chosen from sample independent writing tasks of the Test of English as Foreign Language Internet-Based Test (TOEFL® iBT), a commonly used test of English proficiency for English as a Foreign Language (EFL) students. Both conditions ended with a thank you video to close the experiment.

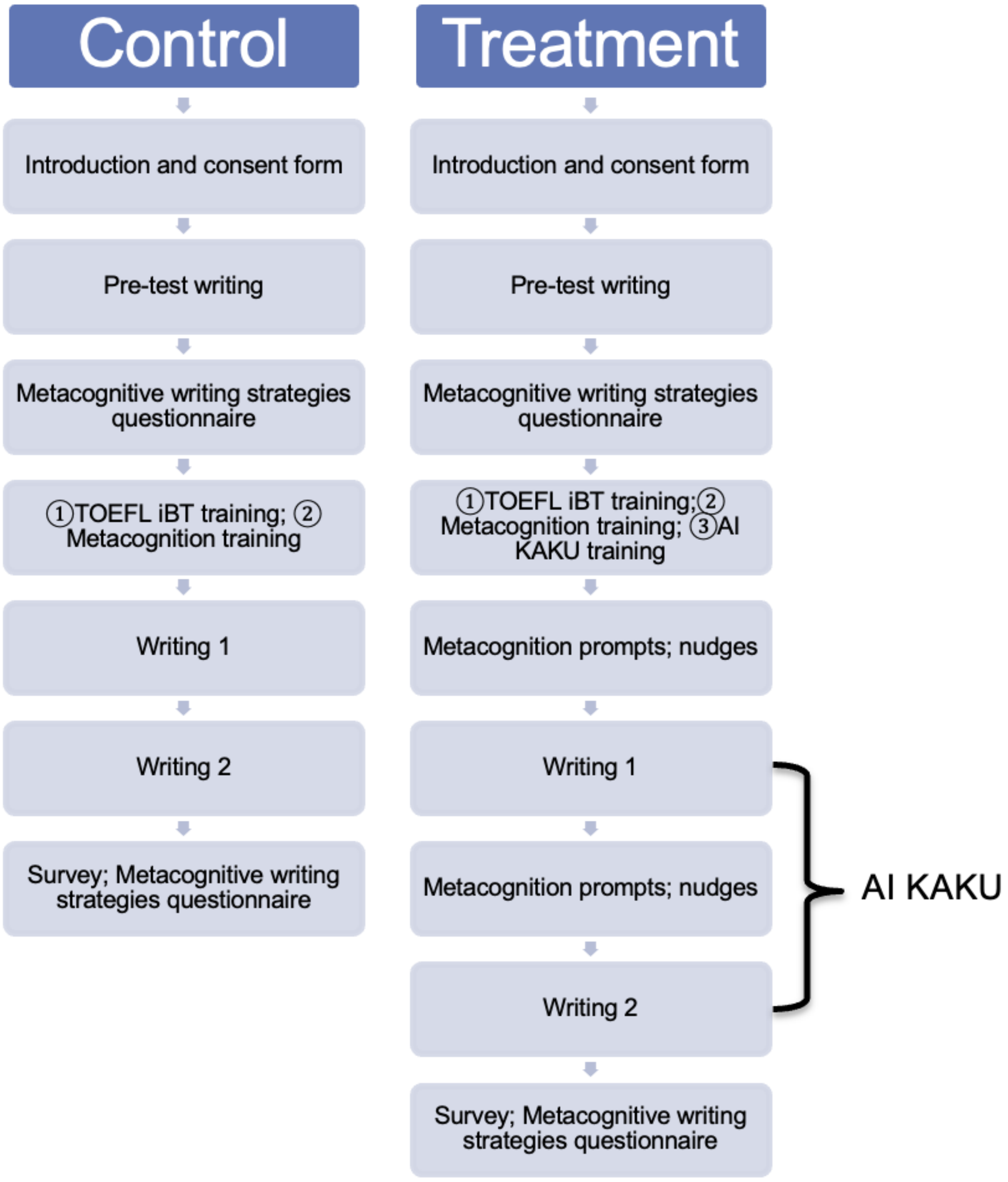

Control
Introduction and consent form
Pre-test writing
Metacognitive writing strategies questionnaire
①TOEFL iBT training; ② Metacognition training
Writing 1
Writing 2
Survey; Metacognitive writing strategies questionnaire
Treatment
Introduction and consent form
Pre-test writing
Metacognitive writing strategies questionnaire
①TOEFL iBT training;② Metacognition training; ③AI KAKU training
Metacognition prompts; nudges
Writing 1
Metacognition prompts; nudges
Writing 2
Survey; Metacognitive writing strategies questionnaire
AI KAKU

Figure 17. Experimental design flow. AI KAKU is the intelligent writing assistant.

While participants under the control condition received the TOEFL iBT and metacognition training, they had to complete the three writing tasks (pre-writing, writing 1, writing 2) without thesaurus, grammatical error feedback, or predictive text functions. Participants under the treatment condition completed the three writing tasks with the AI KAKU writing assistant. This writing assistant features text prediction and reverse translation features intended to help L2 writers in the writing process. Participants were encouraged to write at least 300 words, but the experiment’s software accepted submissions of any length. The writing samples analyzed in this study contained between 51 to 635 words.

4.5.2 Participants

Convenience sampling was used, and Japanese university students were invited to participate in the study from three institutions related to the researcher. The participants’ consent was collected via the experiment’s website, and they were able to retract their consent at any time during the experiment (please see Appendix A (Participant consent form)). The experiment was structured to allow the participants to start and finish the training and writings tasks at the participants’ convenience within the experiment’s duration. The experiment initially received 197 registrations; after checking for incomplete and errant attempts, 121 submissions (control n = 60, treatment n = 61) were deemed acceptable, resulting in 363 writing samples in addition to survey responses available for analysis.

The participants’ incoming English level was estimated by analyzing the writing samples they submitted in the pre-test. Much literature has been published regarding the correlation of writing features and the Common European Framework of Reference for Languages (CEFR) levels (Harsch & Kanistra, 2020; Leontjev et al., 2016; Salamoura & Saville, 2010) therefore the researcher was confident that estimating CEFR based on a participant’s writing was a reliable method. In particular, this was accomplished by using the machine learning based CEFR level checker implemented by Cathoven A.I. (2022). Using machine learning techniques to estimate English proficiency level is gaining traction as positive correlations between the techniques and established classification methods are established (Schmalz & Brutti, 2021). Table 8 shows the resulting CEFR levels of the participants and a scatterplot showing individual data points is shown in Figure 18. The results show the participants in this study are tightly grouped around CEFR B1-A2 or EIKEN levels Pre-2 and 2. This is in line with supplemental data the researcher obtained from one of the participating universities. The average Test of English for International Communication (TOEIC) score of the representative students from that university was 738 which is just under the CEFR level of B1 (TOEIC 790).

Table 8. Participants estimated CEFR and equivalent EIKEN levels.

CEFR	EIKEN	Treatment	Control	%
A1	3	0	1	0.8%
A2	Pre-2	18	17	28.9%
B1	2	42	42	69.4%
B2	Pre-1	1	0	0.8%

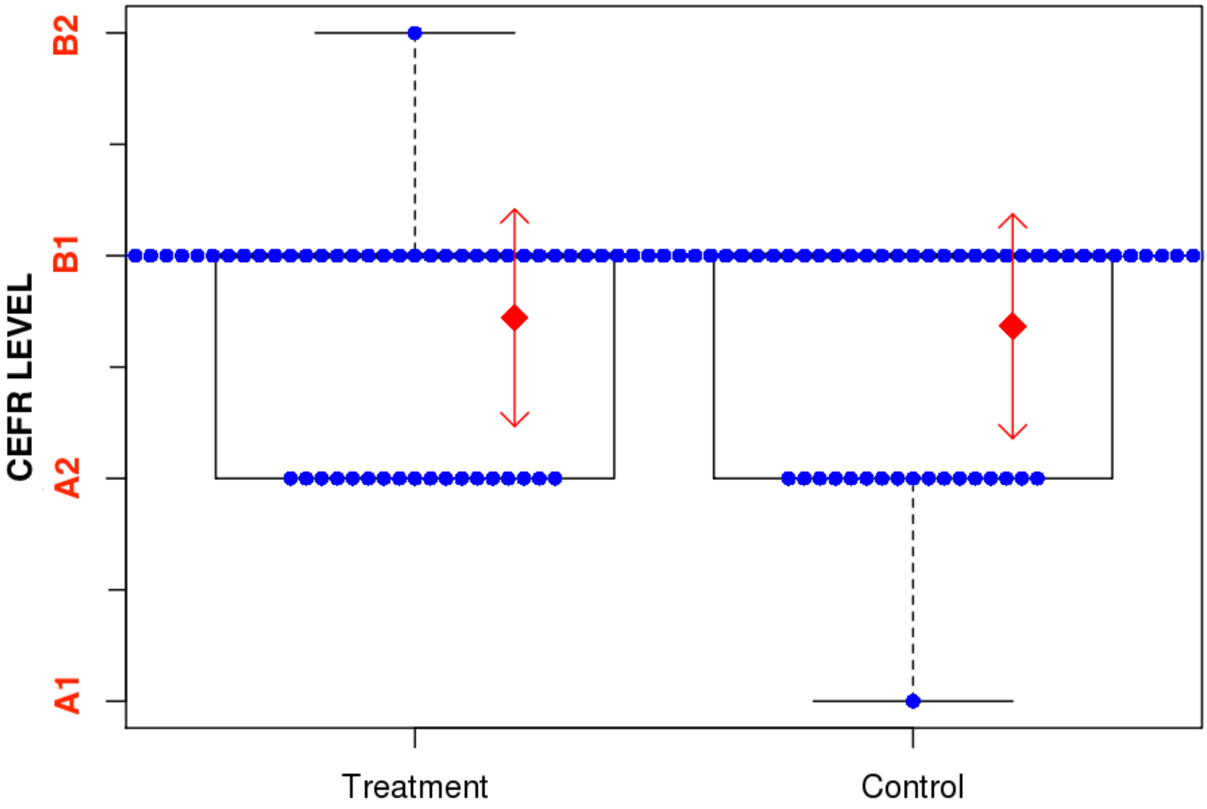


Figure 18. Scatterplot of participants' CEFR levels.

Note: Means and +/- SDs are displayed in red.

The researcher's second experiment described in Chapter 3 asked the participants to self-report their EIKEN levels to get an understanding of their baseline English ability. However, asking the participant to self-report is somewhat problematic as considerable time might have elapsed since they took the exam, and their skill development could have increased or decreased from that point forward. So, in this experiment, a different approach is taken, yet a key-point to consider is that these CEFR estimations are only based on the participant's writing ability. The CEFR level checker by Cathoven A.I. calculates the level by correlating factors of vocabulary level, verb forms used, and sentences structures with a CEFR score. It does not assess other language features such as listening, speaking and reading abilities. In order to address this, further measures were taken in this study to identify any outliers in the participants as described in section [4.6.2] of this chapter.

Participant training

The metacognitive training consists of a ten-minute video recorded by the researcher and largely inspired by Boston University's Teaching Writing Metacognition flipped classroom module (Boston University Teaching Writing, 2020). The training included: 1) a short introduction to metacognition, 2) where the learners might encounter metacognition in the future, and 3) a few tips for writing metacognitively, such as breaking down tasks and outlining. The inclusion of metacognition training is in response to previous research results indicating that knowledge of cognition cannot be developed by metacognitive prompts alone (Carlon et al., 2021) but can be improved with training interventions (Sato & Dussuel Lam, 2021).

Learnability is essential for software to be usable (Nielsen & Molich, 1990). A two-minute video walking through the metacognitive prompts and the main digital writing assistant interface was also prepared to ensure that the participants understood how the digital writing assistant works. This is akin to onboarding tutorials for software applications (Strahm et al., 2018). This training video is only shown to participants in the treatment group since the metacognitive prompts and the digital writing assistant are visible to them only.

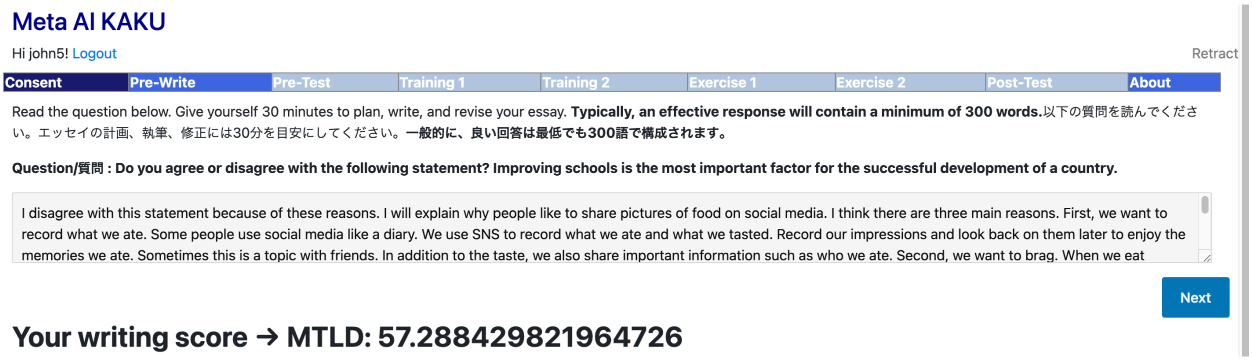
4.5.3 Treatment software

Next word prediction and reverse translation

This study employs a digital writing assistant called "AI KAKU" that has been previously used by the researcher in two empirical studies (Gayed, Carlon, & Cross, 2022; Gayed, et al., 2022) that showed potential in improving the writing performance of the EFL participants. This study expands upon those initial studies to add writing training, metacognitive training, prompting, and the intelligent agent's writing assistance. The web-accessible tool (https://www.aikaku.app/) was designed with L2 writers as its primary target demographic. AI KAKU has several unique features; a text prediction engine that displays word suggestions with confidence scores based on the user's input. These word suggestions are based on a language model developed by OpenAI (GPT-2) and implemented by the Allen Institute for AI (Radford et al., 2019).

Secondly, a reverse-translate output field that translates the users' inputted English into their chosen first language. This is intended to encourage the L2 writer to continue writing in the L2 without resorting to commonly used tactics such as wholesale machine translation from the user's L1. In addition, the simultaneous reverse translation is intended to provide a mental bridge back to the user's L1, allowing them to check briefly if what they are writing in the L2 is what they intended to write. The word suggestions and updated reverse translation only appear after a 2.5-second pause in typing. This creates space for user agency and does not interrupt the writing process when participants are pausing/thinking.

In addition to the intelligent word suggestions and reverse translation features, AI KAKU also provides writing feedback for the user in the form of a Measure of Textual Lexical Diversity (MTLD) score (see Figure 19). This measure of lexical diversity has been shown in a study by Treffers-Daller, Parslow and Williams (2018) to equate to the CEFR B1 (IELTS level 4, TOEFL iBT 42-71) level when a score of 70.14 is achieved. Importantly, AI KAKU does not feature spelling/grammar correction as feedback given to the user while writing. Research has shown that corrective feedback on mechanical aspects of writing does little to improve student writing. At the same time, more emphasis should be placed on strategy, and formative feedback to students (McCarthy et al., 2022).


Meta AI KAKU

Hi john5! Logout

Read the question below. Give yourself 30 minutes to plan, write, and revise your essay. **Typically, an effective response will contain a minimum of 300 words.**以下の質問を読んでください。エッセイの計画、執筆、修正には30分を目安にしてください。**一般的に、良い回答は最低でも300語で構成されます。**

Question/質問 : Do you agree or disagree with the following statement? Improving schools is the most important factor for the successful development of a country.

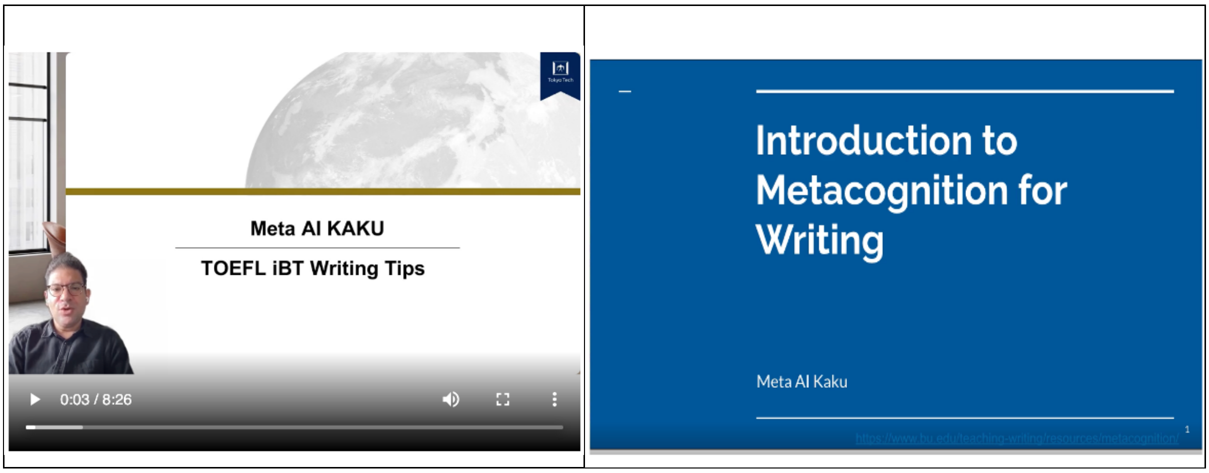
I disagree with this statement because of these reasons. I will explain why people like to share pictures of food on social media. I think there are three main reasons. First, we want to record what we ate. Some people use social media like a diary. We use SNS to record what we ate and what we tasted. Record our impressions and look back on them later to enjoy the memories we ate. Sometimes this is a topic with friends. In addition to the taste, we also share important information such as who we ate. Second, we want to brag. When we eat

Your writing score → MTLD: 57.288429821964726

Figure 19. Feedback on lexical diversity.

Training videos

All the participants, in both control and treatment groups, received two training videos before they began the post-writing tasks. The videos ranged from 8 to 10 minutes in length and covered the topics of (1) writing strategies for the TOEFL iBT independent writing task and (2) introduced metacognition and how to use metacognition to become a better writer (see Figure 20). Allowing both the control and treatment groups to benefit from the training videos was one of the goals of the researcher to make the learning outcomes as equitable as possible for all the participants.



Introduction to Metacognition for Writing

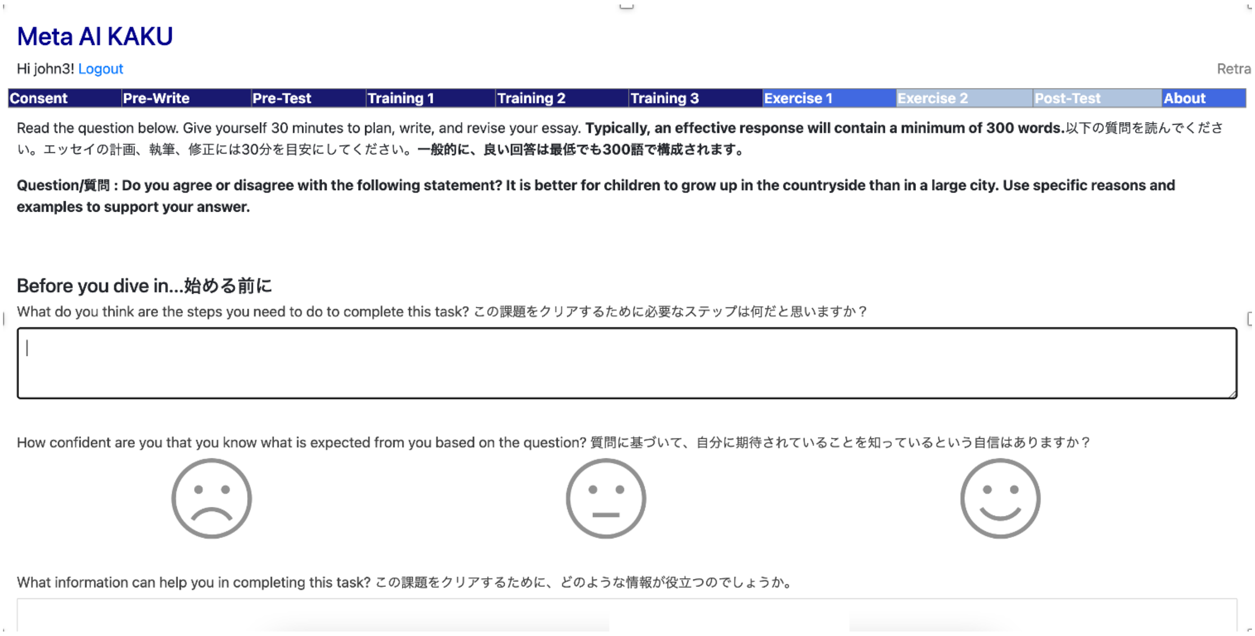
Meta AI Kaku

https://www.bu.edu/teaching-writing/resources/metacognition/ 1

Figure 20. Screenshot of training videos.

Metacognitive prompts and nudges

One metacognitive strategy discussed in the training video is the ability to reflect on the task at hand before starting it. To facilitate this pre-task planning, the treatment group was asked to think about the necessary steps and information needed to complete the writing task. In addition to answering the pre-task planning prompts, the participants self-rate their confidence level by clicking on a sad, neutral, or happy face as seen in Figure 21.



Read the question below. Give yourself 30 minutes to plan, write, and revise your essay. **Typically, an effective response will contain a minimum of 300 words.**以下の質問を読んでください。エッセイの計画、執筆、修正には30分を目安にしてください。**一般的に、良い回答は最低でも300語で構成されます。**

Question/質問 : Do you agree or disagree with the following statement? It is better for children to grow up in the countryside than in a large city. Use specific reasons and examples to support your answer.

Before you dive in...始める前に

What do you think are the steps you need to do to complete this task? この課題をクリアするために必要なステップは何だと思いますか？

How confident are you that you know what is expected from you based on the question? 質問に基づいて、自分に期待されていることを知っているという自信はありますか？

☹ 😐 ☺

What information can help you in completing this task? この課題をクリアするために、どのような情報が役立つのでしょうか。

Figure 21. Example of pre-writing metacognition prompting.

After the participants completed the pre-task planning and reflection prompts, they were presented with one of the following nudges based on the rules displayed in Table 9.

Table 9. Conditions of nudges displayed to user.

Self-evaluation		**Input length**	**Nudge**
Execution	**Confidence**		
Sad or Neutral	Any	Less than 50 characters	You can try breaking down the task into smaller chunks to make them less overwhelming. タスクを小さな塊に分解して、圧倒されないようにすることもできます。
Any	Sad or Neutral	Less than 50 characters	You will feel more confident if you have more information at hand. Feel free to use your life experiences as information source. 手元に多くの情報があれば、自信を持つことができます。自分の人生経験を情報源にするのもいいでしょう。
Sad	Sad	Less than 50 characters	Do not be overwhelmed by the task. Take your time to reflect on your action plan. タスクに圧倒されてはいけません。アクションプランをじっくりと考えてみてください。
Sad or Neutral	Any	More than 50 characters	You may not feel like it, but you are actually to a good start! Feel free to elaborate more on your thoughts. 自分では感じていないかもしれませんが、実は良いスタートを切っているのです。あなたの考えをもっと詳しく聞かせてください。
Any	Sad or Neutral		
Happy	Any	Less than 50 characters	It will be helpful if you can demonstrate more how you break down the task at hand. タスクをどのように分解しているのか、より具体的に示していただけると助かります。
Any	Happy	Less than 50 characters	Listing out the information you need may help make completing the task easier. 必要な情報をリストアップすることで、よりスムーズに作業を進めることができます。
Happy	Happy	Less than 50 characters	Approaching writing tasks in a reflective manner may lead to better future performance. ライティングの課題に反省的に取り組むことは、将来のパフォーマンス向上につながるかもしれません。
Happy	Happy	More than 50 characters	You are off to a good start! Keep it up and remember to reflect on your results. あなたは良いスタートを切ることができました。その調子で、自分の結果を振り返ることも忘れないでください。

4.5.4 Writing quality factors

Tools used

The researcher analyzed the writing samples to gain insight into the linguistic development of the L2 participants. The samples obtained were analyzed via seven measures: six quantitative methods via two web-based tools and one qualitative method via a holistic assessment scale developed by the Educational Testing Service (ETS) corporation (ETS, 2019). The ETS rubric was used to match the writing prompts employed in the study, based on ETS's TOEFL® iBT exam. To measure the dimensions of token count and MTLD (McCarthy, 2005), Mizumoto's (2015) web-based R interface (see also Mizumoto & Plonsky (2016)) accessible via https://langtest.jp was used. The remaining machine assessment dimensions of lexical density (LD), lexical frequency profile (LFP), mean length of T-unit, clause/T-unit were analyzed with the web-based Lexical Complexity Analyzer developed by Lu (2012) and accessible via https://aihaiyang.com/software/lca/ and the web-based L2 Syntactic Complexity Analyzer developed by Ai & Lu (2013) and accessible via https://aihaiyang.com/software/l2sca/. A summary of the writing quality dimensions used in this study and their related descriptions are outlined in Table 10.

Table 10. Dimensions of writing sample quality.

Assessment dimension	**Description**	**Demonstratable writing skill**
Token count	Total number of corrected tokens in the sample.	Fluency
MTLD	The mean length of a sequence of tokens in a text that maintain a given TTR value.	"Richness" and "variety" of writing
Lexical Density (LD)	Ratio of lexical (content) words to the total number of words in the sample.	Informativeness
Lexical Frequency Profile (LFP)	Ratio of number of words written beyond the 2000 word-frequency level.	Sophistication
Mean Length T-unit	Mean number of words in one minimally terminable unit.	Lexical development / Syntactic complexity
Clause / T-unit	Number of clauses (amount of subordination) in one minimally terminable unit.	Syntactic complexity
ETS Holistic rubric	Human assessment using the ETS Independent Writing Holistic 6-point Likert scale.	Task completion / Organization / Coherence

Writing quality dimensions

To summarize the lexical and syntactic dimensions used in this study, a brief description of the seven measures noted above will be clarified in more detail. The first measure of token count is the total number of tokens written for each task given to the participant. Simply put, the volume of written output is commonly seen as an indicator of L2 writing maturity and proficiency (Crossley & McNamara, 2012).

The second quality dimension of MTLD measures the type to token ratio (TTR) after every word until the value of 0.72 is reached, after which a factor is calculated. The TTR measurement starts again with the next token until the next factor is calculated. Finally, the total number of tokens is divided by the total number of factors. Essentially, lexical diversity or lexical variation demonstrates an L2 writer's range of vocabulary. Beginner or elementary L2 writers tend to repeatedly use the same limited set of vocabulary; more advanced writers can use a greater variety of words. There are several methods to measure this dimension of writing proficiency, TTR (corrected, root), the number of different words (NDW), and McCarthy and Jarvis's (2010) D measure are examples. However, those measures are more sensitive to text length; MTLD, on the other hand, is more robust and less sensitive to text length (Koizumi, 2012).

The third measure of lexical density (LD) can show how much Information is in the text, its content density, with more dense texts being able to relay more information than less dense texts, which Breeze (2008) identifies as one measure of language proficiency.

The fourth measure of Lexical Frequency Profile (LFP) shows the proportion of words that are in the 2000 word- frequency level (based on the British National Corpus [BNC] and the Corpus of Contemporary American English [COCA]) and beyond the total number of words written (Laufer & Nation, 1995) Measures of lexical sophistication are correlated to L2 development (Polio, 2001) as Crossley and McNamara (2011) show a relationship between the number of uncommon or advanced words in L2 learners' writing and the learner's language development.

The measures of mean length of t-unit and clause per t- unit are linguistic features that demonstrate syntactic complexity. Again, for this measure, the literature shows that more advanced L2 writers can produce more complex syntactic elements such as subordination or coordination and produce longer sentences. In contrast, lower-level L2 writers tend to compose shorter, less complex sentences (Casal & Lee, 2019).

The last writing quality dimension employed four university EFL educators to serve as raters using ETS's publicly available 6-point holistic assessment rubric (see Table 26 in Appendix D for reference). Human assessment of the writing was considered an important dimension to include in this study as machine assessment cannot fully capture factors such as task completion (relevance to the question), organization, or proper use of language (Wiseman, 2012). In other words, a well written composition is not simply the sum of its parts. The training was provided via videoconferencing, where the researcher provided instructions and the raters scored ten writing samples together with the researcher as a calibration session in addition to discussing how to prioritize and interpret the ETS rubric. The rubric itself follows established L2 essay rubric norms by prioritizing content and ideas, organization, cohesion, vocabulary, grammar, and mechanics as factors in that order (Schoonen, 2005). Table 11 shows the number of samples rated per rater, with all four raters scoring 73 common samples and then each rater scoring an additional 72 to 74 samples independently. Interrater agreement and rating normalization, scaling techniques will be discussed in the Results and Discussion section of the article.

Table 11. Sample distribution across raters.

	R1	**R2**	**R3**	**R4**
# Commonly rated	1 – 73	1 – 73	1 – 73	1 – 73
# Individually rated	74 – 145	146 – 217	218 – 289	290 – 363

4.5.5 Metacognitive measures

The Metacognitive Writing Strategies Questionnaire (MWSQ) (Zhao & Liao, 2021), displayed in English and Japanese, was modified from 18 questions to just ten questions after the review or course staff to reduce the likelihood of participant dropout (see Appendix E for reference). The options were also reduced from six to just five (1 being the worst and 5 the best), which was shown to yield better data quality (Revilla et al., 2014). It was then administered before and after the main writing activities on both the control and treatment groups to gauge the effects of the short metacognition training video and the metacognitive prompts on the participants. The questionnaire measures metacognitive ability specifically associated with writing by asking about ones' understanding of the writing instructions and target audience, the ideas one intends to convey, and their approach to organizing their writing.

4.6 Results and discussion

4.6.1 Inter-rater agreement and scoring normalization

Human assessment of writing samples can take considerable human resources to complete; this study collected over 300 writing samples with an average length of 255 words each. The researcher recruited three raters (including one of the authors of this study, four raters in total) to assess the writing samples using the ETS 6-point holistic rubric. To complete the scoring more efficiently, the researcher split the 363 samples into five chunks: chunk 1 was graded by all the raters and then chunks 2 to 4 were graded independently. Table 12 shows a Spearman's rho correlation matrix for the writing samples that were rated by all the raters with the range of ρ .69 to ρ .82 considered as "moderate" to "strong" in social science research (Akoglu, 2018).

Table 12. Spearman's inter-rater correlation values for commonly rated chunk.

	Rater 3	*Rater 4*	*Rater 1*	*Rater 2*
Rater 3	1			
Rater 4	0.75	1		
Rater 1	0.69	0.80	1	
Rater 2	0.74	0.78	0.82	1

However, a good correlation between raters does not indicate positive agreement. Krippendorff's alpha was calculated at α .72 indicating sufficient agreement between the raters (Krippendorff, 2018). This gave the researcher confidence to move forward and include the human assessment of the writing samples as another dimension of writing quality to be further analyzed. Given the correlation and agreement between the four independent raters was strong, the researcher proceeded to standardize the independently scored samples (see Table 11) by calculating a z-score using *Equation 1*.

$$x_{standardized} = \frac{R1score - \bar{X}(R1scorecommon)}{\sigma(R1scorecommon)} \qquad \textit{Equation 1}$$

Here, R1 is rater 1, R1score is the independently scored sample, R1scorecommon are the samples that were rated by all of the raters. This gave the researcher a standardized score (z-score) for all the independently rated samples. The z-score was scaled back to the original rubric scale using *Equation 2***Error! Reference source not found.**.

$$x_{normalized} = \frac{(x - xmin) \times 5}{xmax - xmin} \qquad \textit{Equation 2}$$

Here, x is the z-score, $xmin$ is the minimum z-score for that rater's independently scored samples, and $xmax$ is the maximum z-score for that rater's independently scored samples.

