John Maurice Gayed, May Kristine Jonson Carlon, Jeffrey Scott Cross

Tokyo Institute of Technology, Tokyo, Japan

johnmaurice.gayed@gmail.com, maykristine.jonson@gmail.com, cross.j.aa@m.titech.ac.jp

The Matthew effect in CALL: Examining the equity of a novel intelligent writing assistant as English language support

Bio data

John is a lecturer at the Institute for International Relations teaching students from the School of Engineering at the University of Hyogo. As a Ph.D. student at the Tokyo Institute of Technology (Tokyo Tech), he has been working on AI KAKU, an AI-based writing assistant for English language learners. His other research interests include defining ethical frameworks for the use of AI in education.

May is a postdoctoral researcher at Tokyo Tech, a software engineer at a medium-scale company in Tokyo with a subsidiary in the Philippines, and an Informatics lecturer at Hosei University also in Tokyo, Japan. Her research revolves around improving online learning experiences, may it be through technology-assisted instruction or learning analytics for course quality assurance.

Jeffrey is a professor in Tokyo Tech's School of Environment and Society and the founder of Tokyo Tech's Online Education Development Office (OEDO) – the office develops MOOCs on edX. OEDO also supports Tokyo Tech faculty through the development of online course content, media production, teaching assistant training, and online course learning analytics. In his lab, he supervises students conducting research in educational technology, machine learning, AI in education, and virtual reality.

Abstract

As 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 this study, we leverage natural language processing (NLP) based transformers to introduce English language support to English as a Foreign Language (EFL) learners while they are in the writing process. A web-based application was created that uses next-word prediction and automatic reverse translation to help EFL participants in their writing. Adult English language learners from professional development language schools participated in a counterbalanced repeated measures study. To understand the presence of the Matthew effect, learners were grouped based on their self-reported EIKEN scores. Their performance according to two 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. This research sets the stage for understanding how emerging tools can support learning without exacerbating Matthew effects. These effects should be considered in both the development and application of educational technology.

Conference paper

Introduction

The roots of AI in education (AIED) can be traced back to the early 1970s. One of the first initiatives of using AI in the field of education was demonstrated via an intelligent teaching platform called "SCHOLAR CAI" in the United States (Collins & Grignetti, 1975). Since then, the rapid progress of AI technologies has seen many developers and institutions implement these systems with the ultimate goal of making learning and teaching more efficient and effective (Roll & Wylie, 2016).

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 spell check 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 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.

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 researchers have 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 researchers explored the equity of using the tool with students with different English skill levels.

Research Questions

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 the 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.

Related Works

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 was 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.

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 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.

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.'s (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 the 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.

Methodology

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 next-generation 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, as seen in Figure 1, 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.

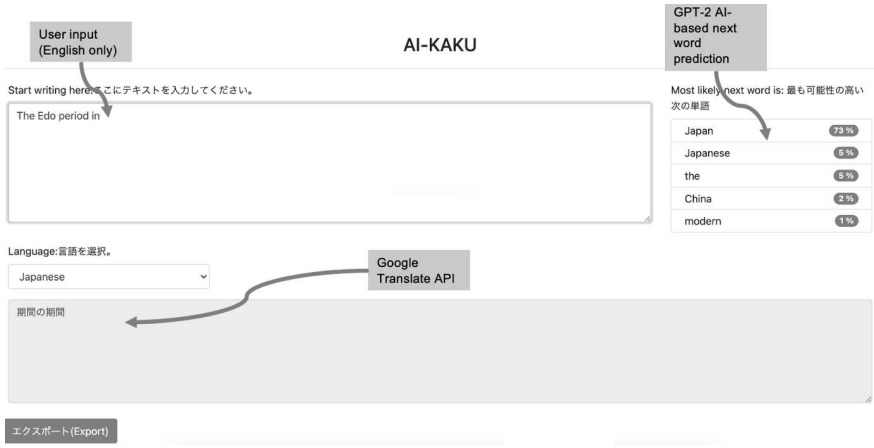

Figure 1. *AI KAKU's interface*

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.

Experimental Design

The researchers 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.

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 Figure 2.

Group A	Group B	Group C	Group D
[T] Topic 1	[C] Topic 3	[T] Topic 3	[C] Topic 1
Likert Survey			
[C] Topic 2	[T] Topic 4	[C] Topic 4	[T] Topic 2
Likert Survey			
[T] Topic 3	[C] Topic 1	[T] Topic 1	[C] Topic 3
Likert Survey			
[C] Topic 4	[T] Topic 2	[C] Topic 2	[T] Topic 4
Likert Survey			

Figure 2. *Experiment design. T = Treatment, C = Control*

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 researchers chose the Independent Writing Task from the test and all of the questions chosen asked the writer their opinion on commonly discussed social topics. By choosing a standardized test source for our writing prompts, the researchers could avoid weighted difficulty differences between writing prompts. In other words, all of 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 researchers 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 thus their ability to turn ideas into written text also improves (Palviainen et al., 2012).