4.6.2 Baseline differences in control and treatment participants

As part of the research design, participants submitted writings for a pre-test before moving on to any of the experiment's training or assisted writing tasks. Descriptive statistics (mean and standard deviation values) are provided to illustrate the effects each condition had on the participants' writing. To check if there were any outliers (high ability or low ability) participants that may skew the data observed in the study, an independent t-test was conducted to determine if there was a significant difference between the control and treatment participants' baseline ability.

Results seen in Table 13 show that in the pre-test phase, there was no statistical difference along all writing quality measures used, giving the researcher confidence that the control and treatment groups started with similar writing abilities before any treatments were applied.

Table 13. Control and treatment pre-test scores.

	Control	*Treatment*	*t-test*
Tokens	244.4 (80)	256.8 (75)	$t(119) = -0.86$ $p = .38$
MTLD	65.4 (18.4)	63.5 (15.1)	$t(119) = 0.62$ $p = .53$
LFP	0.10 (0.03)	0.10 (0.04)	$t(119) = -0.16$ $p = .87$
ETS	3.5 (1)	3.5 (0.94)	$t(119) = 0.32$ $p = .74$
Ldensity	0.52 (0.04)	0.53 (0.04)	$t(119) = -1.75$ $p = .08$
MLTunit	14.1 (3.1)	14.5 (3.4)	$t(119) = -0.69$ $p = .49$
Clause/Tunit	1.6 (0.2)	1.7 (0.03)	$t(119) = -1.13$ $p = .25$

Mean values are displayed for Control and Treatment. Associated SD values are in parentheses.

4.6.3 Effects of treatment and control

A two-way ANOVA was conducted on the seven writing quality factors identified in this study to gain insight into participant performance in the control and treatment conditions. F-ratio and p-values are reported in Table 14. Several dimensions indicate statistical significance between control and treatment conditions and pre-test and writing tasks 1 and 2. According to Cohen's effect size interpretations, the ETS dimension exhibited the largest difference with a large effect between control, treatment (factor A), a small effect between pre-test and writing 1 and 2 (factor B), and a large effect between the interaction of factor A and B (Cohen, 2013).

Table 14. Two-way ANOVA analysis of writing quality factors.

	A	*p.eta²*	*B*	*p.eta²*	*A – B Interaction*	*p.eta²*
Tokens	1.6 (0.20)	-	1.7 (0.17)	-	0.3 (0.7)	-
MTLD	0.01 (0.89)	-	10.8 (0.0000)***	0.08	1.3 (0.25)	-
LFP	1.3 (0.24)	-	35 (0.0000)***	0.22	0.57 (0.56)	-
ETS	16.7(0.0001)***	0.12	4.7 (0.0092)**	0.03	20.6 (0.0000)***	0.14
Ldensity	1.0 (0.30)	-	18.8 (0.0000)***	0.13	3.4 (0.03)*	0.02
MLTunit	0.27 (0.59)	-	4.7 (0.01)*	0.01	0.19 (0.8)	-
Clause/ Tunit	0.73 (0.39)	-	4 (0.02)*	0.03	1.6 (0.2)	-

Note: *p <.05, **p <.01, ***p <.001; F-ratios, p-values displayed in ().
A = (Control – Treatment), B = (Pretest – Writing 1 and 2)

All the factors analyzed in Table 14 are visually represented in Figure 22a-g. For all the dimensions that exhibited statistical significance, a Holm Bonferroni post-hoc analysis (see Table 15) was conducted to determine the direction and strength of the relationships.

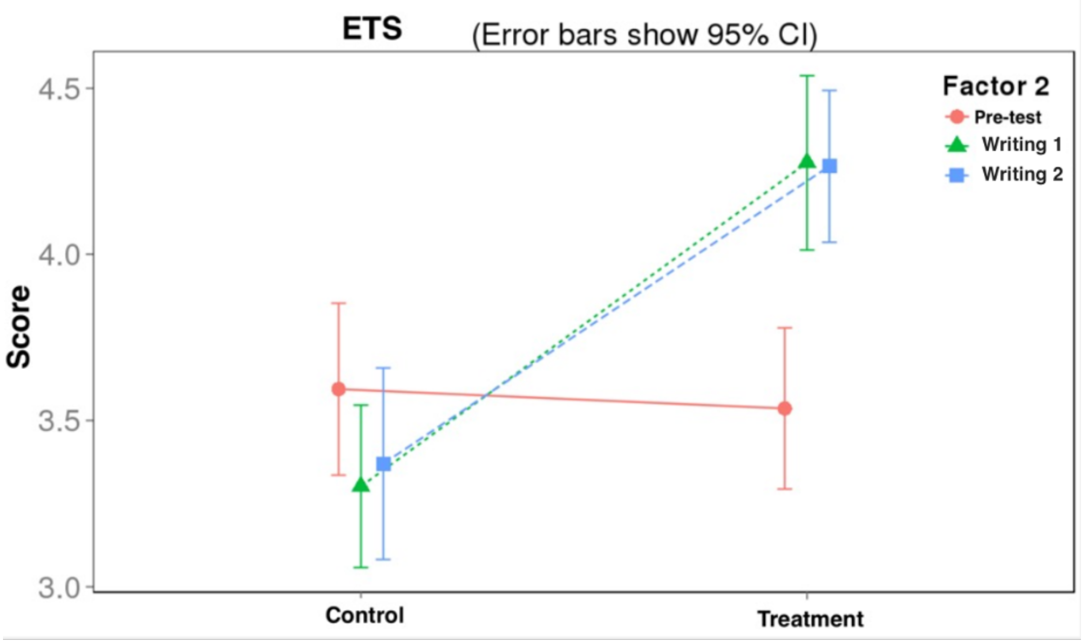


Figure 22a. Impact of Control and Treatment on measures of writing quality.

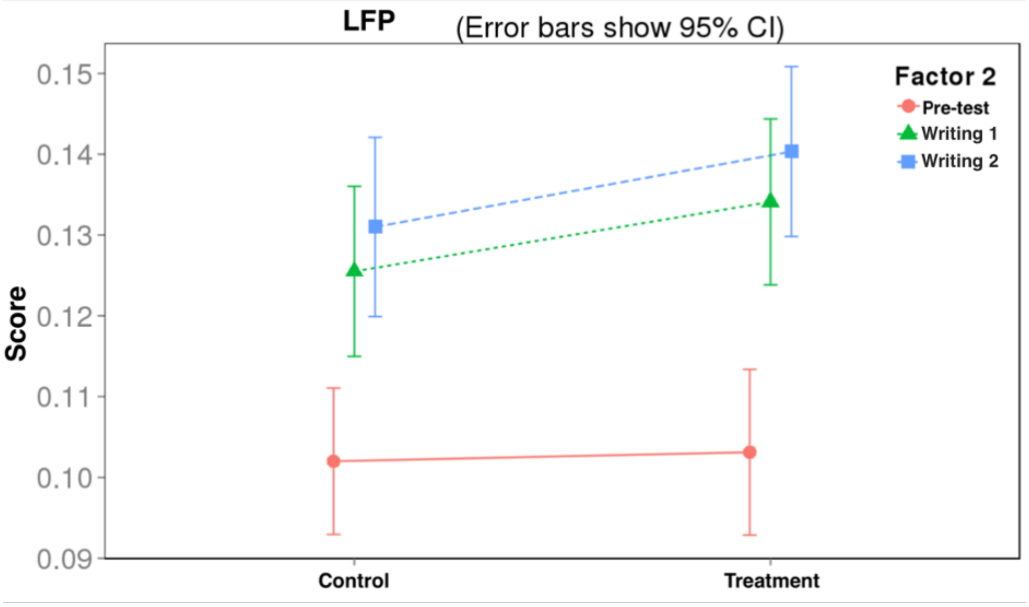


Figure 22b. Impact of Control and Treatment on measures of writing quality.

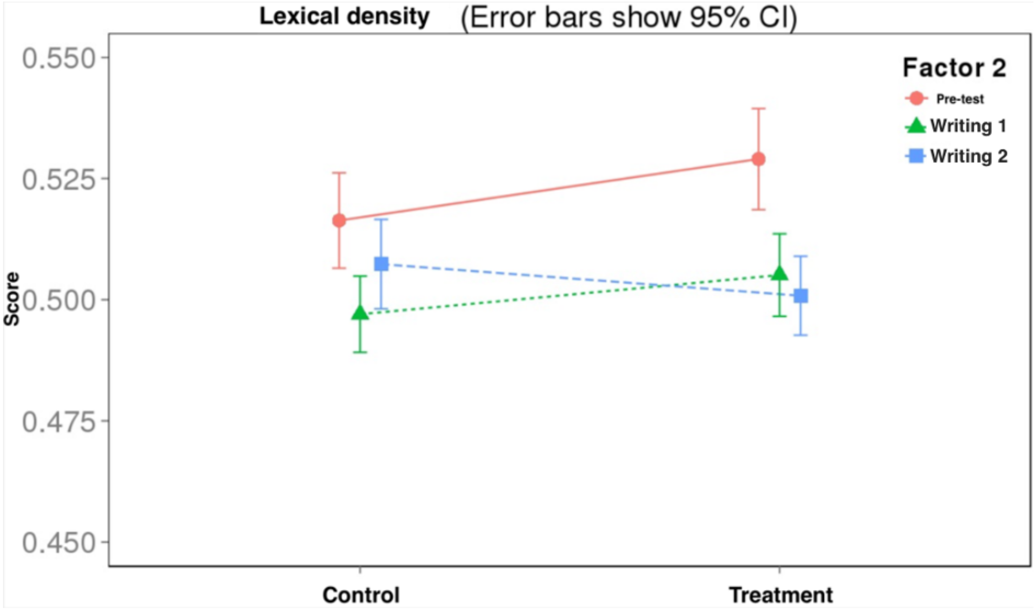


Figure 22c. Impact of Control and Treatment on measures of writing quality.

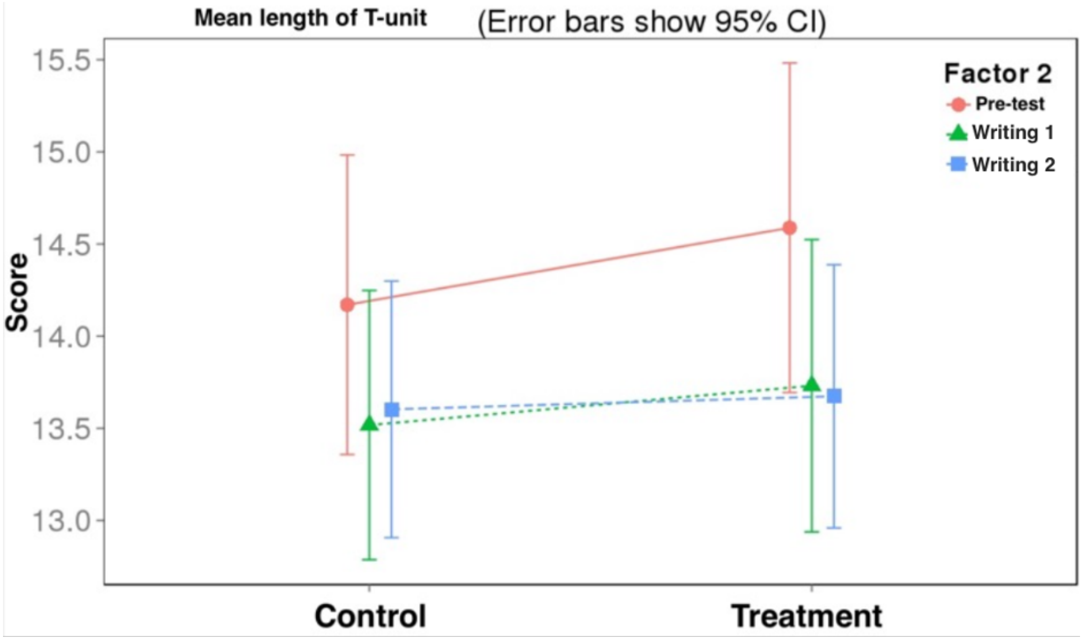


Figure 22d. Impact of Control and Treatment on measures of writing quality.

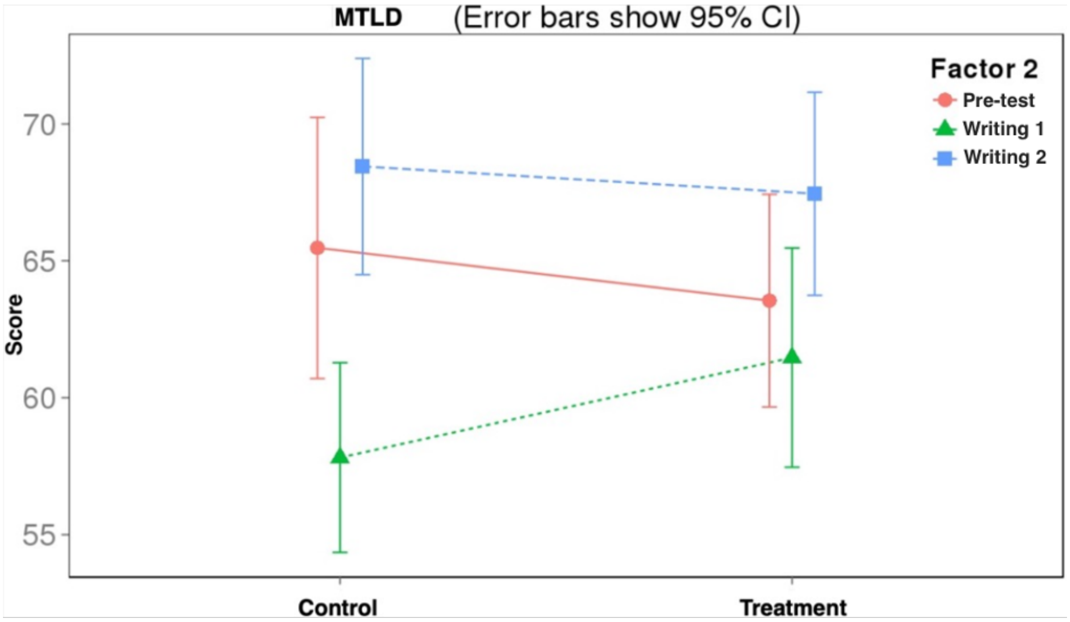


Figure 22e. Impact of Control and Treatment on measures of writing quality.

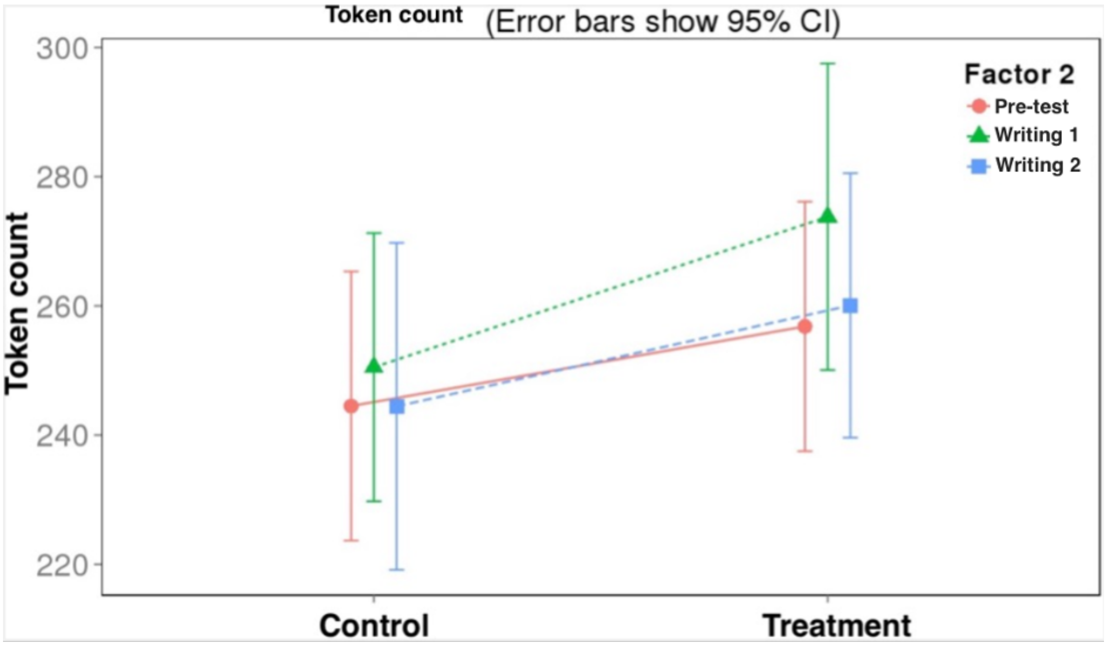


Figure 22f. Impact of Control and Treatment on measures of writing quality.

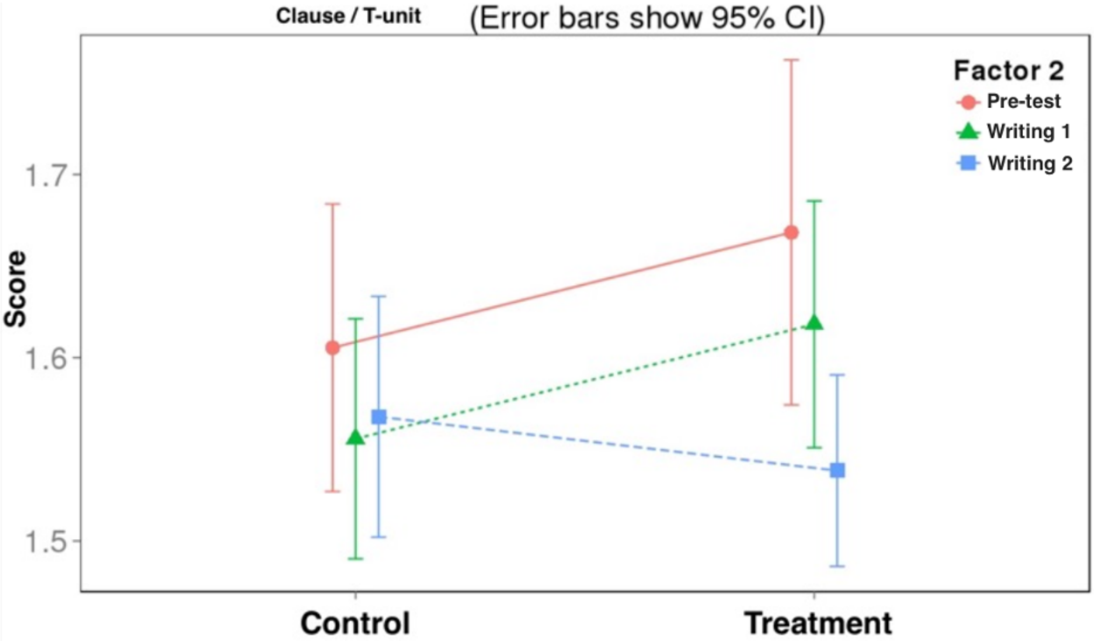


Figure 22g. Impact of Control and Treatment on measures of writing quality.

The post-hoc analysis reveals that the LFP dimension did not show a statistical difference between the control and treatment groups. However, writing 1 and writing 2 outperform the pre-test writing condition. This indicates to the researcher that the training both groups received after the pre-test had some positive effect on the participants' writings. The mean length of T-unit (MLTunit) shows no significant difference between control and treatment groups but indicates better performance with the pre-test writing task than writing 1 and writing 2. This might indicate potential fatigue that some participants reported from doing three writing tasks consecutively.

Table 15. Holm Bonferroni post-hoc analysis of significant factors.

	Pre – writing 1	*Pre – writing 2*	*Writing 1 – Writing 2*
LFP	6.8 (0.000)* <	7.4 (0.0000)* <	1.3 (0.165)
ETS	2.3 (0.01)* <	2.8 (0.004)* <	0.3 (0.7)
ETS@Control	2.2 (0.07)	1.6 (0.09)	0.58 (0.55)
ETS@Treatment	5.4 (0.0000)* <	6.3 (0.0000)* <	0.08 (0.9)
Ldensity	5.3 (0.0000)* >	4.4 (0.0000)* >	0.9 (0.33)
Ldensity@Control	3.5 (0.0007)* >	1.5 (0.13)	2.4 (0.036)* <
Ldensity@Treatment	3.9 (0.0004)* >	4.7 (0.0000)* >	0.9 (0.35)
MLTunit	2.4 (0.03)* >	2.5 (0.03)* >	0.05 (0.95)

Note: *p <.05, **p <.01, ***p <.001; >, < indicates direction of significance. *T*-values, *p*-values are in ().

The ETS dimension demonstrates significance between the control, treatment conditions, and pre-test and writing 1 and writing 2 tasks. Both writing 1 and 2 tasks perform better than the pre-test while under the treatment condition, while there is no significant difference in performance between pre-test and writing 1 / 2 in the control condition. The researcher can infer that the metacognitive training, prompting, and AI KAKU's assistance positively affected the treatment participants.

Lexical density (Ldensity) also shows mixed results with significance between control, treatment conditions, and pre-test and writing 1 / 2 writing tasks, albeit with less consistent directionality. Lexical density is improved in the pre-test compared to writing 1 in the control condition and improved compared to writing 1 and 2 in the treatment condition, while also showing writing 2 outperforming writing 1 in the control condition.

4.6.4 Metacognition training effects

Since MWSQ was modified, translated, and administered to participants that are considerably different from those tested during MWSQ's validation, the modified scale's reliability was tested using Cronbach alpha. Both pre-test (0.809) and post-test (0.83) have Cronbach alpha scores greater than 0.75, suggesting good internal consistency. The normality was also tested using the Shapiro-Wilk normality test, resulting in p-values of 0.0002 and 0.0481 for pre-test and post-test, respectively, indicating normal distribution. A two-way ANOVA test conducted on the survey results (see Table 16 and Figure 23) shows statistical significance between the treatment and control conditions in addition to significance between pre and post-test.

Table 16. Two-way ANOVA analysis of MWSQ results.

	F-ratio (p)	*p.eta²*
(A) Control – Treatment	4.6 (0.04)*	0.20
(B) Pretest – Post-test	40.8 (0.0000)***	0.69
(A)(B) Interaction	1.4 (0.23)	-

Note: *p <.05, **p <.01, ***p <.001

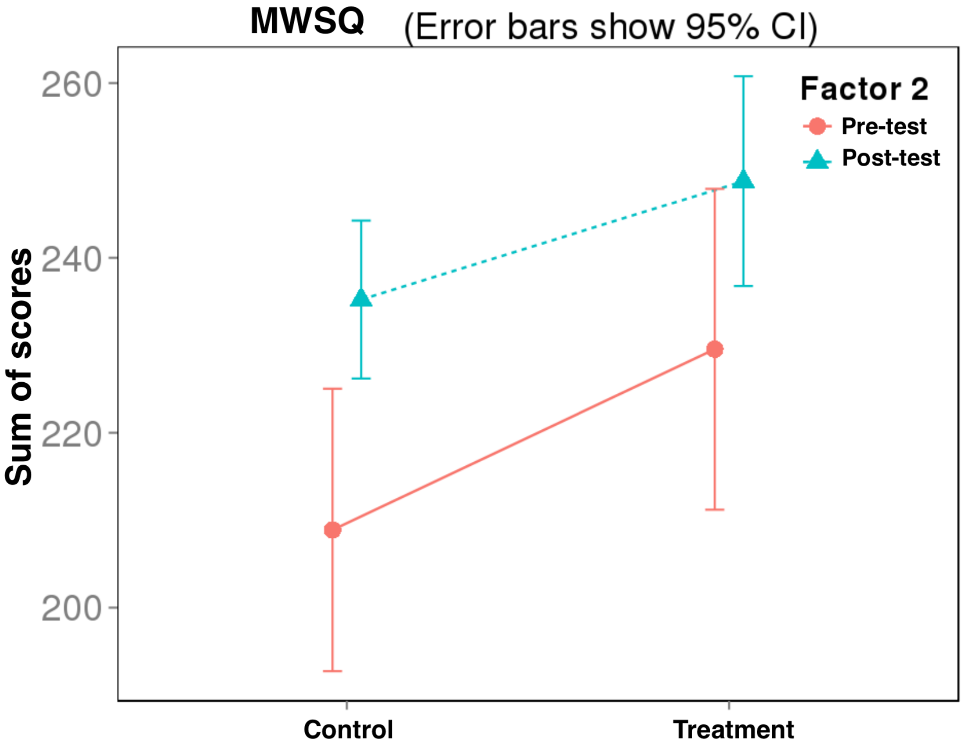


Figure 23. ANOVA results for pre and post MWSQ survey results.

Following Cohen's (2013) guidelines, the difference between both control and treatment and pre-test / post-test show large effect sizes, indicating to the researcher that participants in both the control and treatment groups were able to improve their metacognitive awareness of the writing task after metacognitive training.

4.6.5 Relative Importance Analysis

Taking the results from the writing quality dimensions and the MWSQ inventory, the researcher conducted a relative importance analysis to gain insight into which factors contribute to a higher ETS score. Using the ETS score as the dependent variable, a Random Forest Boruta analysis identifies the variables that are important to the dependent variable. Mizumoto's (2022) seminal work in this area of quantitative research methods identifies the

Random Forest Boruta as a novel type of analysis that clearly outperforms more traditional feature selection techniques such as regression models,

> […] This is because random forests have been suggested as an approach that produces more accurate estimates of predictor importance than do standardized beta coefficients […] Boruta is a novel feature ranking and selection algorithm based on random forests. It runs random forests many times (maximum 100 times by default […] In the same way that random forests are much more accurate than a single decision tree, the Boruta algorithm in general yields more precise estimates of predictor importance than does an ordinary random forest procedure (pg. 19).

According to Mizumoto, this type of analysis has not been widely used in the field of applied linguistics but is growing in popularity due to its performance improvements over older regression methods. As seen in Figure 24, the factors of Tokens (word count) and LFP (lexical frequency profile) are confirmed as factors that were important to the participant achieving a high ETS score. The interpretation of the results from the analysis is intuitive. The more words the participant produced, and the vocabulary level of the words used, were principal factors that led the human raters to assess the writing sample at a higher level.

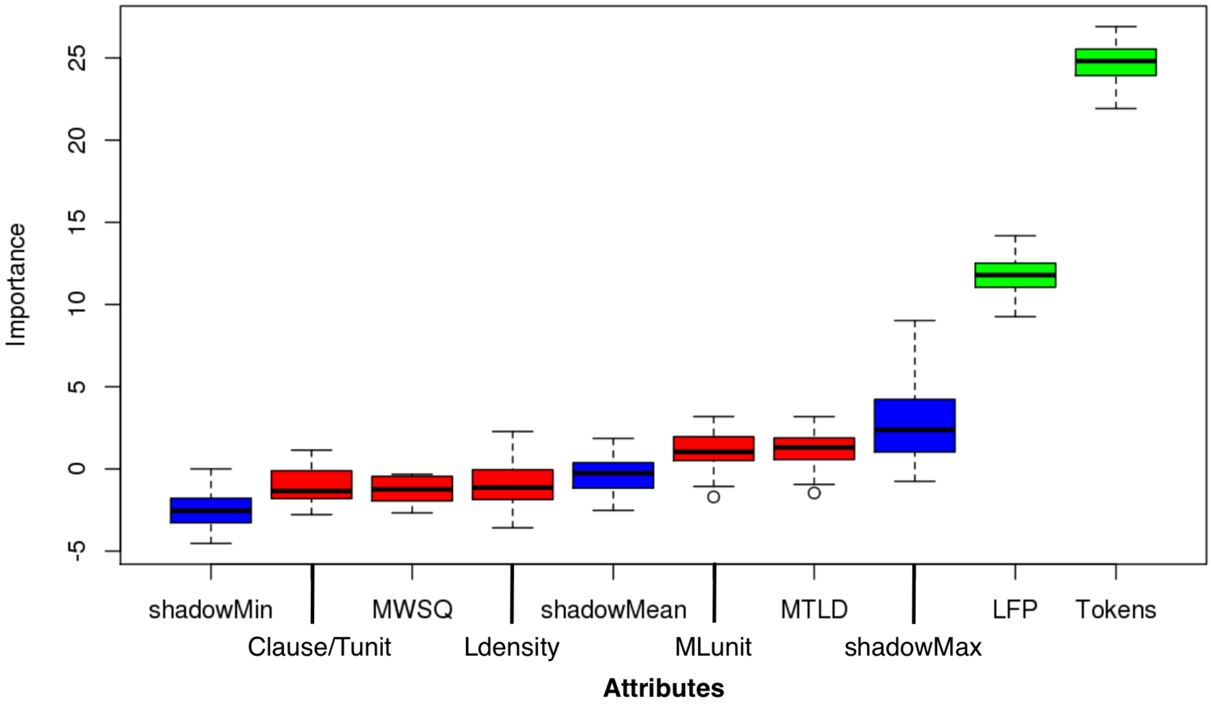


Figure 24. Boruta (Random Forest) analysis of variable importance to ETS score.

While the machine assessment dimensions give the researcher insight into some of the mechanics of writing quality that can be quantitatively analyzed, human assessment is still considered the gold standard when insight into the overall quality of writing is needed. Therefore, understanding which factors lead to better human assessment enables the researcher to fine-tune treatment tools to be used in future studies.

4.7 Participant Feedback

To gain insight into participants' opinions regarding the AI KAKU writing assistant, qualitative data was gathered via a short survey given to the participants at the end of the experiment. Responses to a 5-point Likert survey (N = 60) from the treatment group and open-response comments (N = 44) from all the participants were collected. The two survey questions (5-point: Likert; 1 = "Strongly Disagree"", 5 = "Strongly Agree") regarding AI KAKU's features given in the survey were, "Q1: The word suggestions given to me were useful" and "Q2: The translation of my English displayed to me helped me with my writing". As seen in Figure 25, responses to Q1 were largely positive, with 72% of responses being "agree" to "strongly agree"; similarly, Q2 showed strong positive responses with 77% being "agree" to "strongly agree."

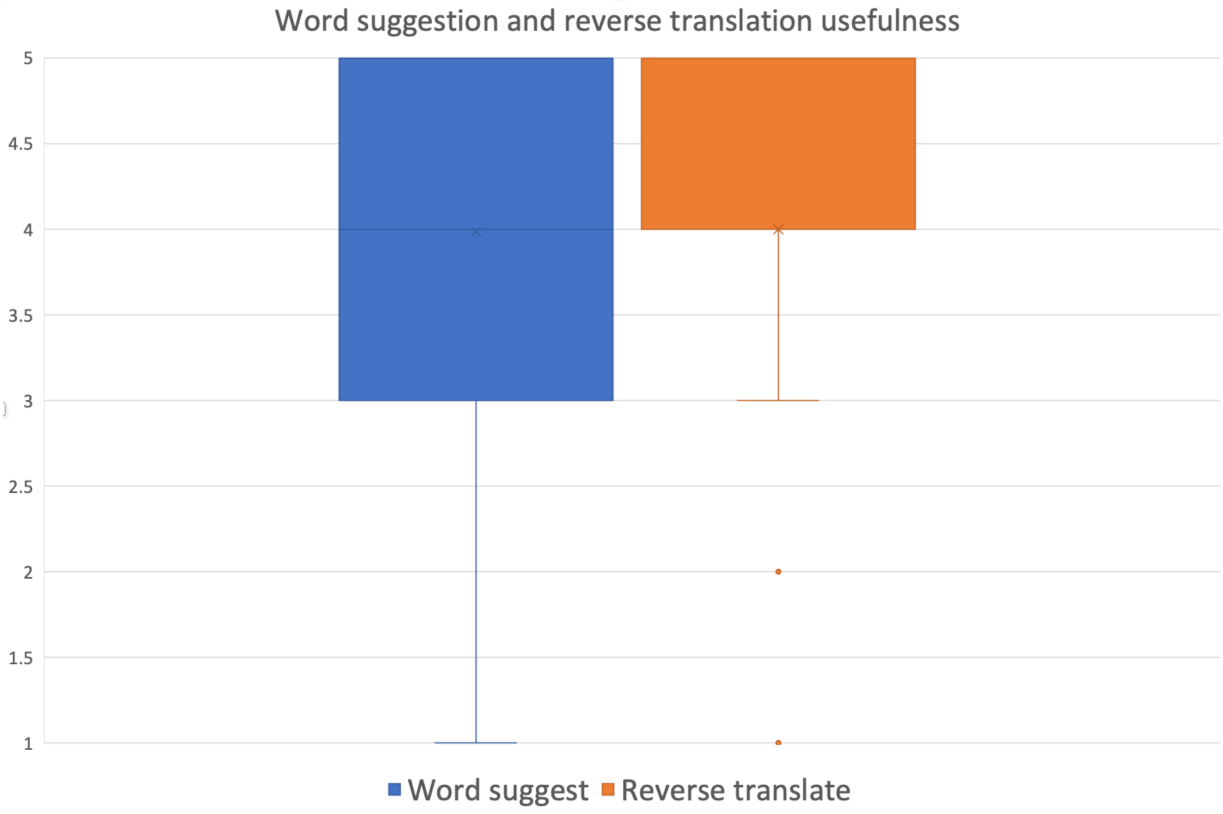


Figure 25. Post-experiment feedback on AI KAKU.

Content analysis of the respondents' qualitative feedback was done via inductive coding protocols where the coding themes are derived from the data (Ezzy, 2013; Hsieh & Shannon, 2005); in addition, the research follows Campbell et al. (2013) and O'Connor & Joffe (2020) guidelines (as demonstrated in Dizon et al. (2022) that allow for single coder analysis of qualitative data. A summary of the themes that surfaced from the feedback are detailed in Table 17. The complete list of 44 comments received from the participants in the study are included in Appendix I (User Feedback on AI KAKU) with additional sentiment analysis (assigned as positive, neutral, and negative). Figure 26 visualizes the top frequency words and some of the relationships between words in the feedback. The word cloud reveals some common themes among the participants, including task difficulty, a desire for improvement, more granular feedback, and pre-task planning.

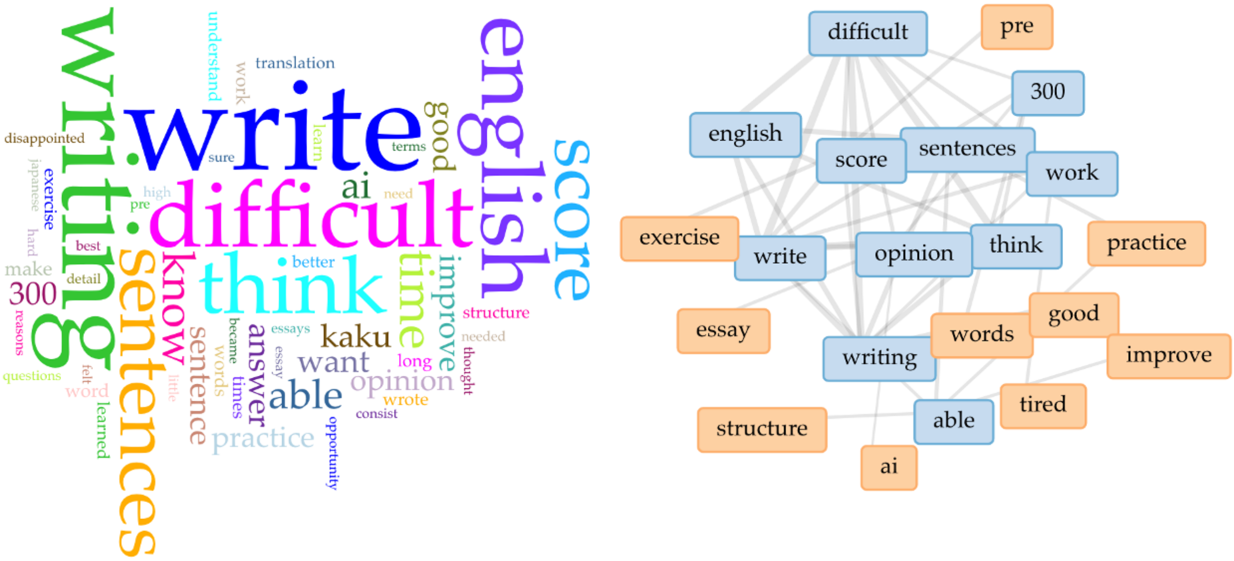


Figure 26. Word cloud and word relationships.

After reviewing the 44 responses, the researcher extracted five themes to code the responses through inductive thematic analysis. Inductive thematic analysis is a process that,

> …involves the identification of themes through "careful reading and re-reading of the data" (Rice & Ezzy, 1999, p. 258). It is a form of pattern recognition within the data, where emerging themes become the categories for analysis. (Fereday & Muir-Cochrane, 2006, p. 82)

The identified themes are: 1. Writing difficulty; 2. Gained writing skill; 3. AI KAKU as valuable; and 4. Insufficient or desire for more support. These themes are then coded as 1. Difficulty; 2. Gain; 3. Valuable and 4. Insufficient. Examples of each theme are given with the theme's ratio to all themes indicated as a percentage. Additionally, the keywords that led the researcher to determine its theme are underlined to show the rationale the researcher took to make those determinations.

Table 17. Coded summary of feedback received.

Theme	Percentage	Example
Difficulty	41	It is difficult for me to write 300 words within 30 minutes.
Gain	48	I was able to be more aware of the paragraph structure and connection than the first time I wrote it, so I think the sentence became more cohesive.
Valuable	16	I learn how to write my opinion thanks to AI KAKU.
Insufficient	20	I wanted [it] to be specific about how to improve my score

4.8 Summary of effects and population implications

The empirical studies in this dissertation include three experiments that measure the effects of the treatment tools developed for this research on a sample population of EFL learners. For all three experiments, the researcher used convenience sampling (Patton, 2002) when recruiting participants for the experiments. The use of convenience sampling is common in the social sciences (Dörnyei & Griffee, 2010; Given, 2008), including research in Computer Assisted Language Learning (CALL) (see Cirocki & Caparoso, 2016; Kılıçkaya, 2022; Zaker, 2015).

However, this is a non-probabilistic sampling method in which participants are selected based on the ease of access to them, rather than by random selection. This means that the sample may not be representative of the larger population, as it may only reflect the characteristics of the individuals who happen to be conveniently available at the time of the study (Andringa & Godfroid, 2020). While the importance of external validity is a topic that has been debated in the literature (Mook, 1983), researchers should keep in mind that the main disadvantage of convenience sampling is that it is prone to bias (self-selection bias, response bias) and thus has the potential to be less representative of the population being studied. Thus, the conclusions that can be drawn from the experimental studies in this research are limited, and the results should be interpreted with caution. In this dissertation and in the published materials associated with it, commonly held recommendations such as describing the selection criteria and the sampling purpose, as well as clarifying the limited scope of the studies were observed (Dörnyei & Csizér, 2012). These constraints are reiterated again in the summary chapter 7.5.1 Constraints and Future Work.

Importantly, since the definition of the representative population in CALL research is difficult to clearly delineate, the researcher followed best practices to strengthen the relationship between the experiment participants and the general population being studied. The three experiments in this research recruited English as Foreign Language (EFL) participants studying in Japan as adults or as university-level students, and thus, the target population of this research reflects this demographic. Even so, the categorization of EFL contains a wide spectrum of learner (Broughton et al., 2002) that could include participants who have never lived outside of their host country to bilingual participants or "returnee" participants who have spent significant time interacting with a second language outside of the host countries. Given the "fluidity" of the general population being studied (EFL learners), the researcher took further steps (repeated measures research design, pre-testing, ANOVA testing) to ensure that outliers in the participants would not disproportionality influence the results (see sections 2.4 Methods; 3.4 Methodology; 4.5.1 Research design) of this research.

Non-probabilistic sampling has been used in Applied Linguistics research due to the very nature of the population the field is trying to study. Some research areas cannot be investigated with randomized sampling, such as quasi-experiments or action research using enrolled students. This prevalence has been documented by Amini Farsani & Babaii (2020) in their systematic review of MA theses spanning thirty years. The results of their research (see Figure 27) shows 91% of Applied linguistics using non-probability sampling.

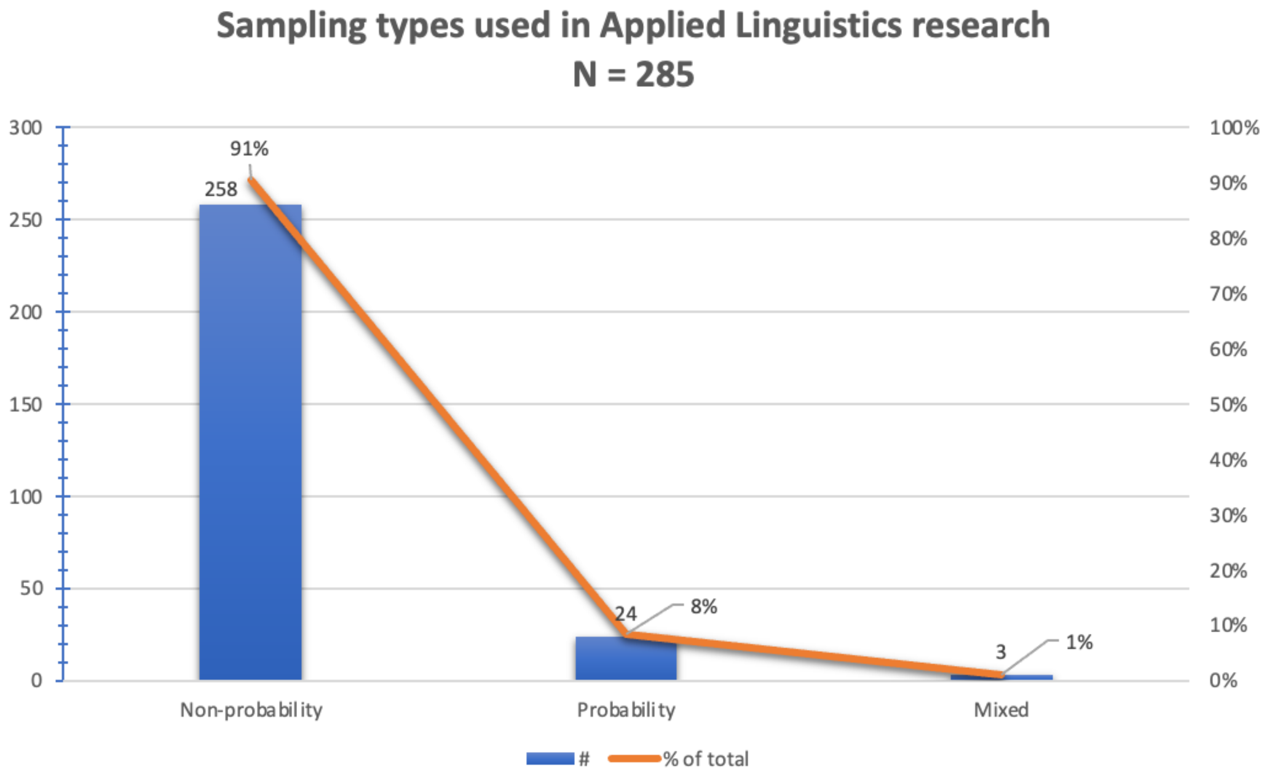


Figure 27[8]*. Sampling techniques used in Applied Linguistics.*

Yet, given the widespread use of this sampling technique, much research has also been done regarding the research outcomes that use convenience sampling. Non-probabilistic sampling can be useful in certain situations where it is not feasible or practical to use other types of sampling methods (Farrokhi & Mahmoudi, 2012), such as when the population is difficult to access, when a large sample size is not required, or when the representative population itself does not have clear boundaries. As Farrokhi & Mahmoudi (2012) point out, using convenience sampling has its limitations but those limitations can be mitigated when the researcher clearly describes the methodology and experimental factors used in their research. One such good practice they advocate for is leaving a proper "audit trail", regarding this they elaborate specifically on the sampling stage of research by stating,

> […] This audit trail is perfectly applicable to the sampling stage of quantitative research as well, especially where non-random groups are used for research purposes. Describing the conditions under which the investigation was carried out removes a lot of misinterpretations of or overreadings from the research results… [we] encourage reluctant researchers to conduct inquiries even though they feel that they do not have access to comparable groups. The only thing required, however, is

[8] Adapted from Amini Farsani & Babaii (2020). Applied linguistics research in three decades: A methodological synthesis of graduate theses in an EFL context. *Quality & Quantity, 54*. https://doi.org/10.1007/s11135-020-00984-w

> to precisely report the circumstances in which the research was conducted. (p. 797)

This researcher carefully weighed the advantages and disadvantages of using convenience sampling, and decided it was the best way to conduct the experiments needed with EFL learners. Access to the learners via the researcher's host institution was a key factor that enabled this research to go forward while following Farrokhi and Mahmoudi's (2012), Vitta & Al-Hoorie (2021), and Moranski & Ziegler (2021) guidelines. By being mindful of the limitations of convenience sampling and using it in conjunction with other research methods described earlier, the researcher was able to improve the validity and generalizability of the findings.

The significant outcomes from this research are summarized in Table 18 and the effect size interpretations based on Plonsky & Oswald (2014) guidelines gives the researcher a better idea of which factors may been seen in similar populations. Factors that demonstrated "large" effect size have a greater likelihood of having practical significance outside of the experimental setting.

Table 18. Summary of significant outcomes.

Factor	***Test***	***n***	**p-*value***	***Effect size***	***Effect size interpretation***
Clause/t-unit	U-test	10	.03	1.05 (Cohen's D)	large
Tokens	t-test	90	.004	0.2 (Cohen's D)	small
Intrinsic load	t-test	90	.03	0.13 (Cohen's D)	small
ETS rating	ANOVA	197	.0001	0.12 (p.eta)	large
MWSQ (A/B)	ANOVA	197	.04	0.20 (p.eta)	large
MWSQ (pre/post)	ANOVA	197	.0000	0.69 (p.eta)	large

4.9 Conclusion

This chapter focuses on the combination of the use of a novel digital writing assistant (AI KAKU) with metacognitive training and prompting in an experimental setting. As to the first research question, the study results demonstrate that writing assistance and metacognitive training benefit L2 writers. Namely, results from the human assessment of the writing samples showed that the treatment condition had a strong positive influence. In addition, we can see the lexical sophistication of the writing samples improved on both writing 1 and 2, as shown by the LFP dimension. Lexical density and the mean length of T-unit were not in line with the researcher's expectations, with both dimensions showing better performance in the pre-test task. Further analysis is needed to understand why these measures showed negative results.

However, according to the relative importance analysis, both lexical density and mean length of T-unit were not important factors that led to a high ETS score.

Reflecting on the results regarding the second research question about metacognitive awareness, participants improved on the MWSQ inventory after receiving metacognitive training in both control and treatment conditions. While the relative importance analysis did not show MWSQ performance impacting the participants' ETS score, the participants only received one training session via the experiment's website. Further research is needed into how prolonged and sustained metacognitive training, prompting, and nudging influence writing quality.

Qualitative feedback gathered in the study was largely positive, with participants indicating metacognitive training and AI KAKU as valuable tools that can help improve L2 writing. The researcher intends to develop digital writing aids for L2 users further. Further globalization and internationalization in engineering education can be a catalyst to support EFL students via novel techniques and tools.

4.10 Acknowledgements

The researcher would like to extend his thanks to the three raters who contributed significant time in rating the writing samples, members of Cross laboratory for their support and to the GSEC administrators and faculty at Tokyo Tech who supported this study.

Chapter 5 Policies and applications of AI in education: A perspective from two advanced countries.

Keywords: *AI in education, AIED, education policy, Computers and education*

5.1 Background

Taking a step outside of the narrowly focused empirical studies in Chapter 2, Chapter 3, and Chapter 4, this chapter will discuss AI in education policy as a general topic in order to have a wider view of trends and policy level initiatives surrounding artificial intelligence in society. Specifically, the researcher looks at the application of AI and how the United States and Japan have used these systems. Recent trends in educational technology suggest that AI in Education (AIED) is an emerging and potentially disruptive field that will have a vast impact on both learners and educators. How to use AI to make clear pedagogical progress is still in its infancy. At the same time, broader issues such as how AI will impact learning, and the ethical considerations of human-machine output are also unclear. The author intends to raise awareness around the application of AI in education and call on researchers, developers, and educators to consider the ethical, pedagogical, and human factors as this technology progresses rapidly in the field.

5.2 Introduction

AI (artificial intelligence) as a set of technologies has been given much exposure in the media and an ever-growing investment from governments, non-profit organizations (NPOs), and private businesses. According to a report on venture capital investment in the United States by Venture Beat, "... data from the National Venture Capital Association, 1,356 AI-related companies in the U.S. raised $18.457 billion…[topping] the 1,281 companies that raised $16.8 billion in 2018" (O'Brien, 2020). Moreover, investment in education-related AI is also projected to increase over the next decade significantly. A report by Prescient and Strategic Intelligence, (2020) shows global investment in AI in Education to reach over $25 billion, with most of the capital being invested in North America or the APAC (Asia Pacific) regions. The AI systems we have today are in many ways quite limited as they are confined to performing very narrow tasks with little adaptable intelligence that humans exhibit (Southgate et al., 2019). Nevertheless, advancements in the field are progressing rapidly while its use and influence in education and beyond are growing.

These advancements in AI have given us programs and machines that can perform independently with minimal external input from humans to support or operate them. Artificial intelligence is efficiently improving learning and accelerating access and equality in the education system by providing extensive help to students and teachers (Taguma et al., 2018). This paper will highlight examples where AI is used to enhance courses and materials for students, helping education become more learner-centered and providing students with an environment that is more conducive to knowledge acquisition. Artificial intelligence also helps educators enhance their classroom management and assessment tasks, allowing the educator to focus on their own pedagogical and content knowledge and gives them the latitude to update those skills (Vincent-Lancrin & van der Vlies, 2020). AI has the potential to help educators save time with applications that can automate grading, scheduling, identifying students that need more attention, and allow educators to monitor student progress with a "guide on the side" to help them.

In addition, AI promises to enable universal access to knowledge so that learners can learn without restrictions to time and place. With AI assistants, students are not restricted to be physically present in the classroom as MOOCs and virtual classroom platforms such as edX, Coursera, Canvas, and Moodle are starting to incorporate AI technologies that enhance virtual education. The sudden emergence of emergency remote teaching and learning that the COVID-

19 pandemic brought upon education institutions around the world has accelerated thinking around how educational technologies can supplement, assist or even replace traditional teaching paradigms (Bond et al., 2021; Ferri et al., 2020).

Another trend that has emerged is AI-based tutors that provide learners personal teaching assistance. This allows the student to learn from virtual sources, exposing students to different teaching methods instead of learning from just the traditional "teacher on the stage" paradigm that is often restricted to the teacher's pedagogy. Suppose the student does not understand the essence of a particular course or is not satisfied with their learning and wants to learn more. In that case, they can expand upon their learning with an AI-based tutor, accessing the material via multiple methods for conceivably a much lower cost than traditional in-class instruction.

Indeed, the main objective of these technologies is to improve the education system by efficiently providing a more comprehensive range of content, access, and teaching resources. Students can access content that is available via online platforms, enabling a greater sense of autonomy and agency in their studies. Artificial intelligence in education can create a virtual or hybrid educational platform to enable the greatest number of learners to receive the benefits of education. Far-reaching implications of these technologies can impact students who cannot afford the traditional educational experience or cannot physically attend classroom-based instruction due to limitations imposed on them.

5.3 Research Question

This exploratory chapter is in no way exhaustive of all the topics surrounding AI in education but attempts to give a brief overview of some of the technologies being used. The researcher addresses the questions:

1. What are some of the current trends and potential ways AI impacts education and influences learning?
2. What are the approaches from a policy level regarding AI in the United States and Japan?

The economies which will be covered in this paper are the United States and Japan. These economies were not chosen due to their progress in AI in education but rather for their current level of digitalization with different levels of momentum going forward. It can be argued that a minimum level of digitalization is necessary for the use of AI applications in the field. The countries selected have advanced digitalization levels, as seen in Figure 28. According to the report's authors, "Stand Out" economies currently exhibit high levels of digitalization and have strong momentum in investment and deployment of digital initiatives. "Break Out" economies are economies that do not currently have highly developed digital infrastructures but are digitalizing at a rapid rate, with China being a notable extreme with dramatic levels of demand and innovation. Japan falls under the "Stall Out" category which is characterized as an economy that has well-developed digitalization but does not show the same level of momentum as "Stand Out" or "Break Out" economies. Lastly, "Watch Out" countries are characterized as economies that face significate challenges with currently low levels of digitalization and relatively low momentum when compared to their peers. (Chakravorti et al., 2020).

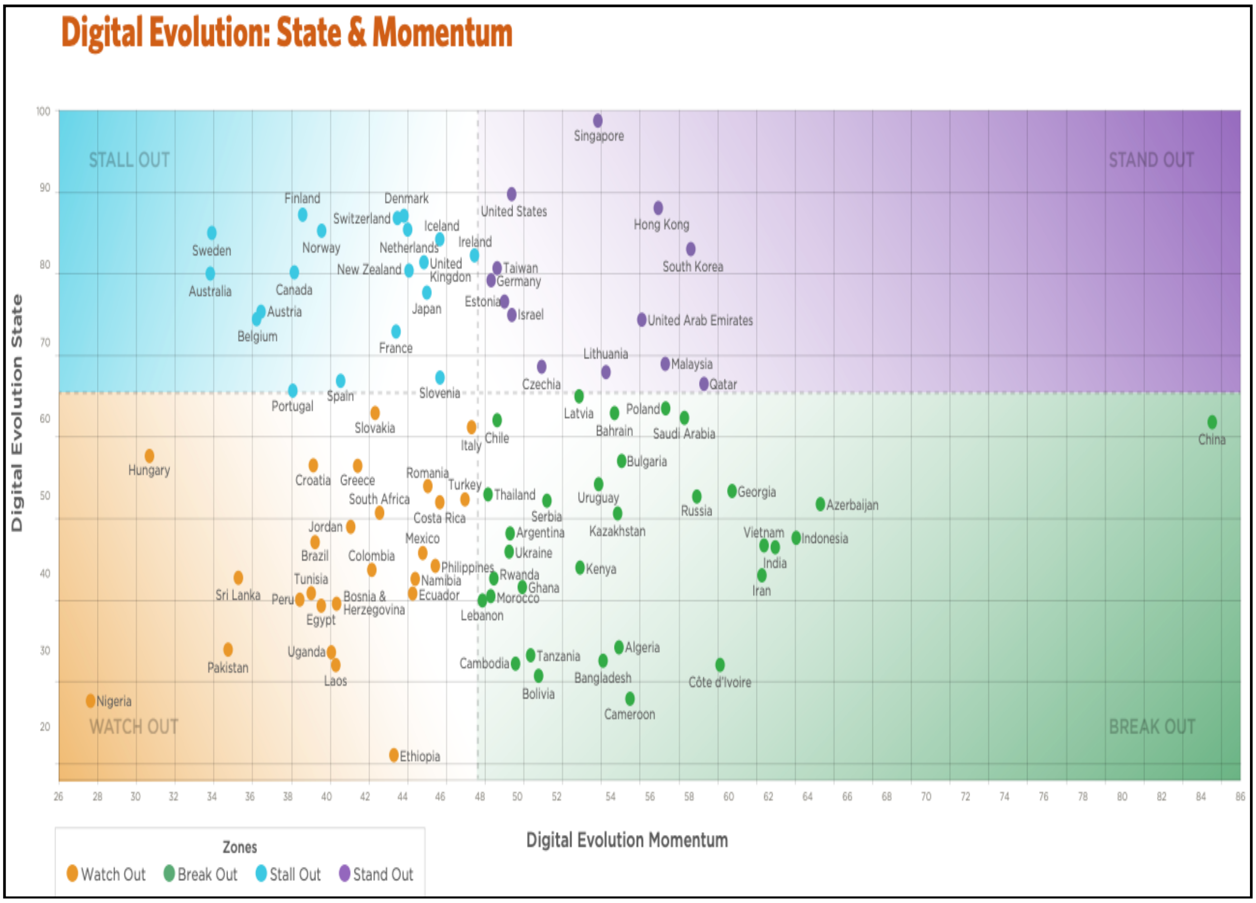


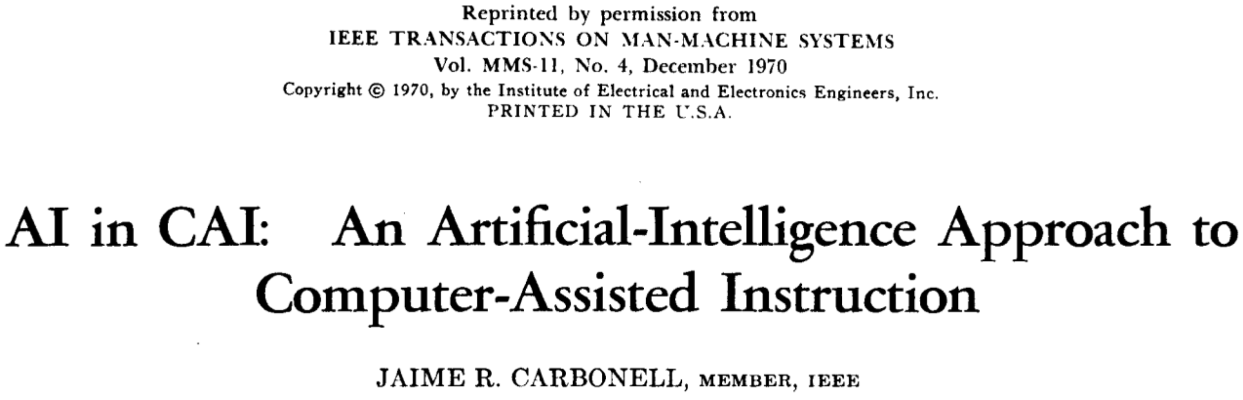
Figure 28[9]. Digital in the time of COVID.

5.4 Artificial Intelligence in Education (AIED)

Artificial intelligence is widely thought of as a new technological development in education, with progress and attention to this area only happening recently. However, some of its roots

[9] By Chakravorti, B., Chaturvedi, R. S., Filipovic, C., and Brewer, G. The Fletcher School at Tufts University, 2020. p.8. Trust in the Digital Economy and Its Evolution Across 90 Economies as the Planet Paused for a Pandemic. Retrieved from https://sites.tufts.edu/digitalplanet/files/2020/12/digital-intelligence-index.pdf. Reprinted with permission.

can be traced back to the early 1970s. One of the first initiatives using artificial intelligence in the field of education was conducted in the United States by the name of "SCHOLAR" CAI (Computer-assisted instruction; see Figure 29), an intelligent teaching system (Carbonell, 1970; Collins & Grignetti, 1975).



Reprinted by permission from
IEEE TRANSACTIONS ON MAN-MACHINE SYSTEMS
Vol. MMS-11, No. 4, December 1970
Copyright © 1970, by the Institute of Electrical and Electronics Engineers, Inc.
PRINTED IN THE U.S.A.

AI in CAI: An Artificial-Intelligence Approach to Computer-Assisted Instruction

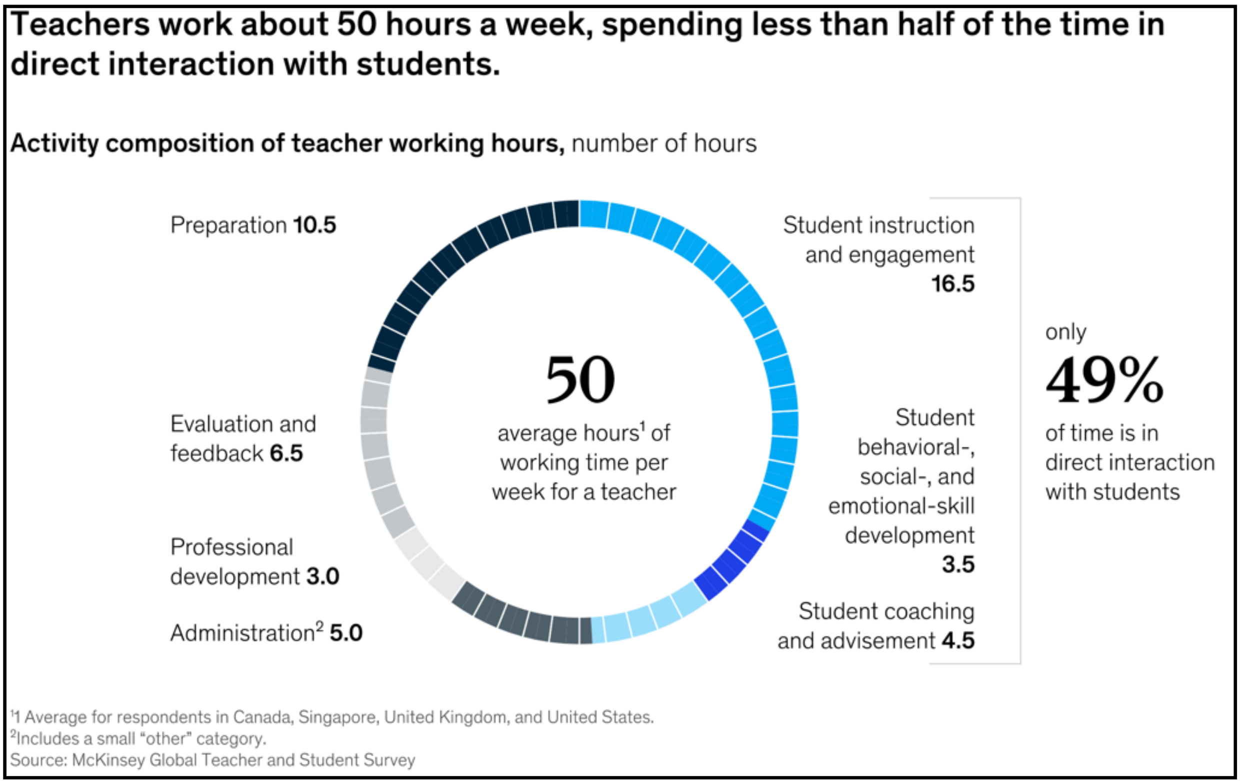
JAIME R. CARBONELL, MEMBER, IEEE

Figure 29. Image of Carbonell's paper published by IEEE, 1970.

This computer-assisted program was developed to determine how human questions and answers were structured to create a model that mimics how human tutors interact with students. State and private actors are implementing artificial intelligence to improve many aspects of society. One of the objectives is to improve the education system, making it more efficient and effective for learners (Roll & Wylie, 2016). The use of AI is beneficial for learners and students, but it has the potential to help educators in the classroom while giving them the latitude to develop new skills.

Teachers can enhance their level of content knowledge and pedagogy of their subject area with artificial intelligence technologies. The education system can become more proficient by using artificial intelligence to modify itself. Education bodies are beginning to use the capabilities of artificial intelligence to accelerate the level of education in the world (Ocaña-Fernández et al., 2019). As artificial intelligence continues to develop over time, we can see its role in the classroom expanding in numerous ways. It enables students to access new/expanded curriculum by using online platforms that incorporate AIs (Popenici & Kerr, 2017). In this way, it helps fill the gaps between the learning outcomes and potential deficiencies in the learning environment by providing teachers and students an avenue to enhance their knowledge and capabilities, increasing their proficiency and credibility in their respective domains (Bates et al., 2020). It promotes student-centered learning, allowing learners to progress on their own time and provides them multiple options for understanding by providing access to a broader variety of content structured to a spectrum of levels. The central vision of artificial intelligence in the education system is to simplify education and improve its effectiveness for students (Bose & Khan, 2020).

As shown in Figure 30, McKinsey and Company's (2020) data collected from K-12 educators across four countries indicates that teachers spend the preponderance of their time on activities that are not directly related to instruction (Bryant et al., 2020). In that regard, we can see using AI in the classroom as having the potential to reduce the load (administration, evaluation, preparation) on educators. Educational institutions play a vital role in enhancing the credibility of using AI by making their students familiar with the technology and educating them about using the technology efficiently and appropriately (Tilak, 2020).



Teachers work about 50 hours a week, spending less than half of the time in direct interaction with students.

Activity composition of teacher working hours, number of hours

[1] 1 Average for respondents in Canada, Singapore, United Kingdom, and United States.
[2] Includes a small "other" category.
Source: McKinsey Global Teacher and Student Survey

Figure 30[10]. McKinsey Global Teacher and Student Survey.

[10] Taken from McKinsey Global Teacher and Student Survey. Data showing time spent on the job from educators in four countries. Reprinted with permission.

5.4.1 Virtual Platforms

Educators spend a considerable amount of time preparing high-quality teaching materials related to their specific courses and modules. Digital platforms that implement AI have been shown to help educators produce and organize curriculum materials. AIs used to analyze target knowledge databases and related works, giving the instructor course outlines and teaching insights that they can implement in their classrooms. It also helps teachers by providing them a platform to develop and polish their abilities to enhance their professional performance (Colchester et al., 2017). Teachers are also challenged with keeping their students up to date on new subject-area concepts, but with the assistance of virtual platforms that implement AI, they can bridge that gap more easily. Some existing virtual platforms that implement AI technologies in this framework include:

- Knewton CO.'s altar is an adaptive learning platform for higher education that allows instructors to teach core academic courses at different levels of difficulty.
- Cognii is a virtual learning assistant that allows open-format answers and provides real-time feedback to students.
- Querium uses AI in its virtual platform that focuses on STEM subjects. The platform analyses student responses and time to answer and relates to data back to the instructor. This gives the teacher more insight into the student's learning and where they need to improve.
- Quizlet uses an AI in its Quizlet Learn product. The platform provides adaptive lesson plans and customized learning for the student. It uses machine learning on large data sets from millions of users to finetune its study suggestion/recommendation engine.

5.4.2 Automated classroom activities

AI is currently being used in systems such as Turnitin and Criterion's e-rater, applications that use machine learning to assess essays written by students automatically. The systems analyze the texts based on the context, credibility, and uniqueness of the content, thus saving the teacher a great deal of time from checking and assessing submitted works manually. In addition, using these systems reduces the chances of human error or bias from the grader when assessing student submissions. Ultimately, giving the educator more time to spend on direct interaction with students and improving course quality and modernity. Education technology often must balance the affordances they provide to students in their studies and potential for misuse. Academic dishonesty is an age-old problem that becomes increasingly difficult to address as technology progresses. Yet, these issues are not being ignored. Technological solutions to these problems need to work with pedagogical and professional development programs (Aaron & Roche, 2013) in order to ensure dishonesty is not the main attraction of an educational technology.

5.4.3 Customized educational programs

Artificial intelligence is being used to customize educational programs. There are numerous software and online platforms available that utilize AI, which is ameliorating education. Educational software is a leading example of the usefulness of artificial intelligence when it can be customized according to the requirements of teachers and learners. Educational institutions are also investing in these systems to customize them for their respective courses and modules. This enhances the institution's credibility by providing an extensive range of accessible material to students that maintain their interest and engagement (Eryılmaz et al., 2019). Some examples of artificial intelligence solutions are Dream Box, Achieve3000, Thinker math, Brainly, Carnegie Learning, and many more. Thinker math, a platform deployed for Turkish speakers, is an excellent example of a system that uses AI to teach students the fundamentals of mathematics in a way that is accessible and can be tailored to students of varying levels. As a mathematician working in AI in education succinctly summarizes,

> *It is difficult to deny that in the modern society of knowledge and information computers have become a valuable tool for teaching and learning. The wealth of information in hands of students, the animation of figures and representations that increases the students' imagination and problem solving skills, the rich variety of data and resources that teachers can use working with their students to keep them engaged in the classroom, etc. are some of the benefits obtained by using the computers in education. (Voskoglou, 2019, pp. 129).*

Artificial intelligence used with learning platforms has equalized access to education to students who had socioeconomic barriers to gaining knowledge from traditional institutions. Those barriers can be addressed by platforms such as Brainly, an online platform for both teachers and students that enables users to ask questions that they could not solve independently. Machine learning algorithms used on the platform filter and identify spam to help the system users receive quality answers without the usual risks of disinformation, racism, sexism, and others that have been pervasive on social networking sites. They also use AI algorithms to target a specific age range for a piece of content and give students questions about what they have previously asked. This is a unique way to individualize learning for students, creating an experience specific to every user and providing a highly optimized, beneficial learning environment.

Lastly, Carnegie Learning is a suite of products that take advantage of AI. Products in the suite provide customized learning materials to learners based on their skills and the course they are enrolled in (Marr, 2018). The system's goal is to help make the process of learning more comfortable by offering personalized education for the learners. It evaluates the learner's abilities and then renders learning content based on their skills, which helps them understand the course and enables educators to generate learning material according to their level of understanding (Subrahmanyam & Swathi, 2018).

To summarize, Table 19 describes some of the types of AI being used in education examined in this study. As research and development in the field of AIED continues, educators will have the opportunity to see AIED in use-cases that are not currently deployed.

Table 19. Summary of use-cases of AIED.

Type	Student/educator focused	Example
Virtual Assistant	Student	Cognii
Automation	Educator	Turnitin
Customized Edu	Both	Carnegie Learning

With the increase in the adoption and deployment of AI technologies in and outside the classroom, state-level policies defining and framing the possible risks, negative effects, training protocols, and best-use scenarios need to considered and enacted (Raman & Rathakrishnan, 2019). Doing so will positively impact how this technology contributes to society.

5.5 AI Governance

This chapter’s second research question addresses state-level approaches and policies regarding the use of AI in education. The author believes that while innovation in the field is progressing rapidly, public policy regarding the use of these technologies is still in its beginning stages. The leading market research firm Gartner publishes yearly reports on the perceptions surrounding the state of technologies being used in numerous industries. One such area they report on is artificial intelligence and its development. We can see from projections done by Gartner (2021) that “AI Governance” is still in its very early “Innovation Trigger” stage. The stages as described by Gartner are reproduced graphically in Figure 31. According to Gartner’s research report, it is projected to take five to ten years to reach some form of maturity.

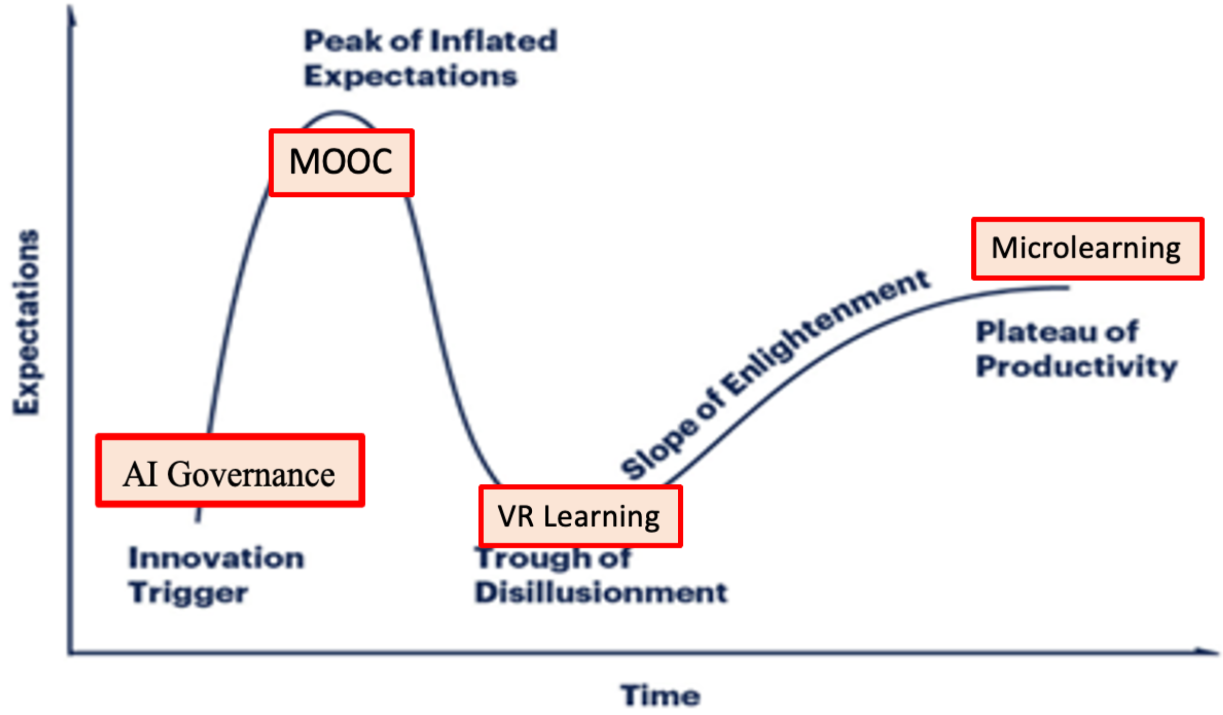


Figure 31[11]*. AI Governance in relation to other Edtech: Gartner's Hype Cycle.*

Digital technologies like the internet of things and artificial intelligence, along with information and computer technology, provide opportunities in improving the procedure and process of education. As stated earlier, global investment and development in this area are experiencing massive growth. This has resulted in having a range of digital solutions for the stakeholders and educational institutions that can provide a better educational experience. Digital technologies increase the value and production of data by creating new opportunities for improving educational policies. There remains, however, new challenges for institutions and stakeholders while society begins to implement these technologies in the classroom. These policies are being formed independently at the state level through the available research and guidance of non-governmental organizations such as the OECD.

5.5.1 Governance in the USA

The president of the United States in 2019 issued an executive order for launching an AI Initiative. The executive order elaborated that the Federal Government play a significant role in facilitating research and development into AI. In addition, it promoted training along with protecting national interest, the changing workforce, and security. The executive order

[11] Adapted from Gartner Methodologies, Hype Cycle, https://www.gartner.com/en/research/methodologies/gartner-hype-cycle.

highlights “American leadership in artificial intelligence” with the intent of enhancing collaboration with allies and foreign partners (Allen & Chan, 2017). The American initiative towards AI involves principles which are: training workers; protecting American values by incorporating privacy and civil liberties and fostering confidence and public trust in AI; driving technological breakthrough, protecting US technological advantage in AI, and supporting and promoting innovation in an international environment; followed by driving development for technical standards in the education system (Horowitz et al., 2018). This order calls on the NSTC (National Science and Technology Council) for selecting a committee on artificial intelligence to coordinate the initiative of American AI. The agencies and departments that are deploying and developing AI are requested to adhere to those objectives.

Regulating AI, providing grants, and guiding AI are all encouraged to adhere to the six objectives. These objectives cover the investment in research and development of AI along with accessing federal data, computing resources, models, reducing barriers to such technologies, minimizing vulnerability to malicious attacks, ensuring technical standards, implementing an action plan for national security interest, and protecting US economy, along with training AI researchers (Berendt et al., 2020). Research and development in AI are priorities for the US that has enjoyed broad support from governmental officials (Furey & Martin, 2019). The committee and NSTC also released an updated version of a strategic plan that included eight strategies. NITRD (The Networking and Information Technology Research and Development) in 2019 released a supplement to the initiative in the form of an FY2020 budget highlighted to be one billion dollars per year (Parker, 2018). The memo from the president's executive office regarding this named it the second-highest priority in the fiscal year 2020. This was prioritized because of security purposes, as AI development was considered central to national security and a major disruptive technology in the world market.

Artificial intelligence was also featured in the 2018 National Security Strategy to help the country lead in innovation and technology, including the education system, followed by statecraft, surveillance, and weaponization (Hoadley & Lucas, 2018). Interestingly, in 2017, the Department of Homeland Security put out a report by the name of “AI Risk to Critical Infrastructure” that analyzed the narratives relating to AI for understanding the benefits and threats of the adoption of artificial intelligence. Some of the policies surrounding the success of AI involve building post-basic education and training, frameworks for digital skills, computational thinking, higher education with AI, along vocational and technical training. The skills are intended to strengthen education surrounding the development of AI (Williamson & Eynon, 2020). AI under the Obama Administration also included measures regarding AI policy in the United States. The former president launched in 2016 along with the White House Office of Science and Technology Policy (OSTP), a series of workshops along with a "Subcommittee on Machine Learning and Artificial Intelligence" for monitoring technological advances which help in coordinating the activity of the AI use on the federal level. These activities led to the formation of three reports that influenced thinking around AI across the globe. These reports were "Preparing for the Future of Artificial Intelligence," "The National Artificial Intelligence Research and Development Strategic Plan," and "Artificial Intelligence, Automation, and the Economy."

As for focusing on local and state policy regarding AI, various bills are being introduced at local and state levels. An example is the California State passing a resolution in 2018 in support of the Asilomar AI Principles, containing twenty-three guidelines for beneficial and safe use and development of AI (*Assembly Concurrent Resolution No. 215. Asilomar AI Principles*, 2018). As national policies are made, we can see that the support of local governments is also essential to disseminate the issues surround the use of AI properly.

5.5.2 Governance in Japan

The amount of information regarding AI and policy in Japan is certainly much greater than this researcher was able to access. Due to this researcher's limited ability in Japanese, only reports that were published in English were analyzed and this section is only intended as a brief overview. Further examination of reports in the native Japanese would give the researcher a deeper understanding of national policy on the topic of AI.

In 2016, Japanese Prime Minister Shinzō Abe required the Japanese government to create the "Artificial Intelligence Technology Strategy Council," whose aim was to increase the number of specialists in AI and support them with funding (Garcia, 2019). Shortly after, as Garcia reports, Japan published its "Draft AI R&D Guidelines." This recognizes "the enormous benefits that AI will bring for people as well as for the society and the economy, making important contributions to solve different difficulties that people, local communities, countries, and the world are faced with. They also identify certain risks such as lack of transparency and loss of control" (Garcia, 2019, p. 29).

Additionally, in 2019, Japan was the host of the G20 Artificial Intelligence (AI) Principles forum. The forum and working-group aim were to produce a human-centered approach to AI, furthering public trust and confidence in AI technologies to maximize their potential (Vincent-Lancrin & van der Vlies, 2020). Japan has put considerable effort into transforming its education system by reforming its national curriculum standards of education.

However, as seen in Figure 32, the factors involved in applying AI and the number of actors working on its advancement and implementation require a top-down view better to grasp the complex reality (Dirksen & Takahashi, 2020).

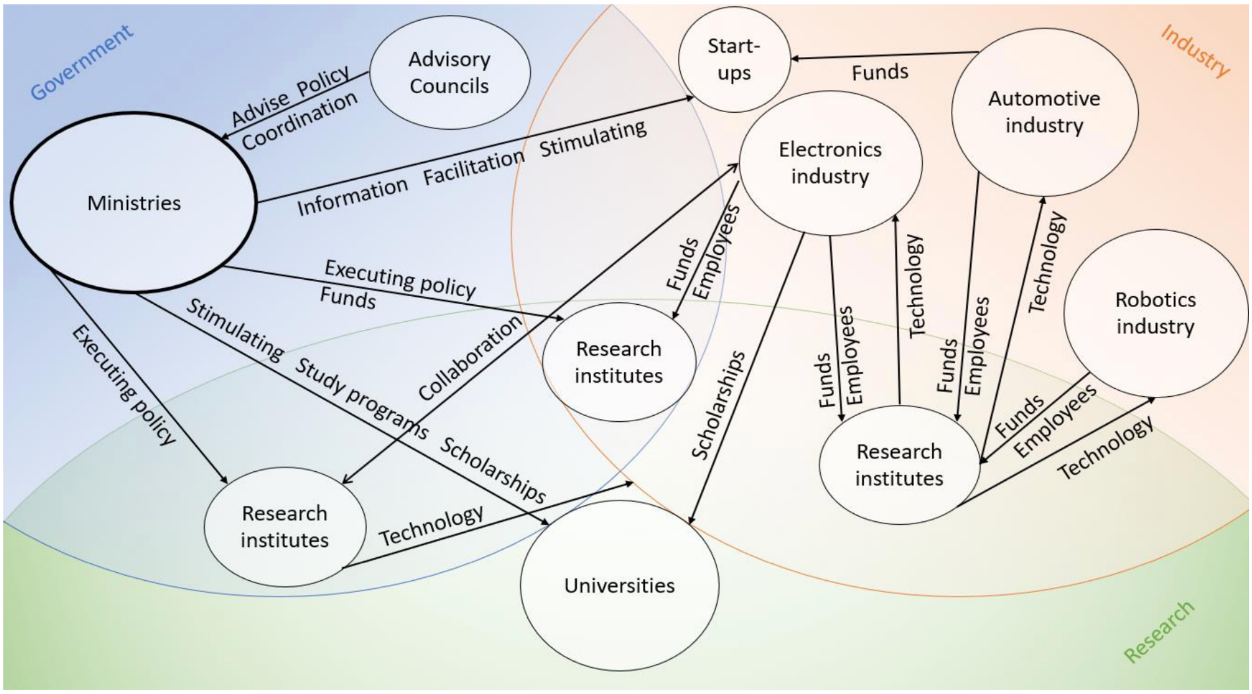


Figure 32[12]*. Artificial Intelligence development in Japan as described by Dickerson.*

While Japan has traditionally been slower to develop and participate in AI frameworks due to cultural factors (Fujii & Managi, 2018), there are still initiatives to develop and incorporate these technologies into society. An umbrella initiative that serves as a guideline for developing new systems is termed "Society 5.0". According to Hayashi et al., (2017), with Society 5.0, Japan seeks to,

> […] create new values by collaborating and cooperating with several different systems and plans standardization of data formats, models, system architecture, and others and the development of necessary human resources. In addition, it is expected that enhancements of intellectual properties development, international standardization, IoT system construction technologies, big data analysis technologies, artificial intelligence technologies, and others encourage Japan's competitiveness in a "super smart society" (p. 264).

Higher education in Japan emphasizes the development of artificial intelligence in the education system to accelerate the competitiveness of academia and industry. One of the main

[12] Dickerson, 2020. Artificial Intelligence in Japan. The complex interaction between Government, Industry, and Research. Reprinted with permission.

motivating factors of furthering AI education is developing more qualified professionals and enhancing the domain of artificial intelligence (Zeng et al., 2018). According to a Japan Times article written by Takamitsu Sawa, the vice director of the International Institute for Advanced Studies, The Ministry of Education, Culture, Sports, Science, and Technology (MEXT) is working with industry and academic professionals to reform university education. Specific measures include revising university entrance exams to reflect better hard skills students will need in their academic and professional lives. MEXT produced a report that emphasized the need to support human resources capable of AI and data science and improve basic liberal arts education (Sawa, 2019). Much of the resource investments are being applied to integrating artificial intelligence-based education to improve Japan's AI research capabilities. This can also broaden Japan's reach into overseas research projects, improving Japan's standing globally as a center of AI research, development, and deployment (Ishii et al., 2020). The Japanese, traditionally seen as leaders in robotics and automation, position itself to leverage artificial intelligence as one of the most significant technological developments in this century.

5.6 Conclusion

This chapter has highlighted some of the current developments in artificial intelligence and its application in education. In addition, the chapter addressed specific policies and applications of AI being used in the United States and Japan. The pedagogical implications of using AI in the classroom are vast, with opportunities for better learning and a more efficient learning model. It can provide numerous opportunities for learners to understand better the concepts they might struggle to master in a traditional classroom setting due to many factors. AI-based assessment and classroom management systems help saving educator's time spent on non-instruction activities. AI can transform traditional learning methods into new and innovative paradigms that are more granular, accessible, and equitable.

The second part of this paper covered some of the initiatives and policies being developed around artificial intelligence in the United States and Japan. The researcher was able to identify research and policies being formulated surrounding artificial intelligence in general, but with very little policy research being implemented in education specifically. As these economic leaders push the way forward, we can gain insight into the net gains and potential problem areas and pitfalls when using AI in society. The author found a surprising lack of literature regarding policies and guidelines on national or local levels regarding the application of artificial intelligence in the education sector. Further research needs to be conducted and disseminated in this area as the ramifications of using AI assistance in learning are potentially transformative.

The next chapter of this dissertation gives some insight into how artificial intelligence is perceived by in-service educators. Specifically, what do they think about AI-based assistance in the classroom. As research and development into these digital aids progresses at a breathtaking pace, the opinions of educators on how these technologies are impacting their field should be considered. At the end of the day, it will be educators' technological knowledge and use of these digital aids which will make them most impactful to learners.

Chapter 6 Artificial intelligence in education: Educators' perspective on an emerging technology

Keywords: *AI in education, AIED, educator training, Computers in education, mixed methods research*

6.1 Background

The penultimate chapter of this dissertation turns its focus to in-service educators to gain some insight into their outlook and opinions of artificial intelligence and its use in the classroom. Increasingly, educators and education policymakers have started to recognize the potential for artificial intelligence to be a disruptive technology in the classroom. However, few studies have investigated educators' knowledge and attitudes surrounding the use of these technologies. This chapter uses a mixed-methods approach to gain qualitative (interviews, open response questions) and quantitative (Likert survey; sentiment analysis) data to gain an understanding of educators' perspective on the concept of artificial intelligence, its use in the classroom, and their outlook on how the technology may impact learning outcomes and teaching practice.

6.2 Introduction

> *The teacher of 2010 will rarely spend a day lecturing but will be primarily a facilitator and coach. ... the teacher will coach students through video lectures, educational television programs, and artificial intelligence-based programs. Only occasionally will teachers instruct classes themselves. Instead, they will be freed up to deliver the personalized instruction critical to educational achievement. (Cornish, 2004, pp. 9-10).*

While Cornish's prediction quoted above from 2004 has not fully come into fruition, some of the basic tenets of learner autonomy, agency, and educator empowerment have seen much progress aided by advances in edtech. The rapid advancements in machine learning (ML) and other AI based educational technologies have begun to gain mindshare among educators and students. The field of artificial intelligence in education (AIED) itself has seen expanded interest as seen by the launching of dedicated refereed journals (Computers and Education: Artificial Intelligence in 2021; International Journal of Artificial Intelligence in Education) and the establishment of special interest groups (SIGs) in CALL societies such as the Intelligent CALL (ICALL) SIG in the European Association for Computer Assisted Language Learning (EUROCALL) association. Inside the classroom, students are increasingly using these technologies (machine translation, voice-to-text) to complete assignments, and educators are also becoming exposed to digital tools that might use some forms of AI (plagiarism detectors, grammar/style checkers).

This study seeks to gain insight into the perceptions of educators regarding these technologies and their potential impact on students and the teaching profession itself. The researcher, having developed and tested a unique AI-based writing assistant in empirical studies (Gayed, Carlon, & Cross, 2022; Gayed et al., 2022), would like to understand in more detail potential issues and affordances these tools present from the perspective of a population that will be impacted by these technologies.

A survey based largely on newly created constructs that were created and distributed via online forms such as Facebook, LinkedIn, and other social media platforms. Participants from twenty-nine countries responded to the survey delivered via the Sogolytics online survey platform. In addition to the survey data, qualitative data was gathered via follow-up interviews to give the researcher deeper insight into the opinions of the respondents. Some of the results of this research show that educators have a largely positive view of student use of AI-based technologies in the classroom. However, institutional support for training and implementing edtech is seen to be lacking to some extent.

6.3 Literature review

The researcher has found a paucity of investigative studies in the literature that examine the views of educators on the topic of emergent artificial intelligence technologies. Chounta et al. (2022) examine the opinions of Estonian K-12 educators via survey instruments on their use and perception of artificial intelligence in education. The researchers found that knowledge and use of AI tools is increasing in Estonian classrooms. However, teachers' professional development is identified as a key driver that would enable higher efficacy of AI in the classroom. While systematic reviews such as Chen et al. (2020); Tang et al., (2021); and Zhai et al. (2021) identify trends in artificial intelligence in education, the systematic reviews do not identify any contributions to the literature that focuses on educator's views. Instead, they focus on topics such as course assessment, learning analytics, and knowledge tracing among others. Similarly, Roll & Wylie (2016) conducted a systematic review of 47 papers on AI in education but largely don't find any studies that measure educators' views on the technology. While Roll & Wylie do attempt to identify quantitative or qualitative studies that examine classroom practices, they only briefly mention the role of the teacher/educator as a dimension to consider.

In this researcher's search of the literature, a number of studies did consider the role of the teacher, but only when the perspective was "teaching AI" not "teaching with AI" (Lee & Lee, 2020; Wollowski et al., 2016), a decidedly different realm that this researcher is not touching upon. Other research areas include surveying the general public about their views of AI being used in education and the surrounding ethical issues of its use (Latham & Goltz, 2019) or narrowly focused studies on specific applications (e.g., machine translation) (Briggs, 2018).

This study will elucidate on an area that has been largely ignored: how educators feel about the use of AI technology in the classroom. Specifically, this researcher is interested in AI-based agents that impact EFL students' writing proficiency, therefore the survey questions used in this study will be focused on this rather narrow aspect of artificial intelligence use in education.

6.4 Methods

A mixed methods approach was used to gain quantitative and qualitative data for this study. The data gained from this study was also approved via the researcher's host institution's human subject research ethics application (see Appendix C (Ethics approval for human-subjects research)). A questionnaire using newly developed constructs (Planning Technology-

Supported Instruction; Technology and Assessment) and specific questions concerning AI agents being used to assist student writing was formulated. The survey was delivered via the Sogolytics online platform (https://www.sogolytics.com/online-survey-tool/) and was distributed via online social networking sites such as LinkedIn, Facebook, and Twitter. The researcher accepted responses over a six-month period between April 2021 and December 2021. The Likert data and open-ended questions received from participants are analyzed with Mizumoto's (2015) web-based R interface (see also Mizumoto & Plonsky (2016)) accessible via https://langtest.jp. Thirty questions were contained in the survey and are detailed in Appendix G (Survey to Educators).

The follow-up interviews were conducted online via the Zoom video conference software. In total, ten participants indicated agreement for a follow up interview. Out of those ten the researcher was able to successfully arrange an interview with six candidates. After the interview was recorded and the audio tracks separated from the video file, Otter.ai (https://otter.ai/) was used to transcribe the interview. The transcription was then checked manually by the researcher to fix any major errors by listening to the interview again while going through the machine-produced transcription.

When identifying themes in the participants' written responses and the post-survey interviews, hand-coding of the responses was avoided due to volume of transcripts. Instead, content analysis was conducted via 1) frequency analysis of top keywords and 2) Latent Dirichlet Allocation (LDA). LDA is an unsupervised probabilistic model that can identify latent themes in unlabeled data (Blei et al., 2003). Other works using LDA topic modeling are numerous in fields inside and outside of the social sciences (see Gurcan et al., 2021; Momtazi & Naumann, 2013; Xue et al., 2020). This technique allows the researcher to analyze large amounts of data and distill topics and themes from that data that would otherwise require significant human capital to label and code. The LDA analysis in this chapter is conducted with David Mimno's (mimno, 2022) implementation accessed via https://github.com/mimno/jsLDA. This implementation processes data that was formatted by the researcher into Mallet's input schema and automates the pre-processing steps of data preparation (symbol removal, stop-word removal). After pre-processing the input data, the application allows users to indicate the number of topics to model and the number of iterations to run the LDA analysis. For this study, *k* (number of topics to model) was set to 4 and model iterations was set to 1,500. Iterations above 1,500 showed modest improvement to the topics so the analysis was stopped at 1,500.

6.4.1 Participants

After the response period was closed in December 2021, n = 134 responses were collected and exported into the statistical analysis package. Twenty-nine countries are represented in the study which gave the researcher the confidence that a wide perspective from different educators would be collected on the topic. The countries where participants are currently working are described in Figure 33 and the influence of this researcher's working country (Japan), home country (USA), and professional network on the makeup of the respondents is clearly evident in the figure. This is a confounding factor that the researcher's personal network effect influenced the type of respondent who answered this survey. The demographics of the participants are described in Table 20 in more detail. The demographic makeup of the respondents shows us that they are predominately male, teaching at university with many years of experience (seven years +) and have a Master's degree or greater.

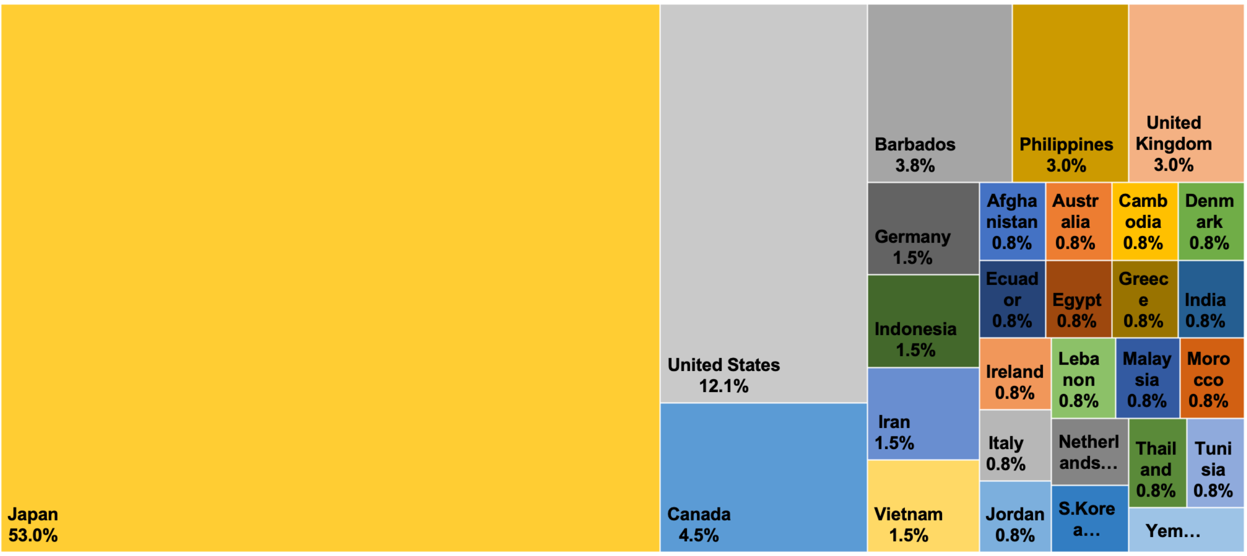


Figure 33. Self-reported working locations of respondents. n = 134.

Table 20. Demographic factors of participants.

Demographic Factors	Number	Percentage	Demographic Factors	Number	Percentage
Gender			**Teaching experience**		
Male	93	69	Less than one year	3	2
Female	36	27	1-3 years	4	3
Other	4	3	4-6 years	11	8
			7-10 years	17	13
Profession			11-15 years	27	20
Teaching	129	96.3	16-20 years	24	18
Non-teaching	5	4	21-25 years	19	14
			More than 26 years	27	20
Working institution type					
University / College	105	78	**Grade level of students**		
High School	15	11	Adults (above university age)	38	28
Elementary School	5	4	University / College students	106	79
Private Language School	11	8	High-school students	20	15
Other	10	8	Junior high-school students	14	10
			Elementary school students	16	12
Subjects taught			Kindergarten students	10	8
Business and economics	14	10	Other	4	3
Arts and humanities	74	55			
Engineering and technology	10	7	**Highest level of education**		
Life / Physical sciences	3	2	Doctorate	42	31
Social sciences	31	23	Master’s degree	72	54
Creative Art and Design	3	2	Bachelor’s degree	16	12
Law	1	<1	Associate degree	1	<1
Travel and Hospitality	6	4	Technical or occupational certificate	1	<1
Other	48	36			

6.5 Results

6.5.1 Existing edtech knowledge and practice

Participants were asked to define the term artificial intelligence as one of the first questions they received on the survey. One hundred eleven responses were received written in English, two in German, and six responses were written in Japanese. Representative typical responses of all the responses are listed below for reference:

1. "A computer program that performs limited reasoning, pattern recognition, and/or predictive functions that to some extent can learn from a complex data set and become better at its function."
2. "Technological entities with at least a modicum of human-like skills."
3. Original German ---"KI ist ein selbstlernendes System, welches aber immer auf den kontinuierlichen Input von Menschen angewiesen ist und basierend darauf weiter lernt und sich weiter verbessert. KI ist wissensmäßig dem menschlichen Gehirn überlegen, aber hinsichtlich emotionaler Intelligenz meines Erachtens unterlegen."

   Translated --- "AI is a self-learning system, but it always relies on continuous input from humans and continues to learn and improve based on that. AI is superior to the human brain in terms of knowledge, but inferior in terms of emotional intelligence in my opinion."
4. Original Japanese --- "膨大なデータを総合的にみて的確に予測・推測し展開を図るために役立てるシステム"

   Translated --- "A system that comprehensively looks at a huge amount of data and makes accurate predictions and inferences for deployment."

The very definition of artificial intelligence is one that is commonly debated (Monett & Lewis, 2018; P. Wang, 2019), but we can see the responses from this survey lean toward defining AI as a system that learns from data. More specifically, after using LDA (see section 6.4 for methodology) analysis on all the responses (N = 111), the themes that emerged are described in Table 21.

Table 21. LDA analysis of written responses.

Topic	Keywords	Interpretation
1	intelligence, human, machine, tasks, computer	Intelligent agents or software that assist in tasks
2	data, computer, learn, algorithms, programmed	Computer generated algorithms that analyze data and produce a response
3	technology, human like, software, use, autonomous	Algorithms that improve autonomously
4	students, user support, knowledge, tools	The use of computer intelligence to improve tasks or make tasks more effective.

Before asking about artificial intelligence being used in the classroom, the researcher wanted to gain some insight into the participants existing knowledge and use of technology in the classroom. To that effect, two constructs were devised to measure "Planning Technology-Supported Instruction" and "Technology and Assessment". Each construct contained three questions. To measure the constructs' validity exploratory factor analysis was conducted. For "Planning Technology-Supported Instruction" the factor loadings were, 0.79 for Q1, 0.65 for Q2 and 0.60 for Q3, in addition, Cronbach's coefficient alpha was also calculated at 0.70, showing the researcher that the three questions can reliably measure the construct. For "Technology and Assessment" the factor loadings were 0.51 for Q1, 0.84 for Q2 and 0.71 for Q3, Cronbach's coefficient alpha was calculated at 0.72, again demonstrating strong reliability. The visualization of the responses to both of those constructs are shown in Figure 34 and Figure 35. This indicates to the researcher that participants in this study are comfortable using technology in the classroom with 90%+ of the responses being along the "favorable agreement" bands of "strongly agree, moderately agree, and mildly agree". In light of this, the researcher believes the reponses from this survey to be heavily biased towards educators who already have a strong edtech background.

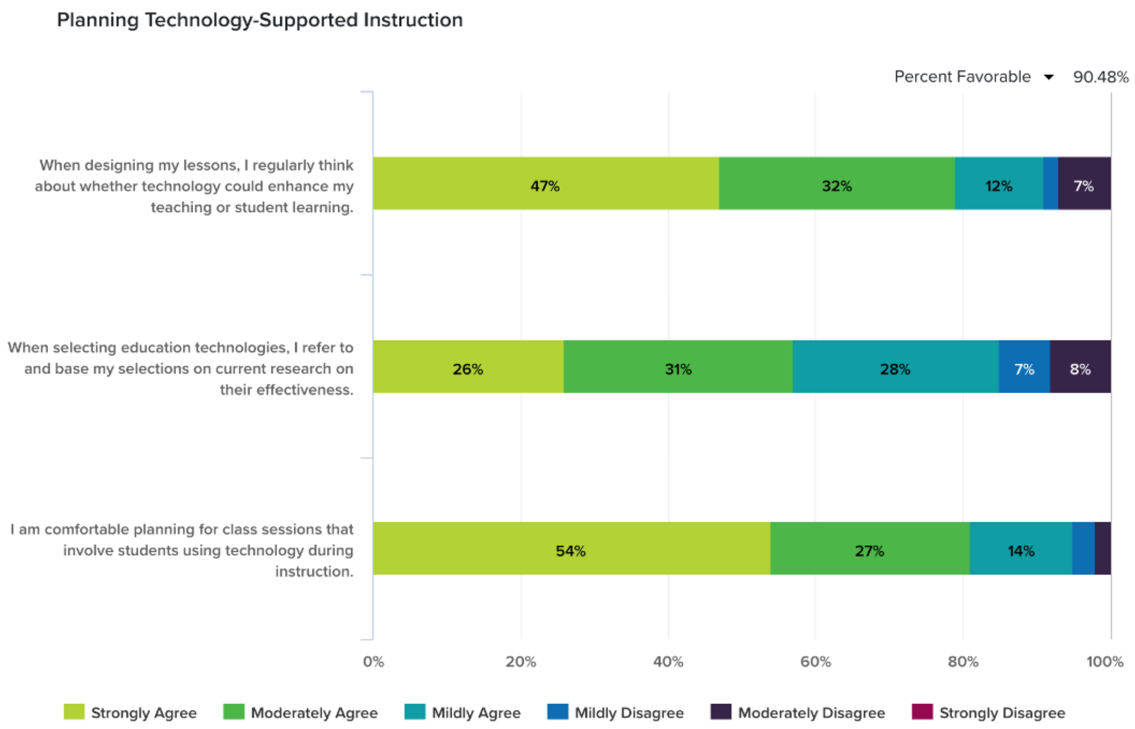


Figure 34. Respondents' Planning Technology-Supported Instruction.

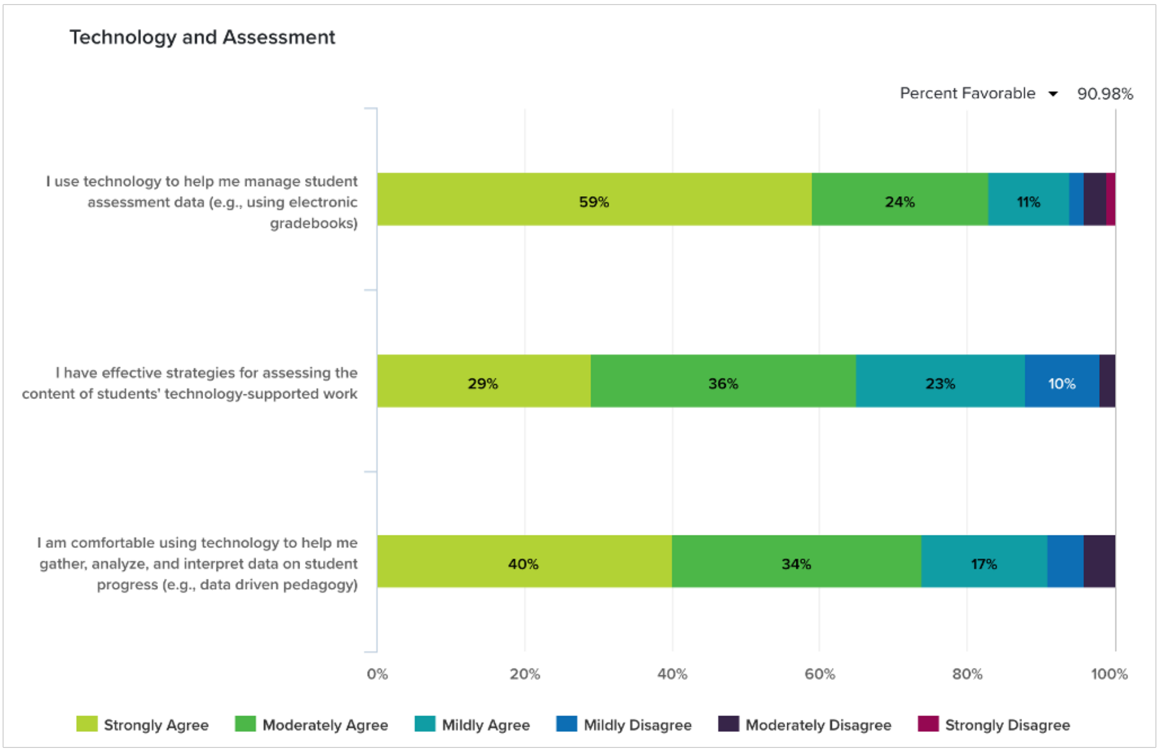


Figure 35. Respondents' Technology and Assessment.

When asked about the types of technologies used in their classrooms, the responses indicated a not surprising combination of computers, learning management system (LMS) and smartphones. Of note, only very few respondents (2%) indicated that they did not use any kind of technology in the classroom. Figure 36 summarizes the responses regarding the types of technology used in the classroom.

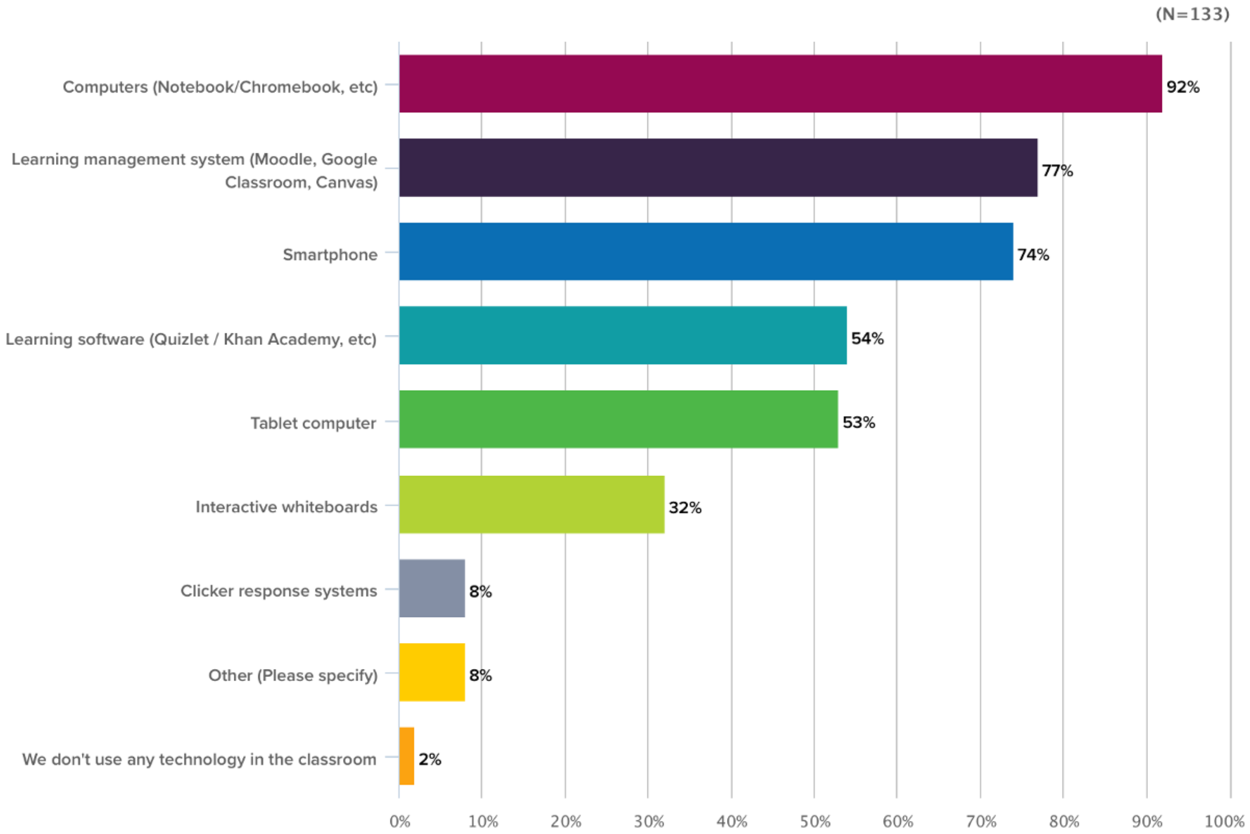


Figure 36. Types of technology used in the classroom.

6.5.2 AI in Education

Turning the focus to artificial intelligence technology being used in the classroom, this section details the respondent's current knowledge of AI tools. Responses to the questions: Q15 "I am familiar with current AI technologies being used by my students in my class", and Q16 "To your knowledge are you using any form of AI or AI-enabled technology as a teaching aid (e.g., Alexa, Siri, Grammarly, etc.)", Q22 "Do you think AI is influencing your approach to teaching" and Q24 "Do you think AI is impacting your students' learning outcomes?". Cronbach's coefficient alpha calculated for Q22 and Q24 show a value of 0.56, which is not entirely surprising as learning outcomes and approach to teaching are not necessarily aligned with each other. The results of those questions shown in Figure 37a-d do not show a strong indication that the participants in this study were familiar, were actively using, felt that AI was influencing their approach to teaching or that AI was influencing the learning outcomes of their students.

15. I am familiar with current AI technologies being used by my students in my class

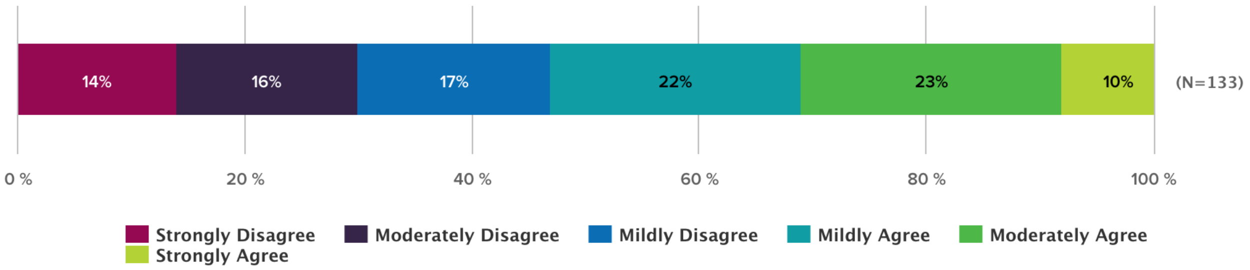


Figure 37. Familiarity with AI technology in the classroom.

16. To your knowledge, are you using any form of AI or AI-enabled technology as a teaching aid (e.g., Alexa, Siri, Grammarly, etc.)?

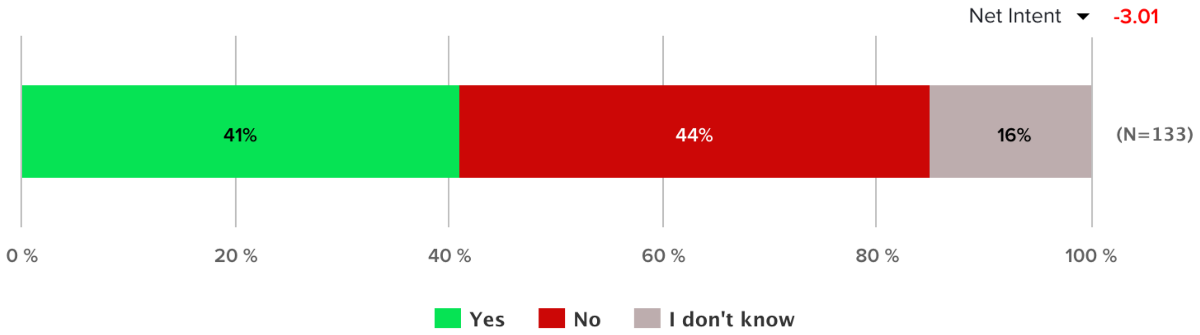


Figure 38. Usage and impact of AI tools in the classroom.

22. Do you think AI is influencing your approach to teaching?

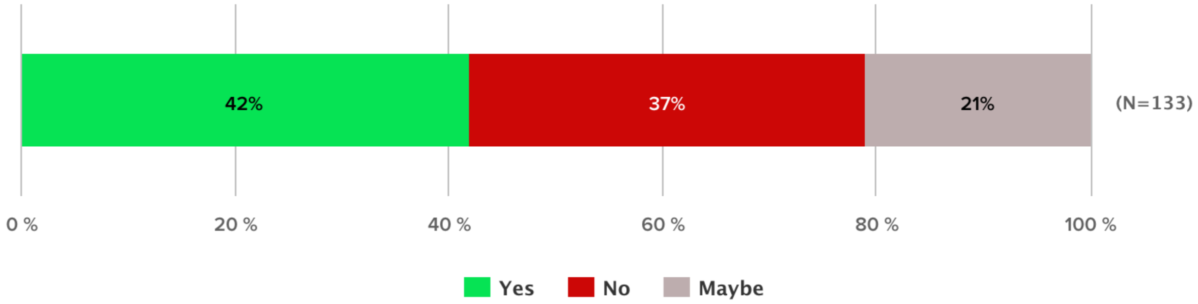


Figure 39. Perception of AI tools influencing teaching.

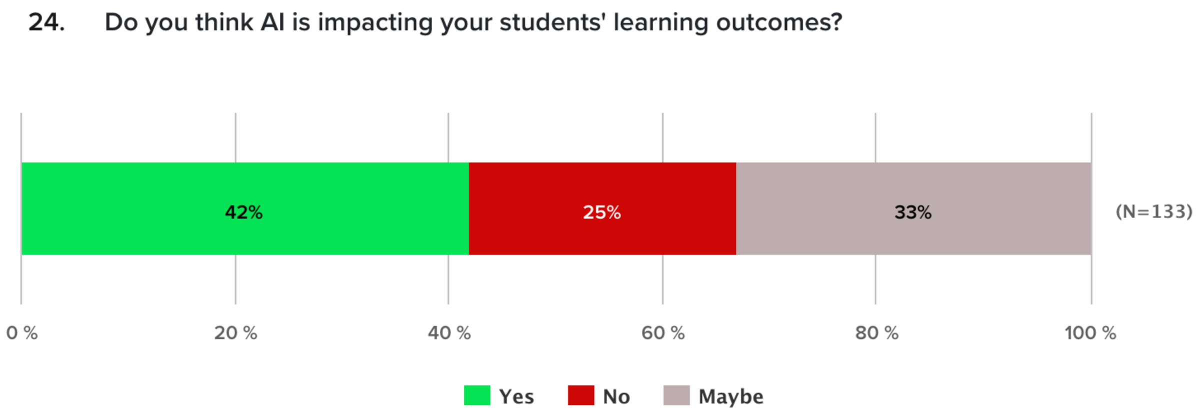


Figure 40. Perception of AI impacting learning outcomes.

Frequency text analysis (see Figure 41) of all the seventy-nine open responses given to the question, “If yes, how are you using the AI aid in your classroom?” showed the term “Grammarly” as the top word used being mentioned twenty-five times, demonstrating the recent popularity of the application and its use in the classroom. Other similar AI-based tools created by big technology companies such as Google were also mentioned. Some representative responses mentioning these tools include:

1. “I have used an intelligent writing assistant (Grammarly) for assistance with English writing and an intelligent personal assistant (Alexa) for English speaking/listening practice my classes.”

2. “Google Translate, Google Voice Recognition, I use Alexa to run my listening stations as a teacher, but the students do not use it.”

3. “Trying to implement dialogue systems for language learning as out-of-class meaningful practice activities. Currently working with an in-development game for language learning involving dialogue with non-player characters. I also strongly recommend my students to use Grammarly for instance.”

4. “We have google home [sic] ready to use for chat bots in activities, and students sometimes use translation software such as deepl or google classroom for assisted writing.”

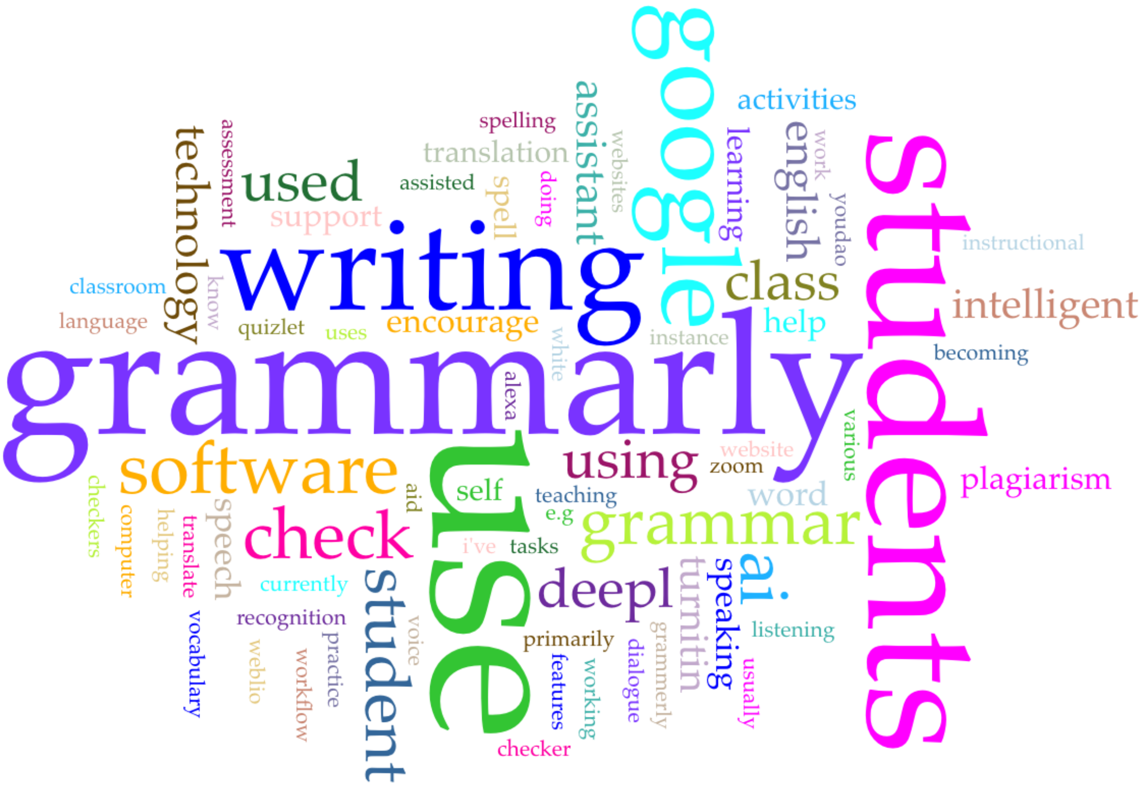


Figure 41. Word cloud of AI tools being used.

For the question, “Please explain how AI is influencing your approach to teaching,” LDA analysis on all the responses (N = 79) was conducted, revealing the themes described in Table 22. Some representative responses demonstrating those themes include:

1. “It helps me consider innovative ways to help my students improve their English.”

2. “I'm aware that students may be using AI in submitting homework and other assignments. I have more concerns about machine translation than other forms of AI.”

3. “I have to be aware of the pros and cons. We want students to think for themselves yet give them a little help. Too much help might be using google translate to translate three pages. That would be like using steroids to hit a home run.”

4. "The data we get, for example, from Quizizz enables us to identify which questions/areas students find easy/difficult, and what might need to be reviewed in class. It also indicates learning behavior (e.g. when, how often, and how long students engage with apps). This can then be used to give feedback to students, identify learning needs, discuss strategic approaches to learning, etc."

5. “Indem auch ich als Lehrkraft auf Vorschläge der KI eingehe, ändert sich mein Lehrverhalten wahrscheinlich. Müßte ich testen, ob ich ohne die Vorschläge von KI genau die gleichen didaktischen Entscheidungen treffen würde.”

Translated: By also responding to AI suggestions as a teacher, my teaching behavior is likely to change. I would have to test whether I would make exactly the same didactic decisions without AI's suggestions.

6. I am reliant on algorithm-based feedback which assesses students in a number of linguistic areas. This offers far more comprehensive feedback than could be realistically expected from an instructor.

These responses indicate student-focused considerations as the respondents are indicating how AI is influencing their teaching, with an emphasis on feedback systems as a way that AI may impact their teaching. Aside from the first theme in Table 22 about AI writing aids, which could have been influenced by the preceding questions, the other themes dealt with how to best use technology in an educational setting.

Table 22. LDA analysis on how AI is influencing teaching.

Topic	Keywords	Interpretation
1	Grammarly, Google Translate, writing, aid, writing, assistants	Using AI writing aids like Grammarly or Google Translate to assist in writing
2	data, learning, outcomes, software, MOOC	AI-based learning platforms and tools that collect data on student performance and learning behavior
3	improve, learning, voice, face, interactive, fun	Creating more interactive, engaging learning experiences in and out of classroom
4	fairness, pros, teacher, training, concerns	Adapt strategic and cautious approach to implementation, benefits and limitations of the technology

6.5.3 AI writing assistance

Narrowing the focus further, the researcher gained insight into participants' views when it came to AI assistance for writing tasks. This was especially pertinent to the researcher considering the researcher's interest in developing AI-based software applications that help students in an EFL context. The questions of Q18 "A student using AI assistance in school is unfair to other students who don't use such assistance" and Q21 "AI augmented writing (writing where the student is using AI assistance) is a form of plagiarism as the writer is using external assistance to complete an assignment" both show a net rejection of those notions with 60% to 70% of respondents indicating "mildly to strongly" disagreement with the statement (see Figure 42;

Figure 43). Internal reliability for Q18 and Q21 measured via Cronbach's alpha resulted in a value of 0.56 indicating only a weak relationship between perceived fairness and the notion of AI assisted writing as a form of plagiarism. This is in line with the researcher's expectations.

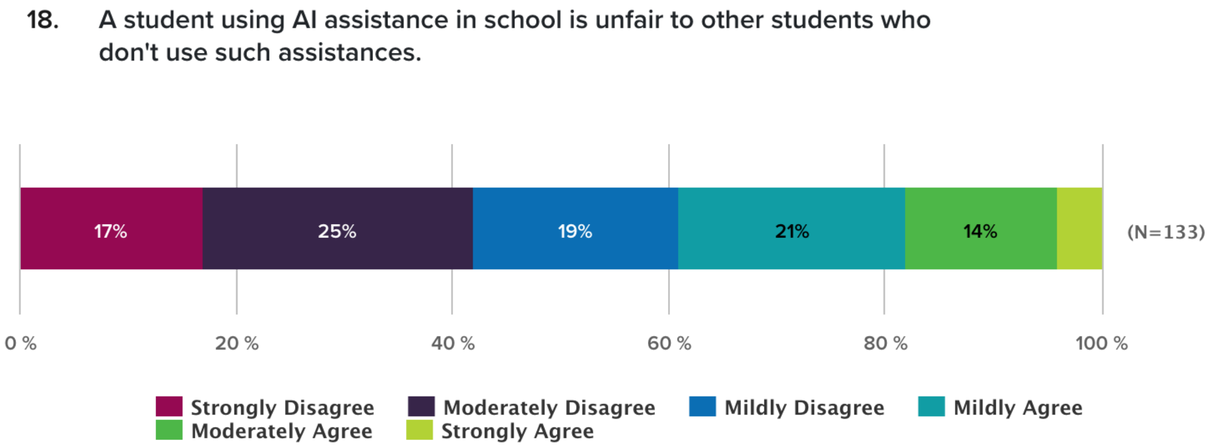


Figure 42. Respondents' views on fairness of AI assisted writing.

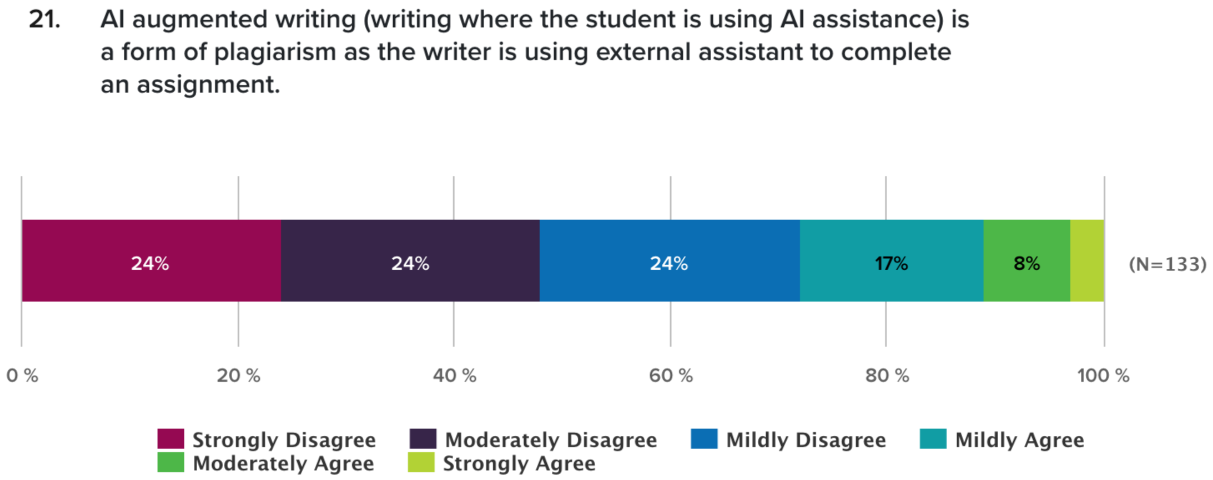


Figure 43. Respondents' views on AI assistance as form of plagiarism.

Responses that commonly appeared to Q18 and Q21 include:

1. I think it is smart and 21st century. I teach ALL my students how to use the affordances that will lead them to be effective professionals in the 21st century.

2. I disagree with the statement. It's just another tool that students can use to help in their learning and achieve mastery.

3. It is fair. It is the way of the future and students ought to be familiar with HOW to use it effectively, so AI wouldn't be viewed as a lazy way of encouraging predictive learning.

4. It is fair as long as all students are provided with the same opportunity.

5. I am usually teaching technology subjects so it would be hypocritical if I prevent my students from using tech to make their lives easier

Here the responses demonstrate an overall rejection of both notions that AI-assisted writing is either a form of plagiarism or an unfair advantage to students who use it. Emphasis is placed on understanding the capabilities of the technology and how to implement those capabilities into instruction and student learning.

6.5.4 Post survey interview

As mentioned earlier, some of the participants agreed to follow-up interviews. Six interviews averaging one-hour in length were successfully completed over the online video conferencing application Zoom. Some descriptive information about the interviewees is shown in Table 23. This information is included to give a general sense of the background of the interviewees without revealing any personal information about them that could potentially identify them to the general public.

Table 23. Interviewees' demographic factors

#	Education level	Career level	Subject	Workplace
1	PhD	Departmental Director	Humanities	University
2	PhD candidate	Associate Professor	EFL	University
3	M.Ed.	Lecturer	Business Admin	University
4	MA	Lecturer	EFL	University
5	PhD	Associate Professor	EFL	University
6	MS	Head Teacher	EFL	Private Language School

After conducting the semi-structured interviews, the sessions were transcribed for analysis using a combination of speech to text software and manual correction by the researcher. The questions asked during the interview included:

1. Do you feel artificial intelligence will become widely used in education? Or do you feel AI will be a niche technology? Please explain your reasons.

2. What would you consider as a possible successful application of AI in education?

3. What are some of the critical ethical issues surrounding AI being used in education?

4. From a professional perspective, do you see AI as a potential threat to your career? Why or why not?

5. Have you witnessed any improvements in the learning outcomes in your students? Or improvements in your professional workload?

The researcher analyzed all the responses from the interviewees with frequency analysis and the top keywords from all the participants are as follows:

Table 24. Top keywords from interviewees.

Interviewee	Top keywords spoken
1	machine translation, ai, students, plagiarism, technology, Japan, correcting, teacher, learning, working
2	students, ai, terms, technologies, guess, responses, instructor, conversations, teacher, formal education
3	ai, students, technology, education, application, learning, Quizlet, YouTube, learn, ethical issue
4	students, Japanese, technologies, learning writing, kinds, Duolingo, word, reading, ai, translation, question, language, spellcheck, ethical issues
5	ai, data, learners, translation, students, guess, question, education, technologies, elementary schools, translate
6	ai, students, education, writing, plagiarism, guess, check, google translate, artificial intelligence

Word frequency analysis shows some common themes among the respondents, including machine translation, ethical issues, and plagiarism. In addition, LDA analysis was conducted on all the responses to each question, the topics identified from the analysis and the researcher's interpretation of the LDA analysis are shown in Table 25.

Table 25. LDA analysis of post-survey interviews.

Topic	Keywords	Interpretation
1	teachers, language, learn, technology, English, online, working	Using technology in CALL and the new dynamics of remote learning.
2	terms, data, learning, research, language, speaking, listening, assistants	Language research will progress with the development of these technologies.

3	students, correct, word translation, Japanese useful writing, reading	Impact on students especially considering translation and writing.
4	students, class, education, Google, threat, translate, future, learners	Negative impacts on students from private companies.

To give more insight into the responses received for each question, the responses that closely matched the topics identified in the LDA were chosen as representative samples:

[Note: some responses edited for clarity]

Q1 [TOPIC 1,2]:

> "I do think it will be used more frequently, or a greater degree than this now, but I don't see it being mainstream... [but] I only see it being, like, most applicable with large class sizes, like 100 plus… for example, like a MOOC. I think, when you're dealing with, you know, 100, possibly plus students, you need something to be able to give personalized feedback in a timely manner. And obviously, as an instructor alone, or even an instructor with TAs, you know, we can't do that. So, I see the value of AI in that context."

Q2 [TOPIC 3]:

> "I think translation apps, definitely seems like it will be useful for learners. I think my opinion with that is maybe it might be controversial to other people. But I think that probably when the translation apps and things that are good enough, I don't even think we should really waste time teaching students how to write necessarily as we would traditionally, unless they're interested in doing that."

Q3 [TOPIC 3,4]:

> "AI translation, we're getting pretty good now where students student may write it in their native language… for example in Chinese, Vietnamese, Korean or Japanese and then just translate it over? And then say, hey, this is my paper in English, which I think is an ethical issue. So, in a way that's self-plagiarism, right? Yeah. Plagiarism. It's not an ethical problem on the side of Google, but maybe yeah, for the students themselves. Then, for educational purposes the teacher would probably have to change their grading style."

Q4 [TOPIC 1]:

> "For myself, no, I don't see it as a threat. And before I've mentioned it as like a compliment... there's ways to assist you know, maybe an AI might assist in grading, you know, checking 1000s of paper or helping checking grammar or something. So, then the teacher can have more free time to plan or do the work on a curriculum side and other more urgent things. So, I think more of as a compliment, not as a threat. I could see some other colleagues who might say they call it a threat, but even they don't know what the threat is."

Q5 [TOPIC 1,2]:

> "But, yeah, so, yeah, and in terms of research, and I guess, even in terms of like, just anecdotal evidence, it does seem like it encourages them to speak, you know,

> that's just even in the classroom. So, I'm assuming, you know, outside the class that it is going to be that much more useful to them. And I find that if it is pretty fun, it will motivate them to engage more."

6.5.5 Overall outlook

When asked about their overall outlook to technology and AI in education, participants were asked to respond on a scale of 1 – pessimistic to 10 – optimistic. The net majority (84.21%) had an optimistic response as seen in Figure 44. Again, as indicated earlier, the existing edtech background of the respondents in this survey could have heavily biased the responses to this question.

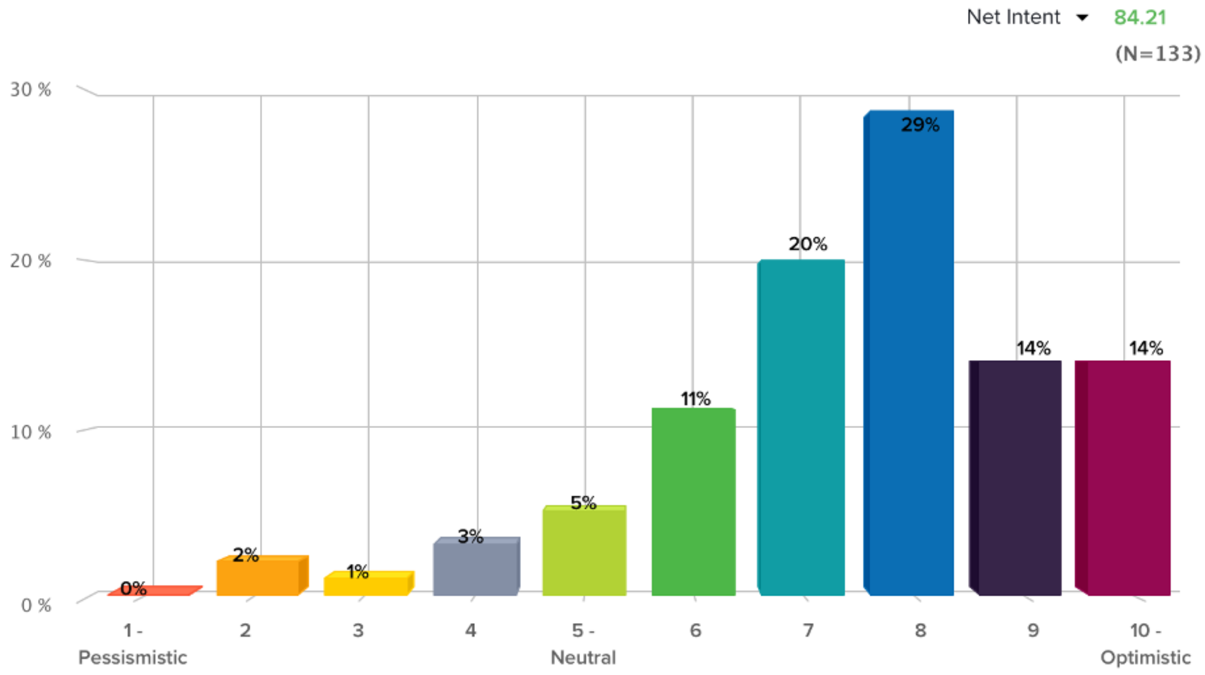


Figure 44. Respondents' overall outlook on AI and edtech.

6.6 Conclusion

This chapter examines how AI based technologies are perceived by educators using quantitative and qualitative analysis. The participants in this study (n = 134) are largely positive with their attitudes regarding AI in education. However, respondent bias is certainly a factor that influenced these results due to the sampling method used in this study. The researcher intends to build upon this research by conducting another mixed-methods study with respondents that have vary degrees of technological savvy. In addition, some responses in this survey gave contradictory opinions on concepts such as plagiarism and fairness. Future studies should define these concepts more concretely to have a more accurate understanding of how educators view artificial intelligence in education.

As Bax (2008) makes note, in the majority of educational settings a highly integrated relationship between education and technology has not yet been attained, with many teachers still seeing technology as an afterthought and a diversion from their typical classroom experience. Further investigation into educators’ professional development and institutional policy should be considered as these technologies advance and are adopted by students and educators.

Chapter 7 Conclusion and Future Work

> *"But the elastic heart of youth cannot be compressed into one constrained shape long at a time. Tom presently began to drift insensibly back into the concerns of his life again. What if he turned his back, now, and disappeared mysteriously? . . . [H]e would join the Indians… He would be a pirate! That was it! Now his future lay plain before him and glowing with unimaginable splendor."*

– Mark Twain[13], The Adventures of Tom Sawyer

7.1 First Experiment's Summary

The increasing use of English as a Lingua Franca (ELF) worldwide has brought attention to tools that can assist English as a Foreign Language (EFL) learners in their journey to fluency. Much research has shown that EFL learners often do not have sufficient latitude to output at a satisfactory level when writing in a second language. In addition, cognitive (working memory) resources are spent on low-level writing tasks (word production, translation) at the expense of time being allocated to higher-level writing tasks such as organization and revision. The researcher's laboratory developed an AI-based web application called "AI KAKU" to assist EFL learners in reducing the cognitive barriers they face when producing written text in English. While there has been much research and discussion on Automated Writing Evaluation (AWE) technologies or older technologies such as spell check and grammar check, few studies have attempted to use AI-based tools as learning aids instead of feedback agents.

The researcher considered the first experiment a success. Moving the application from the first conceptual design wireframes to a usable web-based application was the first accomplishment. Shortly after, a small group of adult EFL participants were recruited in a counter-balanced experiment to evaluate the potential impact of AI KAKU on student writing. The results of the experiment indicated that this is a potentially useful tool for English language learners who need more structured assistance than traditional word processors.

7.2 Second Experiment's Summary

As the pilot study showed promise, the researcher decided to continue to investigate the impact AI KAKU has on EFL students by conducting a larger study that investigates more aspects of

[13] The Adventures of Tom Sawyer, Chapter 8. Used under Fair Use copyright, https://www.loc.gov/item/20015592/

the writing process. The participant's cognitive functions and how AI KAKU affects participants with different English ability were key research questions in this second experiment.

When practitioners introduce new educational technologies into their classrooms, the potential for unintended outcomes from their use might arise. One such potential negative artifact is an increase in the achievement gaps between learners, where high performers tend to benefit more from newly introduced educational technologies than their peers. This phenomenon is commonly referred to as the Matthew effect. In the second experiment, we leverage the NLP-based assistant to introduce English language support to EFL learners while they are in the writing process. To understand the presence of the Matthew effect, learners were grouped based on their self-reported EIKEN scores. Their performance according to three writing factors as well as their perceived cognitive load while using the tool were measured to establish which groups benefit the most from using the tool. While we see gains among participants while they are using AI KAKU, analysis on how the tool was impacting participants of different levels was inconclusive. Despite the lack of clarity regarding AI KAKU's equity, these effects should be considered in both the development and application of educational technology.

7.3 Third Experiment's Summary

The final experiment investigates second language learners' writing output using an online next- word prediction writing tool after exposure to training and metacognitive prompts to improve their critical thinking. Engineering graduates' writing skills are often deemed lacking by industry standards; this can be even more challenging for EFL learners. This study employs a randomized control trial with university-level participants using AI KAKU and metacognitive prompting and nudging. EFL participants were given question prompts in the TOEFL iBT independent writing task style, and the outputs were assessed (machine and human) using several measures for writing quality.

All participants were shown short explanatory videos for TOEFL writing advice and metacognition training. The treatment group, exposed to the next-word prediction writing aid and metacognitive prompts, performed better than the control group even though both received the same training and writing opportunities. This study indicates there is value in providing writing support and metacognitive thinking practice to improve writing skills and, ultimately, writing output quality. This study was the most involved from a research design perspective and from a data analysis perspective. The researcher attempted a solution to efficiently rate over 360 writing samples in a timely and reliable manner while using the limited amount of human capital that was available to the researcher. In addition, a more multidimensional effort was made to improve participant writing by introducing the concepts of metacognitive writing strategies. The researcher feels this third experiment introduces novel solutions in both the design and novel outcomes that can be applied to larger bodies of work in applied linguistics, computational linguistics, and human, computer interface research.

7.4 Issues of Equity

As mentioned in Chapter 3 of this dissertation, the treatment tools in this study were analyzed for potential positive or negative effects on participants of different English abilities (e.g., high-level L2 learners vs low-level L2 learners). The resulting analysis in that study (see Figure 16 for reference) showed no clear indication that AI writing assistance was hindering or boosting high/low-level users. Yet, issues surrounding what type of learner benefits the most from these technologies should be investigated further as noted by Godwin-Jones (2022) in an overview of intelligent writing assistants. Across a range of technologies such as machine translation (e.g., Google Translate; DeepL), automated writing evaluation (e.g., Criterion) or AWE with synchronous feedback (e.g., Grammarly; Microsoft Word 365) and predictive text (e.g., Grammarly, Smart Compose) the researcher couldn't find consistent empirical evidence of those tools impacting a certain level of a user over another.

However, it is important to further investigate the potential these AI-based tools have on students and further research is needed into mitigating strategies to reduce inequity that arises from using artificial intelligence with our students. Some research such as a study by Chon et al. (2021) found that when machine translation was used as a mediating agent in the writing process it assisted lower-level participants at a greater rate than high-level users. Other studies by Dizon & Gayed (2021) also indicate that lower-level users can benefit from corrective feedback writing assistants such as Grammarly. Yet, as Godwin-Jones points out, the literature also shows contrasting views: lower-level users might not have the linguistic competence to understand the feedback they receive with these writing assistants (Koltovskaia, 2020). As such, researchers need to focus more on the pertinent question of what factors exacerbate inequality that might arise from the use of digital tools such as AI-based writing assistants.

7.5 In Summary

This dissertation attempted to measure the impact of a novel writing assistant on English language learners who are studying the target language in Japan. All three studies included quantitative analyses that demonstrated statistically significant improvement with participants who were using the tools development for this research. Further development and research into these AI-based tools is warranted considering the potential impact they have on student performance.

In addition, Chapter 5 and Chapter 6 give some context to the use of these tools in the classroom by examining educational policy level changes related to the development of AI. Chapter 6 in particular, sheds light on a rather underdeveloped area: the opinions and views of educators when it comes to educational technology that uses AI. The major positive outcomes of this dissertation are visualized in Figure 45.

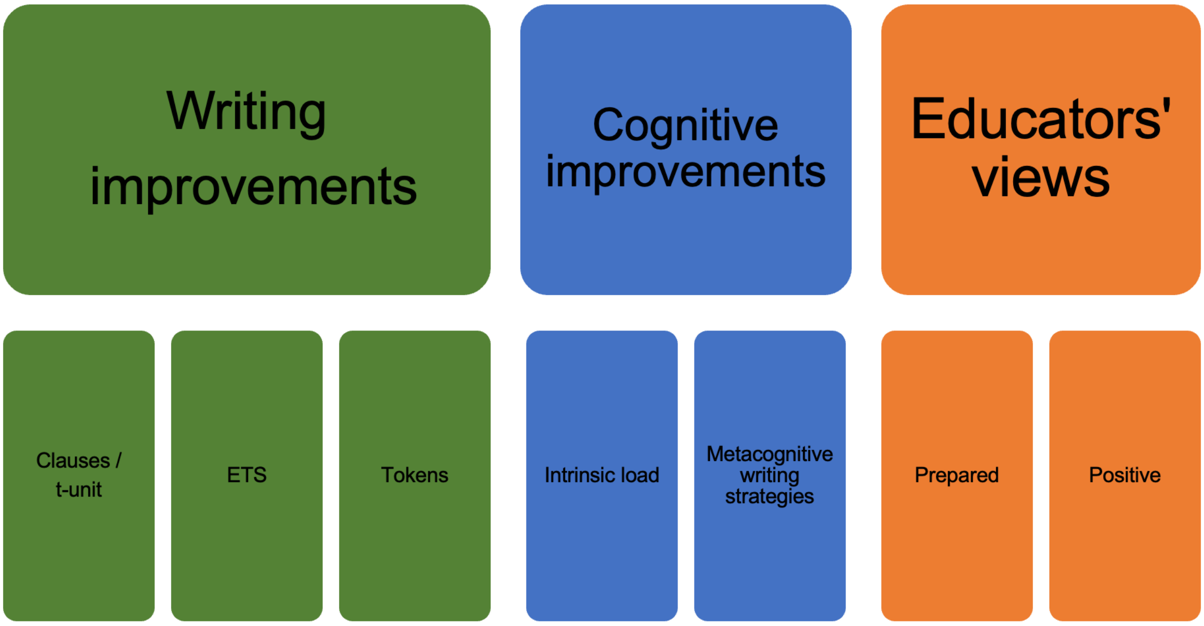


Figure 45. Major outcomes of this research.

7.5.1 Constraints and Future Work

As mentioned earlier in section 4.8 the non-probabilistic sampling method used in the three empirical studies in this research have limitations that should be highlighted despite the prevalence of using such methods (see Figure 27) in CALL research. The experiments in this research recruited English language learners studying in a university or language school context in Japan. Importantly, the researcher decided to conduct the experiments across multiple sites in order to satisfy one of the requirements of bolstering the validity of using non-probabilistic sampling. In addition, participants' demographics and reported English proficiency levels are described in section [2.4 Methods] for experiment one, [3.4.2 Experimental design] for the second experiment and [4.5.2 Participants] for the third and last experiment.

While the researcher followed methodological best practices to mitigate the disadvantages (e.g., response bias) of convenience sampling, the conclusions that can be drawn from the experimental studies in this research are limited to the demographic described in this research, and the results should be interpreted while keeping these limitations in mind.

Despite the fact that the majority of CALL research is conducted in classrooms with students, further research is justified using expanded methods. To gain clearer insight into the impact treatment measures have on representative populations, wider scale randomized sampling techniques can be used in order to gain more generalizable results. Furthermore, research designs incorporating diverse sampling groups, and long-term follow-up can build upon previous research and give researchers and policy-makers better insight into digital tools and their effects on second language learning and teaching.

Other constraints include indications in Chapter 4 that the predictive text feature of AI KAKU may potentially disrupt the writing process for some users deserves more attention. While the majority of the qualitative feedback received from the participants in the experiments described in Chapter 2, Chapter 3, and Chapter 4 resulted in positive reactions to the features of AI KAKU, the researcher would like to further investigate more detailed human-computer interaction with the tools described in this research. The voices of a few participants that indicated potential disruption to their writing process should not be ignored and the researcher would like to investigate biometric factors such as eye tracking and keystroke logging in future studies to determine how students of different proficiency interact with intelligent writing assistants such as AI KAKU.

Other potential constraints and negative effects such as mental degeneration (e.g., spell-check has negatively influenced society's ability to spell words) should also be considered when technologies employed by AI KAKU become more widespread in society. Other concerns of AI assistance discouraging originality or encouraging homogeneity due to an overreliance on these systems should be carefully considered by developers and researchers. Issues of creativity and agency in EFL writing in academic settings are not new (Alghamdi & Alnowaiser, 2017; Wei, 2020) and educators should consider how to implement these new tools in the classroom while ensuring student agency and creativity are encouraged in the classroom.

The writing assistant described in this research was intentionally designed to maintain user agency and creativity. Word suggestions are only shown after a 2.5 second pause in writing (indicating a mental pause in the writing process) and only one word at a time is shown to the user instead of giving phrase or sentence level suggestions. This discourages over-reliance on word suggestions as the user is still responsible for formulating the bulk of the text. AI KAKU can be categorized as a system described by Dale & Viethen (2021) as a “short-leash” AI assistant in that the amount of assistance given to the user is limited.

Interestingly, going back to the quotation first introduced in this dissertation by Plato, we can find that at each inflection of “cognitive outsourcing”, humanity has taken advantage of that outsourcing in order to improve on other areas of interest. Danaher (2018) is instrumental in this regard with this work on ethical frameworks for AI assistants. He argues that mental resources are finite and difficult and having the latitude to use higher order thinking skills is a benefit to society. In that sense, he argues for having a new form of “algorithmic cognitive outsourcing” which will help us reduce that mental labour.

The researcher is hoping to show that students will use these “digital assistants” as part of a “new normal”. Moreover, the advancements in natural language processing and machine learning have led to the development of more sophisticated intelligent writing assistants which offer synchronous feedback to the writer compared to traditional text editors. One of the aims of this research is to show how a potentially disruptive technology (AI-assisted writing assistance) impacts users while they are using the technology with the knowledge that these kinds of tools may become a part of everyday life in the future. To the same extent that computer assisted spell- grammar-check has permeated writing in the modern age, this researcher believes AI-assistance will also be something that cannot be separated from the learning and writing process.

Appendix A (Participant consent form)[14]

About AI-based writing assistant’s impact on English language learners’ writing proficiency
【 AI-based writing assistant’s impact on English language learners’ writing proficiency】の研究について

This course will be used as an experiment for the research titled “AI-based writing assistant’s impact on English language learners’ writing proficiency." We would like to request your consent in participating in this experiment. Please read the research details below and respond to the consent form. Thank you.

本講座は、" AI-based writing assistant’s impact on English language learners’ writing proficiency "と題した研究の実験として使用させていただきます。この実験に参加するにあたり、同意をお願いしたいと思います。下記の研究内容を一読いただき、ご確認の上同意書にご記入くださいますようお願いします。

(1) Research summary・研究概要について
Doctoral students at the Cross Lab in the Transdisciplinary Science and Engineering Department, Tokyo Institute of Technology (Tokyo Tech), are investigating the application of an AI-based predictive next word writing tool that is specifically aimed at assisting ESL university students. Progress in machine learning (ML) and natural language processing (NLP) have given us tools that can help ESL learners in their struggle with producing text written in English. Having real-time word-choice suggestions based on the context of the users' input can help students produce richer text and focus more on higher-level thinking skills. These higher-level thinking skills are said to be better utilized by those who have high metacognition, or the ability to think about thinking.
東京工業大学（東工大）環境社会理工学院クロス研究室の博士課程グループでは、ESL の大学生を支援することを目的とした AI ベースの単語予測分析ライティングツールの応用を研究しています。機械学習（ML）や自然言語処理（NLP）の進歩により、英語で書かれた文章を作成するのに苦労している ESL 学習者を支援するツールが提供されるようになりました。ユーザーが入力した文脈に基づき、リアルタイムで単語選択の提案をすることが可能になることで、学生は豊かな文章表現を作成し、より高いレベルの思考力に集中することができます。これらの思考スキルは、メタ認知度が高く自信を持った人が多いのが特徴です。メタ認知能力とは、人とのコミュニケーション、仕事や目標を定める能力に優れていると言われています。

(2) Significance and goals of research・研究の意義と目的について
There has been little research into the impact of recently developed AI technologies on the writing proficiency of second language learners. The AI-based technologies that have emerged range from Google’s Smart Compose writing assistant that helps users complete their sentences to Co-Writer Universal which claims to offer a full suite of writing assistance that integrates spell checking, grammar checking and word prediction under the framework of a word processor. Predictive text seems to be a potentially useful tool for second language (L2) learners (in particular, this research will focus on English language learners (ELLs)), as the

[14] This consent form was given to participants who joined the three experiments described in this dissertation.

technology offers real time word-choice suggestions to writers based on the context of the words in a given sentence and the first words typed.
近年開発された AI 技術の分野において、第二言語学習者の文章作成能力に与える影響についての研究はほとんど行われていません。ユーザー側の文章完成を補助する Google Smart Compose ライティングアシスタントから、ワードプロセッサーの基礎構成によるスペルチェック、文法チェック、単語予測を統合したライティング支援を提供する Co-Writer Universal などが学者向けの主な AI 技術です。よって、予測テキストは今後も第二言語（L2）学習者（特に本研究では英語学習者（ELL）に焦点を当てています）にとって必要なツールとなる可能性があると考えられます。

Aside from using assistive digital tools, ELLs can benefit from metacognition in activating their higher-level thinking skills needed for producing content in a foreign language. Studies have shown that metacognition benefits learning both in academic and non-academic settings across different age levels. Developing metacognition and responsible digital tool use in English instruction can empower learners to become competitive and competitive in the highly technical world.
ELL は、デジタルツールのサポートを使用する以外に、メタ認知を活性化することにより外国語でコンテンツを作成するために必要な高度な思考スキル得ることができます。 研究によると、メタ認知は、さまざまな年齢レベルの学問的および非学問的環境の両方で学習するのに役立つことが示されています。 英語教育におけるメタ認知と正しいデジタルツールの開発や活用をすることで、ハイレベルな国際環境で成長できる競争力を身につけることができます。

(3) Research methods・研究方法について

In this experiment, you would be interacting with a system called AI-KAKU while you are answering a TOEFL prompt. After each writing assignment, you will additionally be asked to answer a questionnaire to rate your experience using the system.
この実験では、TOEFL のプロンプトに答えながら、AI-KAKU と呼ばれるシステムと対話します。それぞれのライティング課題の後、システムの使用感を評価するためのアンケートに回答していただきます。

(4) Storage of data and their use in other research・個人情報の管理、他研究での利用について

No personal data will be stored, and your writing, consent form, and surveys will be submitted via AI-KAKU. All activities to this research are voluntary and no personal data will be collected.
個人データは保存されません。また、提供いただいた文章、同意書、アンケートは AI-KAKU を使用して提出されます。この研究に対する全ての活動・参加は任意であり、個人情報が収集されることはありません。

(5) Forecasting results (merits and demerits)・予測される結果（メリットとデメリット）について

The proposed writing tool might make it easier for you to complete your writing as you type. Because the system is non-invasive, there should not be any demerits to using the system. In addition, if you do not wish to participate, then your performance in this course will not be adversely affected.

提案されているライティングツールを利用することで、文章を作成する過程において難なく完成させせることができるのではないでしょうか。システムを使用することによるデメリットはないと考えられます。また、参加を希望しない場合でも、本コースの成績に影響はありません。

(6) Cooperation with the research is voluntary and retraction of consent is possible at any time・研究協力の任意性と撤回の自由について
You have the complete freedom to participate or not participate in this research. Furthermore, if you no longer wish to cooperate even after having previously given consent, as soon as a request for retraction is received, the further experiment will be canceled and data whose sole purpose is for research will be destroyed. The retraction form will be introduced during explanation outside of class hours and can be requested by contacting the researcher. The retraction will not penalize you in any way; in particular, the retraction will not affect your grade in the class.
本研究への参加については個人の自由です。また、事前に同意を得ていたにもかかわらず協力しない場合は、撤回の申し出があった時点で、それ以降の実験は中止され、研究目的としたデータは破棄されます。撤回申請書については担当教員までお問い合わせください。取り消し後のペナルティはなく、授業の成績に影響はありません。

(7) Protection of personal information・個人情報の保護について
Because the name of the research subject is anonymized, personal information regarding the research subject can in absolutely no way be leaked outside of the research team’s control.
研究対象の名前は匿名化されているため、研究対象に関する個人情報を研究チームが外部に漏洩することは絶対にありません。

(8) Publication of the research results・研究成果の公表について
Research results may be publicized through academic associations in educational and computational fields such as the Japan Society for Educational Technology; committees of specialists; international meetings; and in educational and computational journals. In such cases as well, absolutely no identifiable information specific to participants are released.
研究成果は、日本教育工学会などの教育および計算分野の学会、専門委員会、国際会議、教育及び計算ジャーナルを通じて公表される可能性があります。そのような場合でも、実験参加者を識別できるような情報は絶対に公開されません。

(9) Expenses・費用について
The research subjects will bear absolutely no supplementary expenses for the tests and analysis that accompany the research. There is no remuneration for the participants.
研究に伴う測定・解析によって研究対象者が負担する付加的な費用は一切ありません。また、実験への協力に対する謝礼も一切発生しません。

(10) Compensation for adverse health effects・健康被害の補償について
No adverse health effects are anticipated. Should problems arise, please do not hesitate to contact the designated person for this research (contact details below).
実験による健康への悪影響は考えられにくく、また予想されません。 問題が発生した場合は、指定された担当者宛にご連絡ください（下記の連絡先）。

(11) For inquiries regarding this research・本件に関する問い合わせ先:

Contact regarding the research・研究についての連絡先:
School of Environment and Society, Department of Transdisciplinary Science and Engineering, Tokyo Institute of Technology
東京工業大学　環境・社会理工学院　融合理工学系
John Maurice Gayed
Email・メール: [xxx]@m.titech.ac.jp

Appendix B (Writing prompts given to participants)

Questions prompts given in experiment 1

1. **Read the question below.**
   Give yourself 30 minutes to plan, write, and revise your essay. Typically, an effective response will contain a minimum of 300 words.
   以下の質問を読んでください。
   エッセイの計画、執筆、修正には 30 分を目安にしてください。一般的に、効果的な回答は最低でも 300 ワードを必要とします。

   Question: Do you agree or disagree with the following statement? Overall, the widespread use of the internet has a mostly positive effect on life in today's world. Use specific reasons and examples to support your answer.
   質問です。以下の文章に賛成ですか、反対ですか？

   全体的に見て、インターネットの普及は、今日の世界の生活にほとんど良い影響を与えている。

   具体的な理由や例を用いて、あなたの答えを裏付けてください。

2. **Read the question below.**
   Give yourself 30 minutes to plan, write, and revise your essay. Typically, an effective response will contain a minimum of 300 words.

   Question: Do you agree or disagree with the following statement? It is better for children to grow up in the countryside than in a large city. Use specific reasons and examples to support your answer.

   以下の質問を読んでください。
   エッセイの計画、執筆、修正には 30 分を目安にしてください。一般的に、効果的な回答は最低でも 300 ワードを必要とします。

   質問です。以下の文章に賛成ですか、反対ですか？
   子供にとっては、大都市よりも田舎で育つ方が良い。

   具体的な理由や例を用いて、あなたの答えを裏付けてください。

Questions prompts given in experiment 2

Read the question below.
Give yourself 30 minutes to plan, write, and revise your essay. Typically, an effective response will contain a minimum of 300 words.
以下の質問を読んでください。エッセイの計画、執筆、修正には30分を目安にしてください。一般的に、効果的な回答は最低でも300ワードを必要とします。

1. **Question: Do you agree or disagree with the following statement? Overall, the widespread use of the internet has a mostly positive effect on life in today's world.**
   Use specific reasons and examples to support your answer.
   質問: 以下の文章に賛成ですか、反対ですか？全体的に見て、インターネットの普及は、今日の世界の生活にほとんど良い影響を与えている。

   具体的な理由や例を用いて、あなたの答えを裏付けてください。

2. **Question: Do you agree or disagree with the following statement? It is better for children to grow up in the countryside than in a large city.**
   Use specific reasons and examples to support your answer.
   質問: 以下の文章に賛成ですか、反対ですか？子供にとっては、大都市よりも田舎で育つ方が良い。

   具体的な理由や例を用いて、あなたの答えを裏付けてください。

3. **Question: Do you agree or disagree with the following statement? To succeed in school or work, the ability to adapt to changing conditions or circumstances is more important than excellent knowledge in a job or a field of study.**
   Use specific reasons and examples to support your answer.
   質問: 以下の文章に賛成ですか、反対ですか？学校や仕事で成功するためには、仕事や専門分野の優れた知識よりも、変化する条件や状況に適応する能力の方が重要である。

   具体的な理由や例を用いて、あなたの答えを裏付けてください。

4. **Question: Do you agree or disagree with the following statement? Improving schools is the most important factor for the successful development of a country.**
   Use specific reasons and examples to support your answer.
   質問: 以下の文章に賛成ですか、反対ですか？学校の改善は、国の発展を成功させるための最も重要な要素です。

   具体的な理由や例を用いて、あなたの答えを裏付けてください。

Questions prompts given in experiment 3

Read the question below.
Give yourself 30 minutes to plan, write, and revise your essay. Typically, an effective response will contain a minimum of 300 words.
以下の質問を読んでください。エッセイの計画、執筆、修正には 30 分を目安にしてください。一般的に、効果的な回答は最低でも 300 ワードを必要とします。

(1) Pre-test writing prompt:
Do you agree or disagree with the following statement? Improving schools is the most important factor for the successful development of a country.
Use specific reasons and examples to support your answer.
質問: 以下の文章に賛成ですか、反対ですか？学校の改善は、国の発展を成功させるための最も重要な要素です。

具体的な理由や例を用いて、あなたの答えを裏付けてください。

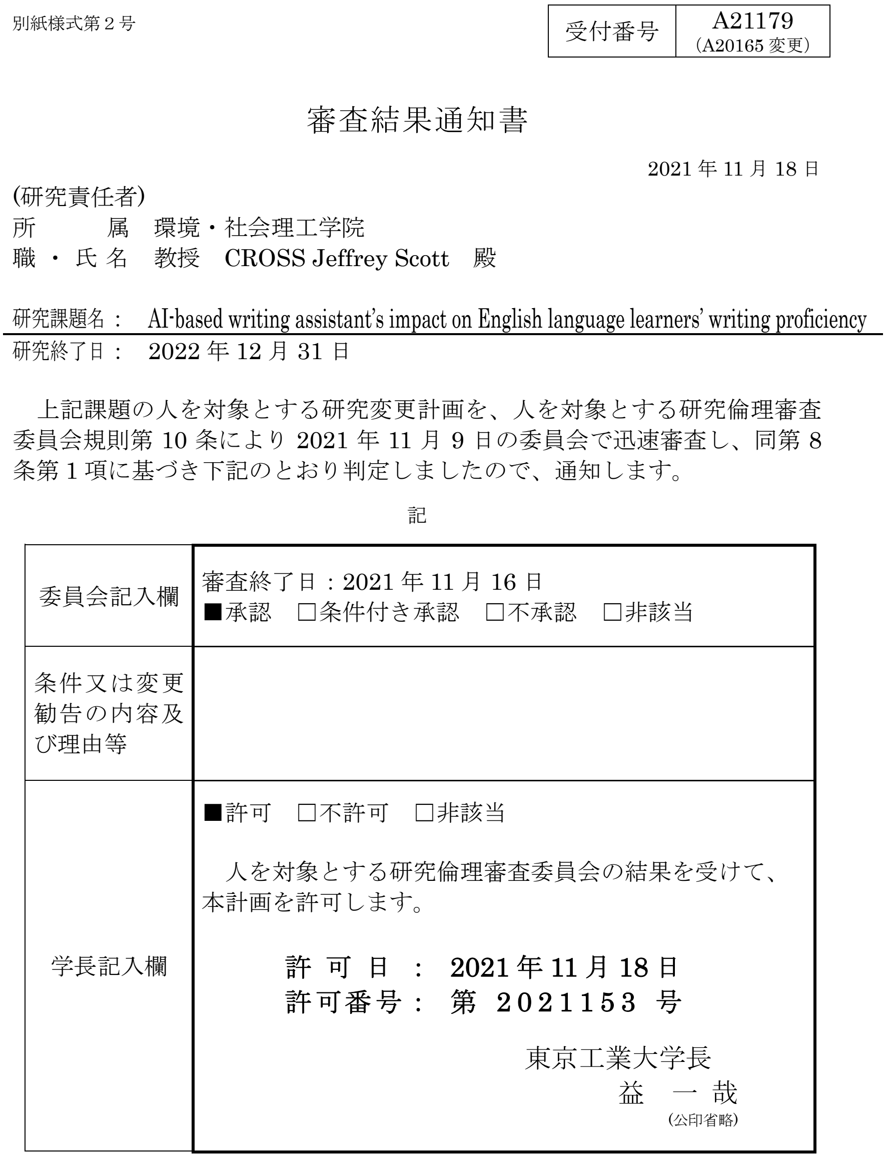
(2) **Writing prompt 1:**
Question: Do you agree or disagree with the following statement? It is better for children to grow up in the countryside than in a large city.
Use specific reasons and examples to support your answer.
質問: 以下の文章に賛成ですか、反対ですか？子供にとっては、大都市よりも田舎で育つ方が良い。

具体的な理由や例を用いて、あなたの答えを裏付けてください。

(3) **Writing prompt 2:**
Question: Do you agree or disagree with the following statement? Overall, the widespread use of the internet has a mostly positive effect on life in today’s world.
Use specific reasons and examples to support your answer.
質問: 以下の文章に賛成ですか、反対ですか？全体的に見て、インターネットの普及は、今日の世界の生活にほとんど良い影響を与えている。

具体的な理由や例を用いて、あなたの答えを裏付けてください。

Appendix C (Ethics approval for human-subjects research)



別紙様式第２号

受付番号	A21179 （A20165 変更）

審査結果通知書

2021 年 11 月 18 日

(研究責任者)
所　　　属　環境・社会理工学院
職・氏名　教授　CROSS Jeffrey Scott　殿

研究課題名：　AI-based writing assistant's impact on English language learners' writing proficiency
研究終了日：　2022 年 12 月 31 日

上記課題の人を対象とする研究変更計画を、人を対象とする研究倫理審査委員会規則第 10 条により 2021 年 11 月 9 日の委員会で迅速審査し、同第 8 条第 1 項に基づき下記のとおり判定しましたので、通知します。

記

委員会記入欄	審査終了日：2021 年 11 月 16 日 ■承認　□条件付き承認　□不承認　□非該当
条件又は変更勧告の内容及び理由等	
学長記入欄	■許可　□不許可　□非該当 人を対象とする研究倫理審査委員会の結果を受けて、本計画を許可します。 許可日：　2021 年 11 月 18 日 許可番号：　第　2021153　号 東京工業大学長 益　一哉 （公印省略）

[English translation for reference]

November 18, 2021

(Principal Investigator)
Professor Jeffrey Scott CROSS
School of Environment and Society

Review Result Notification

Research Title：AI-based writing assistant's impact on English language learners' writing proficiency
Ending Date：December 31, 2022

Dear Dr. Cross,

This is to notify that the Human Subjects Research Ethics Review Committee reviewed the above titled research at the expedited committee meeting on November 9, 2021 pursuant to Article 10 of Human Subjects Research Ethics Review Committee Rules and has determined, in accordance with Article 8.1 of the above said rules, as follows:

Committee's Decision	Review completed on November 16, 2021 ■Approved □Conditioned □Disapproved □Not Applicable
Conditions or modification suggestions	
University President's Decision	■Permitted □Denied □ Not Applicable I hereby permit the Principal Investigator to conduct the above titled research based on the approval made by the Human Subjects Research Ethics Review Committee. **Date of Permission: November 18, 2021** **Permit Number: 2021153** Kazuya Masu President of Tokyo Institute of Technology (Official seal omitted)

*In the event of inconsistency or discrepancy between the Japanese version (if any) and any other language version, the Japanese language version shall prevail.

Appendix D (ETS Independent Writing Rubric)

Table 26. ETS Independent Writing Rubric.

Score	*Task Description*
5	An essay at this level largely accomplishes all of the following: • Effectively addresses the topic and task • Is well organized and well developed, using clearly appropriate explanations, exemplifications and/or details • Displays unity, progression and coherence • Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice and idiomaticity, though it may have minor lexical or grammatical errors
4	An essay at this level largely accomplishes all of the following: • Addresses the topic and task well, though some points may not be fully elaborated • Is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications and/or details • Displays unity, progression and coherence, though it may contain occasional redundancy, digression, or unclear connections • Displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form or use of idiomatic language that do not interfere with meaning
3	An essay at this level is marked by one or more of the following: • Addresses the topic and task using somewhat developed explanations, exemplifications and/or details • Displays unity, progression and coherence, though connection of ideas may be occasionally obscured • May demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning • May display accurate but limited range of syntactic structures and vocabulary
2	An essay at this level may reveal one or more of the following weaknesses: • Limited development in response to the topic and task • Inadequate organization or connection of ideas • Inappropriate or insufficient exemplifications, explanations or details to support or illustrate generalizations in response to the task • A noticeably inappropriate choice of words or word forms • An accumulation of errors in sentence structure and/or usage

1 An essay at this level is seriously flawed by one or more of the following weaknesses:

- Serious disorganization or underdevelopment
- Little or no detail, or irrelevant specifics, or questionable responsiveness to the task
- Serious and frequent errors in sentence structure or usage

0 An essay at this level:

- Merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

Appendix E (Metacognitive Writing Strategies Questionnaire)

Table 27. Metacognitive Writing Strategies Questionnaire (MWSQ).

#	*Read each statement carefully. Consider if the statement generally applies to you. 5-point scale from strongly disagree (1) to strongly agree (5).*
1	I read the directions carefully before writing. 私は文章を書く前にじっくり説明を読んだ。
2	I knew what I was required to do for the writing task. 私はこの文章作成の課題について必要とされていることが何かを理解していた。
3	I planned on what ideas or things I should include in my essay before I started writing. 私は小論文を書く前に内容やアイデアについて事前に考えていた。
4	I planned on the organization of my essay before I started writing. 私は小論文を書く前に、どのように構成するか事前にまとめていた。
5	I mainly thought in my language first and then translated my thoughts from my language into English. 私はまず母国語で考え、それから英語に通訳をした。
6	I tried to use some complex sentence structures. 私は高度な文章構成を使用する努力をした。
7	I often stopped to read my own writing and think about what to write next. 私は度々書いた内容を読むため一旦中断し、次に書くことを考えた。
8	I revised to fix grammar mistakes and/or other language issues I noticed. 私は文法やその他の間違いを修正した。
9	I tried to improve the content and/or ideas in my essay. 私は小論文の内容またはアイデアをより良くするための努力をした。
10	I reorganized some part(s) of the essay to make it more coherent. 私は小論文がよりまとまり、一貫性のあるものになるよう所々の箇所で再編成をした。

Appendix F (Software usefulness questionnaire)

Table 28. Software usefulness questionnaire.

#	*Read each statement carefully. 5-point scale from strongly disagree (1) to strongly agree (5).*
1	I could understand the question. 質問の内容は理解できました。
2	I was able to express my opinions when writing for this task. この課題のライティングでは、自分の意見を述べることができました。
3	The writing was stressful for me. このライティング作業は、ストレスを感じました。
4	When applicable: The word suggestions given to me were useful. 該当する場合：私に示された提案という言葉は役に立ちました。
5	When applicable: The translation of my English displayed to me helped me with my writing. 該当する場合。表示された私の英語を翻訳してくれたことで、文章を書くのに役立ちました。

Appendix G (Survey to Educators)

1. Contact info: John Maurice Gayed Tokyo Institute of Technology Email - gayedsensei@gmail.com If you would like to be contacted about this research or would like to participate in a follow-up interview, please enter your information below.
   a. First name
   b. Last name
   c. Email address
2. Please indicate your country of residence
3. Please indicate the gender you identify with
4. Are you a teacher, trainer, tutor or professional working in the education and training sector?
5. If you are not a teacher, trainer, tutor or professional working in the education and training sector, please indicate your profession and your interest in completing this survey.
6. What type of institution do you work for? Check all that apply.
   a. Adults (above university age)
   b. University / College students
   c. High-school students
   d. Junior high-school students
   e. Elementary school students
   f. Kindergarten students
7. What types of students do you work with? Check all that apply.
   a. Adults (above university age)
   b. University / College students
   c. High-school students
   d. Junior high-school students
   e. Elementary school students
   f. Kindergarten students
8. What is your highest level of education?
   a. Doctorate
   b. Master's degree
   c. Bachelor's degree
   d. Associate degree
   e. Technical or occupational certificate
9. Including this year, how many years have you taught?
   a. Less than one year
   b. 1-3 years
   c. 4-6 years
   d. 7-10 years
   e. 11-15 years
   f. 16-20 years
   g. 21-25 years
   h. More than 26 years
10. What subject do you teach? Check all that apply.
    a. Business and economics
    b. Arts and humanities

c. Engineering and technology
d. Life / Physical sciences
e. Social sciences
f. Creative Art and Design
g. Law
h. Health and Medicine
i. Travel and Hospitality
j. Other

11. How do you define artificial intelligence?

12. Planning Technology-Supported Instruction

(1) Strongly Agree (2) Moderately Agree (3) Mildly Agree (4) Mildly Disagree (5) Moderately Disagree (6) Strongly Disagree

a. What designing my lessons, I regularly think about whether technology could enhance my teaching or student learning
b. When selecting education technologies, I refer to and base my selections on current research on their effectiveness.
c. I am comfortable planning for class sessions that involve students using technology during instruction.

13. Technology and Assessment

(1) Strongly Agree (2) Moderately Agree (3) Mildly Agree (4) Mildly Disagree (5) Moderately Disagree (6) Strongly Disagree

a. I feel comfortable using technology to help me manage student assessment data (e.g., using electronic gradebooks).
b. I have effective strategies for assessing the content of students' technology-supported work.
c. I am comfortable using technology to help me gather, analyze, and interpret data on student progress.

14. What kind of technology is used in your classroom? Check all that apply.

a. Computers (Notebook/Chromebook, etc)
b. Interactive whiteboards
c. Tablet computer
d. Smartphone
e. Learning software (Quizlet / Khan Academy, etc)
f. Learning management system (Moodle, Google Classroom, Canvas)
g. Clicker response systems
h. We don't use any technology in the classroom

15. I am familiar with current AI technologies being used by my students in my class.

(1) Strongly Agree (2) Moderately Agree (3) Mildly Agree (4) Mildly Disagree (5) Moderately Disagree (6) Strongly Disagree

16. To your knowledge, are you using any form of AI or AI-enabled technology as a teaching aid? (e.g., Alexa, Siri, Grammarly, etc.)

a. If yes, how are you using the AI aid in your classroom?

17. A student using AI assistance in school is unfair to other students who don't use such assistance.

(1) Strongly Agree (2) Moderately Agree (3) Mildly Agree (4) Mildly Disagree (5) Moderately Disagree (6) Strongly Disagree

a. Can you explain in more detail why you feel this way?

18. Students should be able to use any type of assistance (Grammarly, spell/grammar check, next word prediction, etc.) to submit written assignments.

(1) Strongly Agree (2) Moderately Agree (3) Mildly Agree (4) Mildly Disagree (5) Moderately Disagree (6) Strongly Disagree

19. AI augmented writing (writing where the student is using AI assistance) is a form of plagiarism as the writer is using external assistance to complete an assignment.
    (1) Strongly Agree (2) Moderately Agree (3) Mildly Agree (4) Mildly Disagree (5) Moderately Disagree (6) Strongly Disagree
20. Do you think AI is influencing your approach to teaching?
    a. Please explain how AI is influencing your approach to teaching.
21. Do you think AI is impacting your students' learning outcomes?
    a. Please explain how AI is influencing your students' learning outcomes.
22. Have you attended any professional development courses that covered issues such as AI in education or the ethical use of AI in education?
23. How prepared do you feel to manage AI-supported learning with your classes?
    (1) Very prepared (2) Prepared (3) Somewhat prepared (4) Not very prepared (5) Not prepared at all
24. Does your school have a vision for how technology should be used by students and teachers to improve teaching and learning?
    (1) No. As far as I know, there is no vision for technology use, written or unwritten.
    (2) I don't know.
    (3) Yes. But it isn't written down, and many teachers (including me) aren't really aware of what the vision is.
    (4) Yes. It isn't written down, but it has been clearly shared with me and other teachers.
    (5) Yes, a formal, written vision, but many teachers have not actually seen it.
    (6) Yes, a formal, written vision that has been shared with myself and other teachers.
25. My school/institution has guidelines about the use of AI in the classroom.
26. To what extent does your school encourage innovative teaching practices?
    (1) Discouraged (2) Tolerated (3) Supported, but not rewarded (3) Rewarded (e.g., through public recognition, equipment, professional development)
27. Please rank the following courses/topics in terms of which you would most like to learn more about in the future.
    (1) AI based digital assistants (tutors) and their application to education
    (2) AI based learning analytics
    (3) AI in school management systems
    (4) AI agents in the smart classroom
    (5) AI in education ethical frameworks
28. It is important to understand ethical, legal and societal issues related to technology use and using technology in ethical ways.
    (1) Strongly Agree (2) Moderately Agree (3) Mildly Agree (4) Mildly Disagree (5) Moderately Disagree (6) Strongly Disagree
29. Overall, with regards to technology, AI, and the future are you:
    (1) Optimistic ---------------- (10) pessimistic

Appendix I (Writing Samples)

Experiment 3 – Pre-test

I agree with this idea. I have some reasons to think that. First, If someone don't know anything such as language or number, it is so dangerous, I think. For example, "Don't enter the area" a place where it has mines says, but If a boy can't read the letter, it is unsafe. And when I go shopping, I can find how much I buy even if shop clerk tells a lie. it is from the result I go to school and I learner calculate. From these situation, I find there is a big gap between some who go to school and other don't go to school. the country many people are in danger will not success development. it is start line at first to go to school and get necessary knowledge to live. Second getting high level knowledge leads getting choices for job, I think. Low education leads less knowledge to work. after all it leads poverty. However high education give a chance to be a doctor or get some job. I am sure that improving schools is important factor for the successful of a country from these reasons.

Tokens	MTLD	Ldensity	LFP	MLTunit	Clause/Tunit
190	78.14	0.54	0.07	12.12	1.93

Rater1	Rater2	Rater3	Rater4
2	2	3	3

Experiment 3 – Writing 1 treatment group

I agree that the widespread use of the internet has a mostly positive effect on life in today's world. I have three reasons to think so. First, by using the Internet, you can quickly find out the information you want to know. You can also get the latest information at any time. Therefore, you can feel free to deepen your knowledge about your hobbies and have a more enjoyable life. If you don't use the internet, it will take time to find the information you want. Therefore, it is possible to use time efficiently by using the Internet. Second, the internet is connected all over the world. Therefore, it is possible to interact with people who are far away while being there just by using the Internet. And if you join a foreigner, you can easily learn about different values and cultures, and your way of thinking becomes more flexible. In addition, you can contact efficiently by using the Internet when you want to make the best use of the characteristics of each place in your work. Third, with the spread of the Internet, many new occupations have increased. Nowadays, there are youtubers who shoot their own videos and upload them to youtube to earn advertising revenue. Their influence is great, and it has become very popular because you can watch only the fields you want to see more easily than on TV, which is a new entertainment for us. From these, the Internet makes our lives more convenient and gives us a lot of entertainment. It is also essential for work where keeping in touch is important. Therefore, I think the spread of the Internet has a positive effect on the lives of today's world.

Tokens	MTLD	Ldensity	LFP	MLTunit	Clause/Tunit
289	73.11	0.51	0.13	15.2	1.5

Rater1	Rater2	Rater3	Rater4
4	5	5	5

Writing Samples 2

<table>
<tr><th colspan="6">Experiment 2 – Control</th></tr>
<tr><td colspan="6">I agree. I think it is better for children to grow up in the countryside than in a large city. There are three reasons. First of all, in the countryside, We can touch nature. Touching nature soothe our mind, such as a forest full of greenery, birds chirping, and the murmuring of a river. We can build a healthy body and develop a rich mind. Actually, when I was little child, I used to go to the river and make mud dumplings. Thanks to this experience, I was able to acquire the power to look for fun and enjoy myself without having to touch the media. Also, we can relax while feeling the four seasons. I think one of the advantages is that you can eat fresh local vegetables right away. Second, we can stay safer than in the city. There are many places where there are few crimes, so we can spend our time with peace of mind. Also, there is less risk of traffic accident. The third reason is that there is a deep community connection. Because the area is small and many people live for a long time, we can deepen our ties with local people. It also improves communication skills, increases smiles, and relaxes our mind.</td></tr>
<tr><th>Tokens</th><th>MTLD</th><th>LFP</th><th>EIKEN</th><th>Cognitive Load</th><th>Intrinsic Load</th></tr>
<tr><td>210</td><td>84</td><td>0.14</td><td>2</td><td>8</td><td>7</td></tr>
<tr><td colspan="6"></td></tr>
<tr><th colspan="6">Experiment 2 – Treatment</th></tr>
<tr><td colspan="6">In my opinion, the ability to adopt to changing condition and circumstance is important than excellent knowledge in a job or a field of study. No matter how capable you are, you cannot succeed unless you adopt to the situation or change. If you are weak in change you will lose to stress and you will not be able to use your abilities. It looks amazing if you have the ability, but I do not think it is meaningless if you do not harmonize well with the people and environment around you. Talent can also be acquired if you study hard many time. Successful people around me try to adopt quickly to the situation no matter how much it changes. And it is because they can adopt it, they are showing their ability. If you are stick to your limited way of thinking and do not listen to other people's opinion at all, you will not be able to do it in companies and schools that value team strength. Also, I think there will be many opportunities in the future for sudden accidents and sudden judgment and responses in my life. At such time, I think people who can respond flexibly to the environment are the ones who are evaluates by others. In recent years, it is also true that more and more companies have hires people to find jobs based on their personalities rather than focusing on academic ability and academic background.</td></tr>
<tr><th>Tokens</th><th>MTLD</th><th>LFP</th><th>EIKEN</th><th>Cognitive Load</th><th>Intrinsic Load</th></tr>
<tr><td>244</td><td>72</td><td>0.2</td><td>2</td><td>7</td><td>7</td></tr>
</table>

Writing Samples 3

Experiment 1 – Control
I disagree. I think children to live large city better than a countryside. Because there are a lot of people there. Why I think that there are a lot of people is better. I think that children should grow to look adult people because looking them teach children a lot of things. For example, how to greed to boss, to work is hard, and so on. Children separate bad or right to look them. However, in a countryside, adult people are less, so

Tokens	Lexical Diversity	Clause/Tunit	
83	31.9	1.3	

Experiment 1 – Treatment
I agree that the internet has a mostly positive effect on life. Because we can see some information with using internet. We can see soon that we want to know and the working is finished soon. It gives us the time in order to do something that we like. And Stress is lessen. But it has bad aspect. There are people who use to bad things. For example there is deceive human. If they are reformed, it is better.

Tokens	Lexical Diversity	Clause/Tunit	
79	60	1.50	

Appendix I (User Feedback on AI KAKU)

Table 29. Participants' feedback on AI KAKU.

Sentiment	*Comments marked as either positive (+), neutral (*) or negative (-)*
+	I was able to be more conscious of paragraph structure and splicing than when I wrote the first time, and I think my writing became more coherent.
+	I learn how to write my opinion thanks to AI KAKU.
+	I understand what I have to write better than before watch video of training 1.
+	It is very helps improve my writing skill.
+	I was able to learn how to compose sentences.
+	At first, I didn’t know what kind of structure to write, but I understood how to write by watching the video.
+	I was able to improve my ability to write sentences.
+	It took me a long time to write my opinion in 300 words. I thought I need to practice more. Writing the outline and then text made it easier for me to write.
*	I felt it was hard to write essays. I think that I want to write better passages.
*	It was a little difficult to think of examples when writing sentences, but I got used to it after a few times.
*	It was hard work for me to write essays even if writing was in Japanese. I prefer more specific topics to write. I needed to determine what I write concretely.
*	I think telling my statement in Japanese easy, but in English, it is so difficult.
-	It is difficult for me to write my opinion in English.
-	I could not use well this Meta AI KAKU system... I want to know the scoring guide more clearly and in detail.
-	I want the feedback of my answer. Such as this part has grammar mistake, you should not use this word so many times etc.
-	Showing the next predicted word was a distraction. I didn’t need it.
-	I was so tired, and this work was very difficult.
-	I solved 3 questions in a row, so I could not keep my concentration.
-	I was a little disappointed that the score was not going up, but that the one I wrote at the beginning was the best.

Bibliography

2020 Report on Test Takers Worldwide—TOEIC Speaking & Writing Tests -. (2020). *ETS Research Report Series*, 60.

Aaron, L. S., & Roche, C. M. (2013). Stemming the Tide of Academic Dishonesty in Higher Education: It Takes a Village. *Journal of Educational Technology Systems*, *42*(2), 161–196. https://doi.org/10.2190/ET.42.2.h

Abrams, L., & Davis, D. K. (2016). The Tip-of-the-Tongue Phenomenon. *Cognition, Language and Aging*, 13–54.

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. *Automatic Treatment and Analysis of Learner Corpus Data*, 249–264.

AI in Education Market Size, Share | Industry Forecast Report to 2030. (n.d.). P&S Intelligence. Retrieved July 23, 2022, from https://www.psmarketresearch.com/market-analysis/ai-in-education-market

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91–93.

Alfaqiri, M. (2018). English second language writing difficulties and challenges among Saudi Arabian language learners. *Journal for the Study of English Linguistics*, *6*(1), 24–36.

Alghamdi, F. M. A., & Alnowaiser, S. A. M. (2017). Achieving Flex in the Inflexible: Dealing with Individual Differences in Highly Structured EFL Preparatory College Courses. *English Language Teaching*, *10*(6), 151–159.

Alisaari, J., & Heikkola, L. M. (2016). Increasing fluency in L2 writing with singing. *Studies in Second Language Learning and Teaching*, *6*(2), 271–292.

Allen, G., & Chan, T. (2017). *Artificial Intelligence and National Security*. Belfer Center for Science and International Affairs.

Amini Farsani, M., & Babaii, E. (2020). Applied linguistics research in three decades: A methodological synthesis of graduate theses in an EFL context. *Quality & Quantity*, *54*. https://doi.org/10.1007/s11135-020-00984-w

Andringa, S., & Godfroid, A. (2020). Sampling Bias and the Problem of Generalizability in Applied Linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142. https://doi.org/10.1017/S0267190520000033

Archibald, A., & C. Jeffery, G. (2000). Second language acquisition and writing: A multi-disciplinary approach. *Learning and Instruction*, *10*(1), 1–11. https://doi.org/10.1016/S0959-4752(99)00015-8

Arnold, K. C., Chauncey, K., & Gajos, K. Z. (2020). Predictive text encourages predictable writing. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 128–138. https://doi.org/10.1145/3377325.3377523

Ashton, T. M. (1999). Spell Checking: Making Writing Meaningful in the Incusive Classroom. *Teaching Exceptional Children*, *32*(2), 24–27.

Assembly Concurrent Resolution No. 215. Asilomar AI Principles. (2018, September 7). Legislative Counsel's Digest. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180ACR215

Aydoğan, H., & Akbarov, A. A. (2014). The four basic language skills, whole language & integrated skill approach in mainstream university classrooms in Turkey. *Mediterranean Journal of Social Sciences*, *5*(9), 672–672.

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction*, *16*(5), 389–400.

Azevedo, R. (2020). Reflections on the field of metacognition: Issues, challenges, and opportunities. *Metacognition and Learning*.

Baker, R. S. J. d., D'Mello, S. K., Rodrigo, Ma. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Basturkmen, H., & Lewis, M. (2002). Learner perspectives of success in an EAP writing course. *Assessing Writing*, *8*(1), 31–46.

Bates, T., Cobo, C., Mariño, O., & Wheeler, S. (2020). Can artificial intelligence transform higher education? *International Journal of Educational Technology in Higher Education*, *17*(1), 42. https://doi.org/10.1186/s41239-020-00218-x

Bax, S. (2008). Bridges, chopsticks and shoelaces: Normalising computers and computer technologies in language classrooms. *Abstract for Keynote Speech at WorldCALL*.

Beare, K. (2020). *How Many People Learn English Around the World?* ThoughtCo. https://www.thoughtco.com/how-many-people-learn-english-globally-1210367

Beers, S. F., & Nagy, W. E. (2009). Syntactic complexity as a predictor of adolescent writing quality: Which measures? Which genre? *Reading and Writing*, *22*(2), 185–200. https://doi.org/10.1007/s11145-007-9107-5

Benda, J. (2013). Google Translate in the EFL Classroom: Taboo or Teaching Tool? *Writing and Pedagogy*, *5*(2), 317–332.

Berendt, B., Littlejohn, A., & Blakemore, M. (2020). AI in education: Learner choice and fundamental rights. *Learning, Media and Technology*, *45*(3), 312–324. https://doi.org/10.1080/17439884.2020.1786399

Bitchener, J., & Basturkmen, H. (2006). Perceptions of the difficulties of postgraduate L2 thesis students writing the discussion section. *Journal of English for Academic Purposes*, *5*(1), 4–18.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bond, M., Bedenlier, S., Marín, V. I., & Händel, M. (2021). Emergency remote teaching in higher education: Mapping the first global online semester. *International Journal of Educational Technology in Higher Education*, *18*(1), 50. https://doi.org/10.1186/s41239-021-00282-x

Bose, D., & Khan, P. F. (2020). Artificial Intelligence enabled Smart Learning. *ETH Learning and Teaching Journal*, *2*(2), Article 2.

Boston University Teaching Writing. (2020). *Metacognition | Teaching Writing*. https://www.bu.edu/teaching-writing/resources/metacognition/

Bougie, R., & Sekaran, U. (2019). *Research methods for business: A skill building approach*. John Wiley & Sons.

Bowerman, C. (1992). Writing and the computer: An intelligent tutoring systems solution. *Computer Assisted Learning: Selected Contributions from the CAL'91 Symposium*, 77–83.

Breeze, R. (2008). Researching simplicity and sophistication in student writing. *International Journal of English Studies*, *8*(1), 51–66.

Briggs, N. (2018). Neural Machine Translation Tools in the Language Learning Classroom: Students' Use, Perceptions, and Analyses. *Jalt Call Journal*, *14*(1), 2–24.

Broughton, G., Brumfit, C., Pincas, A., & Wilde, R. D. (2002). *Teaching English as a foreign language*. Routledge.

Bryant, J., Heitz, C., Sanghvi, S., & Wagle, D. (2020). *How artificial intelligence will impact K-12 teachers*. McKinsey and Company.

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, *42*(3), 294–320.

Carbonell, J. R. (1970). AI in CAI: An Artificial-Intelligence Approach to Computer-Assisted Instruction. *IEEE Transactions on Man-Machine Systems*, *11*(4), 190–202. https://doi.org/10.1109/TMMS.1970.299942

Carlon, M. K. J., Gayed, J. M., & Cross, J. S. (2021, December). Development of open-response prompt-based metacognitive tutor for online classrooms. *2021 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*.

Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. *Journal of Second Language Writing*, *44*, 51–62.

Catford, J. C. (1965). *A linguistic theory of translation* (Vol. 31). Oxford University Press London.

Cathoven A.I. (2022). *ADO Language Hub*. CEFR Checker (Version 0.16.0). https://hub.cathoven.com/?scene=analyser&core=cefr

Ceylan, N. O. (2019). Student perceptions of difficulties in second language writing. *Journal of Language and Linguistic Studies*, *15*(1), 151–157.

Chakravorti, B., Chaturvedi, R. S., Filipovic, C., & Brewer, G. (2020). Digital in the time of COVID. *The Fletcher School at Tufts University*, 80.

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access*, *8*, 75264–75278.

Chen, M. X., Lee, B. N., Bansal, G., Cao, Y., Zhang, S., Lu, J., Tsay, J., Wang, Y., Dai, A. M., Chen, Z., Sohn, T., & Wu, Y. (2019). Gmail Smart Compose: Real-Time Assisted Writing. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2287–2295. https://doi.org/10.1145/3292500.3330723

Chen, M.-H., Huang, S.-T., Hsieh, H.-T., Kao, T.-H., & Chang, J. S. (2012). FLOW: a first-language-oriented writing assistant system. *Proceedings of the ACL 2012 System Demonstrations*, 157–162.

Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in Writing: Generating Text in L1 and L2. *Written Communication*, *18*(1), 80–98. https://doi.org/10.1177/0741088301018001004

Chon, Y. V., Shin, D., & Kim, G. E. (2021). Comparing L2 learners' writing against parallel machine-translated texts: Raters' assessment, linguistic complexity and errors. *System*, *96*, 102408.

Chounta, I.-A., Bardone, E., Raudsep, A., & Pedaste, M. (2022). Exploring Teachers' Perceptions of Artificial Intelligence as a Tool to Support their Practice in Estonian K-12 Education. *International Journal of Artificial Intelligence in Education*, *32*(3), 725–755. https://doi.org/10.1007/s40593-021-00243-5

Chun, H., Lee, S., & Park, I. (2021). A systematic review of AI technology use in English education. *Multimedia-Assisted Language Learning*, *24*(1), 87–103.

Cirocki, A., & Caparoso, J. (2016). Attitudes, motivations and beliefs about L2 reading in the Filipino secondary school classroom: A mixed-methods study. *International Journal of Applied Linguistics and English Literature*, *5*(7), 1–18.

Clark, R. C., Nguyen, F., & Sweller, J. (2011). *Efficiency in learning: Evidence-based guidelines to manage cognitive load*. John Wiley & Sons.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Colchester, K., Hagras, H., Alghazzawi, D., & Aldabbagh, G. (2017). A Survey of Artificial Intelligence Techniques Employed for Adaptive Educational Systems within E-

Learning Platforms. *Journal of Artificial Intelligence and Soft Computing Research*, *7*(1), 47–64. https://doi.org/10.1515/jaiscr-2017-0004

Collins, A., & Grignetti, M. C. (1975). *Intelligent CAI.* BOLT BERANEK AND NEWMAN INC CAMBRIDGE MA. https://apps.dtic.mil/sti/citations/ADA016613

Cornish, E. (2004). *Futuring: The Exploration of the Future*. World Future Society.

Crossley, S. A., & McNamara, D. S. (2011). Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing*, *20*(4), 271–285.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135.

Dabarera, C., Renandya, W. A., & Zhang, L. J. (2014). The impact of metacognitive scaffolding and monitoring on reading comprehension. *System*, *42*, 462–473.

Dai, X., Liu, Y., Wang, X., & Liu, B. (2014). Wings: Writing with intelligent guidance and suggestions. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 25–30.

Dale, R., & Viethen, J. (2021). The automated writing assistance landscape in 2021. *Natural Language Engineering*, *27*(4), 511–518. https://doi.org/10.1017/S1351324921000164

Danaher, J. (2018). Toward an Ethics of AI Assistants: An Initial Framework. *Philosophy & Technology*, *31*(4), 629–653. https://doi.org/10.1007/s13347-018-0317-3

D'Angelo, M. C., & Humphreys, K. R. (2015). Tip-of-the-tongue states reoccur because of implicit learning, but resolving them helps. *Cognition*, *142*, 166–190.

Dimmitt, C., & McCormick, C. B. (2012). Metacognition in education. In *APA educational psychology handbook, Vol 1: Theories, constructs, and critical issues.* (pp. 157–187). American Psychological Association.

Dirksen, N., & Takahashi, S. (2020). Artificial Intelligence in Japan 2020. *Netherlands: Netherlands Enterprise Agency*.

Dizon, G., & Gayed, J. M. (2021). Examining the Impact of Grammarly on the Quality of Mobile L2 Writing. *JALT CALL Journal*, *17*(2), 74–92.

Dizon, G., Tang, D., & Yamamoto, Y. (2022). A case study of using Alexa for out-of-class, self-directed Japanese language learning. *Computers and Education: Artificial Intelligence*, *3*, 100088. https://doi.org/10.1016/j.caeai.2022.100088

Dodigovic, M., & Tovmasyan, A. (2021). Automated Writing Evaluation: The Accuracy of Grammarly's Feedback on Form. *International Journal of TESOL Studies*, *3*(2), 71–88.

Dong, Y. R. (1998). Non-native graduate students' thesis/dissertation writing in science: Self-reports by students and their advisors from two US institutions. *English for Specific Purposes*, *17*(4), 369–390.

Dörnyei, Z., & Csizér, K. (2012). How to design and analyze surveys in second language acquisition research. *Research Methods in Second Language Acquisition: A Practical Guide*, *1*, 74–94.

Dörnyei, Z., & Griffee, D. T. (2010). Research Methods in Applied Linguistics. *TESOL Journal*, *1*(1), 181–183. https://doi.org/10.5054/tj.2010.215611

Doroudi, S., & Brunskill, E. (2019). Fairer but not fair enough on the equitability of knowledge tracing. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, 335–339.

Ducar, C., & Schocket, D. H. (2018). Machine translation and the L2 classroom: Pedagogical solutions for making peace with Google translate. *Foreign Language Annals*, *51*(4), 779–795. https://doi.org/10.1111/flan.12366

Ecke, P., & Hall, C. J. (2013). Tracking tip-of-the-tongue states in a multilingual speaker: Evidence of attrition or instability in lexical systems? *International Journal of Bilingualism*, *17*(6), 734–751.

EF EPI 2021 – EF English Proficiency Index – Japan. (n.d.). Retrieved October 14, 2022, from https://www.ef.com/wwen/epi/regions/asia/japan/

Engelbrecht, J., Llinares, S., & Borba, M. C. (2020). Transformation of the mathematics classroom with the internet. *ZDM*, *52*(5), 825–841. https://doi.org/10.1007/s11858-020-01176-4

Englert, C. S., Wu, X., & Zhao, Y. (2005). Cognitive Tools for Writing: Scaffolding the Performance of Students through Technology. *Learning Disabilities Research & Practice*, *20*(3), 184–198. https://doi.org/10.1111/j.1540-5826.2005.00132.x

Eryılmaz, M., Adabashi, A. M., & Yazıcı, A. (2019). *Artificial Intelligence Methods in E-Learning* [Chapter]. Handbook of Research on Faculty Development for Digital Teaching and Learning; IGI Global. https://doi.org/10.4018/978-1-5225-8476-6.ch015

ETS. (2019). TOEFL Writing Rubrics—Educational Testing Service. In *ETS Independent Writing Rubric*. https://www.ets.org/s/toefl/pdf/toefl-writing-rubrics.pdf

Evmenova, A. S., Graff, H. J., Jerome, M. K., & Behrmann, M. M. (2010). Word Prediction Programs with Phonetic Spelling Support: Performance Comparisons and Impact on Journal Writing for Students with Writing Difficulties. *Learning Disabilities Research & Practice*, *25*(4), 170–182. https://doi.org/10.1111/j.1540-5826.2010.00315.x

Ezzy, D. (2013). *Qualitative analysis*. Routledge.

Farrokhi, F., & Mahmoudi, A. (2012). Rethinking Convenience Sampling: Defining Quality Criteria. *Theory and Practice in Language Studies*, *2*. https://doi.org/10.4304/tpls.2.4.784-792

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, *5*(1), 80–92.

Ferri, F., Grifoni, P., & Guzzo, T. (2020). Online Learning and Emergency Remote Teaching: Opportunities and Challenges in Emergency Situations. *Societies*, *10*(4), Article 4. https://doi.org/10.3390/soc10040086

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. SAGE.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906.

Fowler, H. N., Lamb, W. R. M., Bury, R. G., & others. (1925). *Plato* (Vol. 5). Harvard University Press.

Frankenberg-Garcia, A. (2020). Combining user needs, lexicographic data and digital writing environments. *Language Teaching*, *53*(1), 29–43.

Fujii, H., & Managi, S. (2018). Trends and priority shifts in artificial intelligence technology invention: A global patent analysis. *Economic Analysis and Policy*, *58*, 60–69. https://doi.org/10.1016/j.eap.2017.12.006

Fung, W. J. (2010). A predictive text completion software in Python. *The Python Paper s Monograph*, *2*.

Furey, H., & Martin, F. (2019). AI education matters: A modular approach to AI ethics education. *AI Matters*, *4*(4), 13–15. https://doi.org/10.1145/3299758.3299764

Gama, C. (2004). Metacognition in interactive learning environments: The Reflection Assistant model. *International Conference on Intelligent Tutoring Systems*, 668–677.

Gamper, J., & Knapp, J. (2002). A review of intelligent CALL systems. *Computer Assisted Language Learning*, *15*(4), 329–342.

Garcia, G. (2019). Artificial Intelligence in Japan: Industrial Cooperation and Business Opportunities for European Companies. *EU-Japan Centre for Industrial Cooperation*.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M., & Zettlemoyer, L. (2018). *AllenNLP: A Deep Semantic Natural Language Processing Platform* (arXiv:1803.07640). arXiv. https://doi.org/10.48550/arXiv.1803.07640

Gartner Identifies Four Trends Driving Near-Term Artificial Intelligence Innovation. (2021). Gartner. https://www.gartner.com/en/newsroom/press-releases/2021-09-07-gartner-identifies-four-trends-driving-near-term-artificial-intelligence-innovation

Gayed, J. M., Carlon, M. K. J., & Cross, J. S. (2022). The Matthew effect in CALL: Examining the equity of a novel intelligent writing assistant as English language support. *Proceedings of the XXIst International CALL Research Conference*, 80–93. https://doi.org/10.29140/9781914291050-12

Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing Assistant's impact on English language learners. *Computers and Education: Artificial Intelligence*, *3*, 100055.

Given, L. M. (2008). *The SAGE Encyclopedia of Qualitative Research Methods*. SAGE Publications.

Gnacek, M., Doran, E., Bommer, S., & Appiah-Kubi, P. (2020). The effectiveness of smart compose: An artificial intelligent system. *Journal of Management & Engineering Integration*, *13*(1), 111–121.

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Lang. Learn. Technol*, *26*, 5–24.

Gurcan, F., Ozyurt, O., & Cagitay, N. E. (2021). Investigation of emerging trends in the e-learning field using latent Dirichlet allocation. *International Review of Research in Open and Distributed Learning*, *22*(2), 1–18.

Hamouma, C., & Menezla, N. (2019). The Impact of Digital Literacy Proficiency on EFL Students' Academic Writing Performance: A Case Study of Algerian Third Year EFL Students. *International Journal of Digital Literacy and Digital Competence (IJDLDC)*, *10*(4), 40–55. https://doi.org/10.4018/IJDLDC.2019100103

Harsch, C., & Kanistra, V. P. (2020). Using an Innovative Standard-setting Approach to Align Integrated and Independent Writing Tasks to the CEFR. *Language Assessment Quarterly*, *17*(3), 262–281. https://doi.org/10.1080/15434303.2020.1754828

Hartley, J., & Tynjälä, P. (2001). New technology, writing and learning. In *Writing as a learning tool* (pp. 161–182). Springer.

Hayashi, H., Sasajima, H., Takayanagi, Y., & Kanamaru, H. (2017). International standardization for smarter society in the field of measurement, control and automation. *2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, 263–266. https://doi.org/10.23919/SICE.2017.8105723

Hecker, U., Dutke, S., & Sedek, G. (2000). *Generative Mental Processes and Cognitive Resources: Integrative Research on Adaptation and Control*. Springer Science & Business Media.

Hoadley, D. S., & Lucas, N. J. (2018). *Artificial Intelligence and National Security*. 42.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *7*(2), 174–196.

Horowitz, M. C., Allen, G. C., Kania, E. B., & Scharre, P. (2018). *Strategic competition in an era of artificial intelligence*. Center for a New American Security.

Howitt, D., & Cramer, D. (2020). *Research methods in psychology*. Pearson.

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288.

Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. In *Computers and Education: Artificial Intelligence* (Vol. 1, p. 100001). Elsevier.

Ishii, E., Ebner, D. K., Kimura, S., Agha-Mir-Salim, L., Uchimido, R., & Celi, L. A. (2020). The advent of medical artificial intelligence: Lessons from the Japanese approach. *Journal of Intensive Care*, *8*(1), 1–6.

ISO - 01.110—Technical product documentation. (n.d.). Retrieved August 17, 2022, from https://www.iso.org/ics/01.110/x/p/1/u/0/w/0/d/0

Javadi-Safa, A. (2018). A brief overview of key issues in second language writing teaching and research. *International Journal of Education and Literacy Studies*, *6*(2), 12–25.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working Papers/Lund University, Department of Linguistics and Phonetics*, *53*, 61-79-61–79.

Johnson, A. M., Jacovina, M. E., Russell, D. G., & Soto, C. M. (2016). Challenges and Solutions when Using Technologies in the Classroom. In *Adaptive Educational Technologies for Literacy Instruction*. Routledge.

Kang, G. Y. (2018). Playing With Digital Tools With Explicit Scaffolding. *The Reading Teacher*, *71*(6), 735–741. https://doi.org/10.1002/trtr.1672

Kawauchi, C., & Kamimoto, T. (2000). Distinctive Features of Oral Production by Fluent and Nonfluent EFL Learners. *Language Laboratory*, *37*, 21–36. https://doi.org/10.24539/llaj.37.0_21

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, *1*(1), 1–26. https://doi.org/10.17239/jowr-2008.01.01.1.

Kessler, M. (2020). Technology-Mediated Writing: Exploring Incoming Graduate Students' L2 Writing Strategies with Activity Theory. *Computers and Composition*, *55*, 102542. https://doi.org/10.1016/j.compcom.2020.102542

Kirschenbaum, M. G. (2017). Track Changes: A Literary History of Word Processing. In *Track Changes*. Harvard University Press. https://doi.org/10.4159/9780674969469

Kılıçkaya, F. (2022). Pre-service language teachers' online written corrective feedback preferences and timing of feedback in computer-supported L2 grammar instruction. *Computer Assisted Language Learning*, *35*(1–2), 62–87. https://doi.org/10.1080/09588221.2019.1668811

Knospe, Y. (2018). Metacognitive knowledge about writing in a foreign language: A case study. In *Metacognition in language learning and teaching* (pp. 121–138). Routledge.

Koenig, E., Guertler, K., Żarnowska, D., & Horbačauskienė, J. (2020). Developing English language competence for global engineers. *2020 IEEE Global Engineering Education Conference (EDUCON)*, 242–249.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens. *Vocabulary Learning and Instruction*, *1*(1), 60–69.

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing*, *44*, 100450.

Krashen, S. D. (2003). *Explorations in language acquisition and use*. Heinemann Portsmouth, NH.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Lamb, M. (2011). A Matthew Effect in English language education in a developing country context. In *Dreams and Realities: Developing Countries and the English Language* (pp. 186–206). The British Council.

Landauer, T. K. (2003). Automatic Essay Assessment. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 295–308. https://doi.org/10.1080/0969594032000148154

Latham, A., & Goltz, S. (2019). A Survey of the General Public's Views on the Ethics of Using AI in Education. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), *Artificial Intelligence in Education* (pp. 194–206). Springer International Publishing. https://doi.org/10.1007/978-3-030-23204-7_17

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, *25*(2), 21–33.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322.

Lee, J., & Lee, S. (2020). A Study on Experts' Perception Survey on Elementary AI Education Platform. *Journal of The Korean Association of Information Education*, *24*(5), 483–494. https://doi.org/10.14352/jkaie.2020.24.5.483

Leontjev, D., Huhta, A., & Mäntylä, K. (2016). Word derivational knowledge and writing proficiency: How do they link? *System*, *59*, 73–89. https://doi.org/10.1016/j.system.2016.03.013

Liu, Y. (2020). Relating lexical access and second language speaking performance. *Languages*, *5*(2), 13.

Llach, M. del P. A. (2011). *Lexical errors and accuracy in foreign language writing*. Multilingual Matters.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208.

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, *29*, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003

Maringe, F., & Foskett, N. (2012). *Globalization and internationalization in higher education: Theoretical, strategic and management perspectives*. A&C Black.

Marr, B. (2018). How is AI used in education–Real world examples of today and a peek into the future. *Forbes Magazine*, *25*.

Marra, R. M., Kim, S. M., Plumb, C., Hacker, D. J., & Bossaller, S. (2017). Beyond the technical: Developing lifelong learning and metacognition for the engineering workplace. *2017 ASEE Annual Conference & Exposition*.

McCarthy, K. S., Roscoe, R. D., Allen, L. K., Likens, A. D., & McNamara, D. S. (2022). Automated writing evaluation: Does spelling and grammar feedback support high-quality writing and revision? *Assessing Writing*, *52*, 100608.

McCarthy, P. M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)* [PhD Thesis]. The University of Memphis.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392.

McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing*, *15*(2), 118–129. https://doi.org/10.1016/j.asw.2010.04.002

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, *32*(1), 12–16.

Messer, D., & Nash, G. (2018). An evaluation of the effectiveness of a computer-assisted reading intervention. *Journal of Research in Reading*, *41*(1), 140–158.

mimno. (2022). *JsLDA* [JavaScript]. https://github.com/mimno/jsLDA (Original work published 2013)

Mizumoto, A. (2015). *Langtest (Version 1.0)[Web application]. Kansai University*. https://langtest.jp/#app

Mizumoto, A. (2022). Calculating the Relative Importance of Multiple Regression Predictor Variables Using Dominance Analysis and Random Forests. *Language Learning*. https://doi.org/10.1111/lang.12518

Mizumoto, A., Hamatani, S., & Imao, Y. (2017). Applying the Bundle–Move Connection Approach to the Development of an Online Writing Support Tool for Research Articles. *Language Learning*, *67*(4), 885–921. https://doi.org/10.1111/lang.12250

Mizumoto, A., & Plonsky, L. (2016). R as a Lingua Franca: Advantages of Using R for Quantitative Research in Applied Linguistics. *Applied Linguistics*, *37*(2), 284–291. https://doi.org/10.1093/applin/amv025

Momtazi, S., & Naumann, F. (2013). Topic modeling for expert finding using latent Dirichlet allocation. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(5), 346–353.

Monett, D., & Lewis, C. W. P. (2018). Getting Clarity by Defining Artificial Intelligence—A Survey. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence 2017* (pp. 212–214). Springer International Publishing. https://doi.org/10.1007/978-3-319-96448-5_21

Mook, D. G. (1983). In Defense of External Invalidity. *American Psychologist*.

Moore, K. A., Rutherford, C., & Crawford, K. A. (2019). Supporting postsecondary English language learners' writing proficiency using technological tools. *Journal of International Students, 2016 Vol. 6 (4)*, *6*(4), 857–872. https://doi.org/10.32674/jis.v6i4.321

Moranski, K., & Henery, A. (2017). Helping Learners to Orient to the Inverted or Flipped Language Classroom: Mediation via Informational Video. *Foreign Language Annals*, *50*(2), 285–305. https://doi.org/10.1111/flan.12262

Moranski, K., & Ziegler, N. (2021). A Case for Multisite Second Language Acquisition Research: Challenges, Risks, and Rewards. *Language Learning*, *71*(1), 204–242. https://doi.org/10.1111/lang.12434

Muñoz Martín, R., & Cardona Guerra, J. M. (2019). Translating in fits and starts: Pause thresholds and roles in the research of translation processes. *Perspectives*, *27*(4), 525–551. https://doi.org/10.1080/0907676X.2018.1531897

Nation, P. (2014). Developing Fluency. In T. Muller, J. Adamson, P. S. Brown, & S. Herder (Eds.), *Exploring EFL Fluency in Asia* (pp. 11–25). Palgrave Macmillan UK. https://doi.org/10.1057/9781137449405_2

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

Ngiam, L. C. W., & See, S. L. (2017). Language e-Learning and Music Appreciation. In *Advances in Human Factors, Business Management, Training and Education* (pp. 865–877). Springer.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 249–256.

Niess, M. L. (2011). Investigating TPACK: Knowledge Growth in Teaching with Technology. *Journal of Educational Computing Research*, *44*(3), 299–317. https://doi.org/10.2190/EC.44.3.c

Nunan, D., & Carter, R. (2001). *The Cambridge guide to teaching English to speakers of other languages*. Ernst Klett Sprachen.

O'Brien, C. (2020). AI startups raised $18.5 billion in 2019, setting new funding record. *Venture Beat*.

Ocaña-Fernández, Y., Valenzuela-Fernández, L. A., & Garro-Aburto, L. L. (2019). Artificial Intelligence and Its Implications in Higher Education. *Journal of Educational Psychology-Propositos y Representaciones*, *7*(2), 553–568.

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, *19*, 1609406919899220.

Oh, S. (2020). Second language learners' use of writing resources in writing assessment. *Language Assessment Quarterly*, *17*(1), 60–84.

Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning*, *13*(2), 179–212.

O'Regan, B., Mompean, A. R., & Desmet, P. (2010). From spell, grammar and style checkers to writing aids for English and French as a foreign language: Challenges and opportunities. *Revue Francaise de Linguistique Appliquee*, *15*(2), 67–84.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429.

Palviainen, Å., Kalaja, P., & Mäntylä, K. (2012). Development of L2 writing: Fluency and proficiency. *AFinLA-e: Soveltavan Kielitieteen Tutkimuksia*, *4*, 47–59.

Parisis, N. (2019). Medical writing in the era of artificial intelligence. *Medical Writing*, *28*, 4–9.

Park, J. (2019). An AI-based English grammar checker vs. Human raters in evaluating EFL learners' writing. *Multimedia-Assisted Language Learning*, *22*(1), 112–131.

Parker, L. E. (2018). Creation of the national artificial intelligence research and development strategic plan. *AI Magazine*, *39*(2), 25–32.

Patterson, N. (2007). The Devil in the Machine: Problems with Computerized Writing Assessment. *Language Arts Journal of Michigan*, *23*(1). https://doi.org/10.9707/2168-149X.1143

Patton, M. Q. (2002). Two Decades of Developments in Qualitative Inquiry: A Personal, Experiential Perspective. *Qualitative Social Work*, *1*(3), 261–283. https://doi.org/10.1177/1473325002001003636

Penno, J. F., Wilkinson, I. A., & Moore, D. W. (2002). Vocabulary acquisition from teacher explanation and repeated listening to stories: Do they overcome the Matthew effect? *Journal of Educational Psychology*, *94*(1), 23.

Perry, F. (2021). The use of embedded digital tools to develop English language proficiency in higher education. *Journal of Academic Language and Learning*, *15*(1), Article 1.

Peterson, J. L. (1980). Computer programs for detecting and correcting spelling errors. *Communications of the ACM*, *23*(12), 676–687. https://doi.org/10.1145/359038.359041

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912.

Polio, C. G. (2001). SECOND LANGUAGE DEVELOPMENT IN WRITING: MEASURES OF FLUENCY, ACCURACY, AND COMPLEXITY. Pp. 187. *Studies in Second Language Acquisition*, *23*(3), 423–425.

Popenici, S. A., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, *12*(1), 1–13.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2000). Comparing the Validity of Automated and Human Essay Scoring. *ETS Research Report Series*, *2000*(2), i–23. https://doi.org/10.1002/j.2333-8504.2000.tb01833.x

Purcell, K., Buchanan, J., & Friedrich, L. (2013). The impact of digital tools on student writing and how writing is taught in schools. *Washington, DC: Pew Research Center*, *16*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*.

Raman, A., & Rathakrishnan, M. (2019). *Redesigning higher education initiatives for Industry 4.0*. IGI Global.

Reich, J. (2020). Two stances, three genres, and four intractable dilemmas for the future of learning at scale. *Proceedings of the Seventh ACM Conference on Learning@ Scale*, 3–13.

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research*, *43*(1), 73–97.

Rice, P. L., & Ezzy, D. (1999). Qualitative research methods: A health focus. *Melbourne, Australia*.

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, *26*(2), 582–599.

Saito, K., & Akiyama, Y. (2017). Linguistic correlates of comprehensibility in second language Japanese speech. *Journal of Second Language Pronunciation*, *3*(2), 199–217. https://doi.org/10.1075/jslp.3.2.02sai

Salamoura, A., & Saville, N. (2010). Exemplifying the CEFR: Criterial features of written learner English from the English Profile Programme. *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research*, *1*, 101–132.

Sato, M., & Dussuel Lam, C. (2021). Metacognitive instruction with young learners: A case of willingness to communicate, L2 use, and metacognition of oral communication. *Language Teaching Research*, 13621688211004640.

Sawa, T. (2019, October 16). *Reforming education for Society 5.0*. The Japan Times. https://www.japantimes.co.jp/opinion/2019/10/16/commentary/japan-commentary/reforming-education-society-5-0/

Scardamalia, M., & Bereiter, C. (1987). 4 Knowledge telling and knowledge transforming. *Advances in Applied Psycholinguistics: Volume 2, Reading, Writing, and Language Learning*, *2*, 142.

Schmalz, V. J., & Brutti, A. (2021). Automatic Assessment of English CEFR Levels Using BERT Embeddings. *Http://Ceur-Ws. Org/Vol-3033/*, *3033*.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1–30.

Schoonen, R., Snellings, P., Stevenson, M., & Van Gelderen, A. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. *Writing in Foreign Language Contexts: Learning, Teaching, and Research*, 77–101.

Sevcikova, B. L. (2018). Online Open-Source Writing Aid as a Pedagogical Tool. *English Language Teaching*, *11*(8), 126–142.

Sheehan, S. (2013). *British Council ELT Research Papers* (Vol. 1). British Council.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. *International Encyclopedia of Education*, *4*(1), 20–26.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. *TESOL Quarterly*, *27*(4), 657–677.

Southgate, E., Blackmore, K., Pieschl, S., Grimes, S., Smithers, K., & McGuire, J. (2019). *Artificial Intelligence and Emerging Technologies in Schools Research Report*.

Stanovich, K. E. (2009). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Journal of Education*, *189*(1–2), 23–55.

Stasenko, A., & Gollan, T. H. (2019). Tip of the tongue after any language: Reintroducing the notion of blocked retrieval. *Cognition*, *193*, 104027. https://doi.org/10.1016/j.cognition.2019.104027

Strahm, B., Gray, C. M., & Vorvoreanu, M. (2018). Generating mobile application onboarding insights through minimalist instruction. *Proceedings of the 2018 Designing Interactive Systems Conference*, 361–372.

Subrahmanyam, V., & Swathi, K. (2018). Artificial intelligence and its implications in education teaching and learning in higher education. *Research and Practice in Technology*.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296.

Taguma, M., Feron, E., & Lim, M. H. (2018). Future of education and skills 2030: Conceptual learning framework. *Organization of Economic Co-Operation and Development*.

Tan, B.-H. (2011). Innovating writing centers and online writing labs outside North America. *The Asian EFL Journal Quarterly*, *13*(2), 390–417.

Tang, K.-Y., Chang, C.-Y., & Hwang, G.-J. (2021). Trends in artificial intelligence-supported e-learning: A systematic review and co-citation network analysis (1998–2019). *Interactive Learning Environments*, 1–19.

Tanil, C. T., & Yong, M. H. (2020). Mobile phones: The effect of its presence on learning and memory. *PloS One*, *15*(8), e0219233.

Tardy, C. M., & Matsuda, P. K. (2009). The construction of author voice by editorial board members. In *Written Communication* (Vol. 26, Issue 1, pp. 32–52). Sage Publications Sage CA: Los Angeles, CA.

Text Inspector: Analyse the Difficulty Level of English Texts | Text inspector. (n.d.). Retrieved August 16, 2022, from https://textinspector.com/

Tight, D. G. (2017). *Tool usage and effectiveness among L2 Spanish computer writers*. https://doi.org/10.12795/elia.2017.i17.07

Tilak, G. (2020). *Artificial intelligence: A Better and innovative technology for enhancement and sustainable evolution in education system*.

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), 302–327.

Tsai, S.-C. (2019). Using google translate in EFL drafts: A preliminary investigation. *Computer Assisted Language Learning*, *32*(5–6), 510–526.

Urlaub, P., & Dessein, E. (2022). From Disrupted Classrooms to Human-Machine Collaboration? The Pocket Calculator, Google Translate, and the Future of Language Education. *L2 Journal*, *14*(1). https://doi.org/10.5070/L214151790

Vincent-Lancrin, S., & van der Vlies, R. (2020). *Trustworthy artificial intelligence (AI) in education: Promises and challenges*.

Vitta, J. P., & Al-Hoorie, A. H. (2021). Measurement and sampling recommendations for L2 flipped learning experiments: A bottom-up methodological synthesis. *Journal of Asia TEFL*, *18*(2), 682.

Voskoglou, M. G. (2019). Artificial intelligence as a tool in the modern education. *International Journal of Applications of Fuzzy Sets and Artificial Intelligence*, *9*, 125–138.

Waldron, S., Wood, C., & Kemp, N. (2017). Use of predictive text in text messaging over the course of a year and its relationship with spelling, orthographic processing and grammar. *Journal of Research in Reading*, *40*(4), 384–402. https://doi.org/10.1111/1467-9817.12073

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1–37.

Wang, Y. (2019). Effects of L1/L2 Captioned TV Programs on Students' Vocabulary Learning and Comprehension. *CALICO Journal*, *36*, 204–224.

Wei, X. (2020). Assessing the metacognitive awareness relevant to L1-to-L2 rhetorical transfer in L2 writing: The cases of Chinese EFL writers across proficiency levels. *Assessing Writing*, *44*, 100452. https://doi.org/10.1016/j.asw.2020.100452

Weissberg, R. (2006). *Connecting speaking & writing in second language writing instruction* (Issue Sirsi) i9780472030323).

Wheeler, E., & McDonald, R. L. (2000). Writing in engineering courses. *Journal of Engineering Education*, *89*(4), 481–486.

Why do Japanese have trouble learning English? | The Japan Times. (2017). The Japan Times. https://www.japantimes.co.jp/opinion/2017/10/29/commentary/japan-commentary/japanese-trouble-learning-english/

Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. In *Learning, Media and Technology* (Vol. 45, Issue 3, pp. 223–235). Taylor & Francis.

Wiseman, C. S. (2012). A comparison of the performance of analytic vs. Holistic scoring rubrics to assess L2 writing. *International Journal of Language Testing*, *2*(1), 59–92.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity* (Issue 17). University of Hawaii Press.

Wolfersberger, M. (2003). L1 to L2 writing process and strategy transfer: A look at lower proficiency writers. *TESL-EJ*, *7*(2), 1–12.

Wollowski, M., Selkowitz, R., Brown, L., Goel, A., Luger, G., Marshall, J., Neel, A., Neller, T., & Norvig, P. (2016). A Survey of Current Practice and Teaching of AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1), Article 1. https://doi.org/10.1609/aaai.v30i1.9857

Xiao, G., & Chen, X. (2015). English academic writing difficulties of engineering students at the tertiary level in China. *World Transactions on Engineering and Technology Education*, *13*(3), 259–263.

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. *PLOS ONE*, *15*(9), e0239441. https://doi.org/10.1371/journal.pone.0239441

Yadav, A., Ocak, C., & Oliver, A. (2022). Computational Thinking and Metacognition. *TechTrends*, 1–7.

Zaker, A. (2015). EFL Learners' Language Learning Strategies and Autonomous Learning: Which One Is a Better Predictor of L2 Skills? *Journal of Applied Linguistics-Dubai*, *1*, 27–39.

Zeng, Y., Lu, E., & Huangfu, C. (2018). Linking artificial intelligence principles. *ArXiv Preprint ArXiv:1812.04814*.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*, *2021*, e8812542. https://doi.org/10.1155/2021/8812542

Zhang, L. J., & Qin, T. L. (2018). Validating a questionnaire on EFL writers' metacognitive awareness of writing strategies in multimedia environments. In *Metacognition in language learning and teaching* (pp. 157–178). Routledge.

Zhang, Z. V. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, *43*, 100439.

Zhao, C. G., & Liao, L. (2021). Metacognitive strategy use in L2 writing assessment. *System*, *98*, 102472.

Zheng, B., & Warschauer, M. (2017). Epilogue: Second language writing in the age of computer-mediated communication. *Journal of Second Language Writing*, *36*, 61–67. https://doi.org/10.1016/j.jslw.2017.05.014

Zhu, W. (2004). Faculty views on the importance of writing, the nature of academic writing, and teaching and responding to writing in the disciplines. *Journal of Second Language Writing*, *13*(1), 29–48.