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 researchers' goals while developing AI KAKU was to reduce the problem space for L2 writers, measuring intrinsic load gives the researchers a more granular look into the ability of AI KAKU to address that cognitive burden.

Results and Discussion

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 1 shows the breakdown of the respondents according to group assignment, gender, and reported EIKEN levels.

Table 1. *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%

Impact of Treatment

Lexical Measures

A paired t-test was used to examine the difference between the control and treatment writing conditions. As seen in Table 2, 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 3, 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 2. *Lexical differences between writing conditions. Mean and SD values are shown in ()*

	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

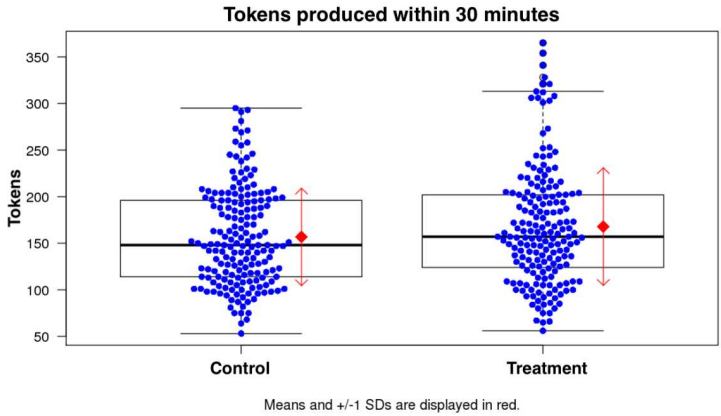

Figure 3. *Scatterplot of Token production under each condition*

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 researchers needed to confirm the reliability of the questions used in this study. The value for Cronbach Alpha for the survey items was $\alpha = 0.57$, which can be interpreted as "acceptable" according to Taber's (2018) meta-analysis of Alpha reliability measures. Results summarized in Table 3 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 4) 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 3. *Cognitive and intrinsic load differences. Mean and SD values are shown in ().*

	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.

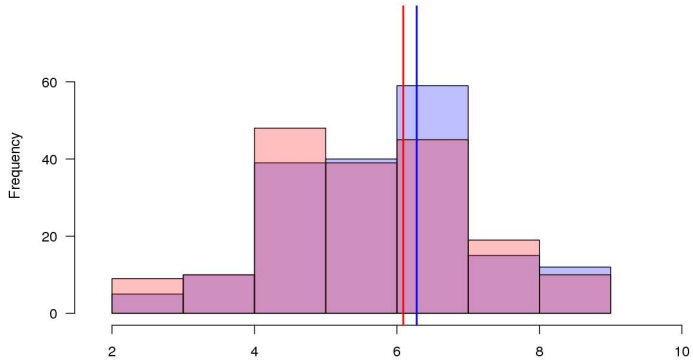

Figure 4. *Impact of control and treatment on intrinsic load*

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 researchers to approach the second research question regarding evidence of the Matthew effect and how the writing assistant impacted participants at different skill levels.

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 researchers. The box plots in Figure 5 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.

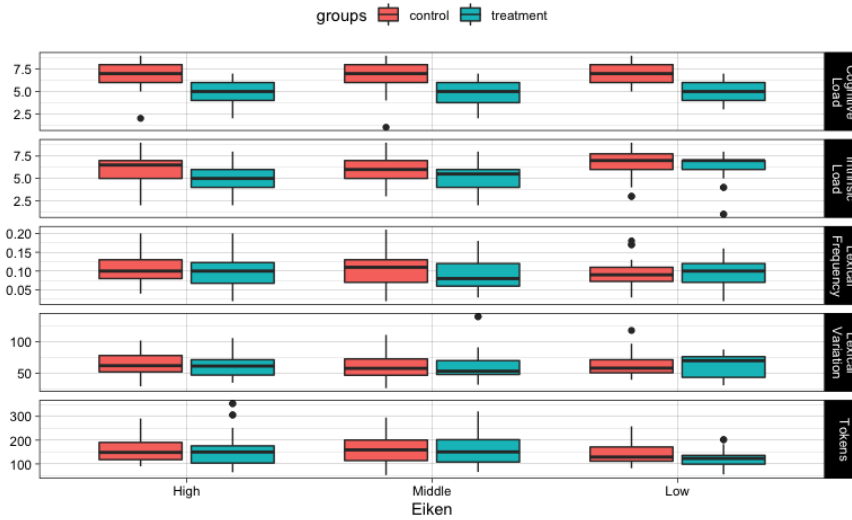

Figure 5. *Performance per cluster*

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 2) between control and treatment conditions (EIKEN levels are disregarded here), the researchers feel 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 researchers 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 6. 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.

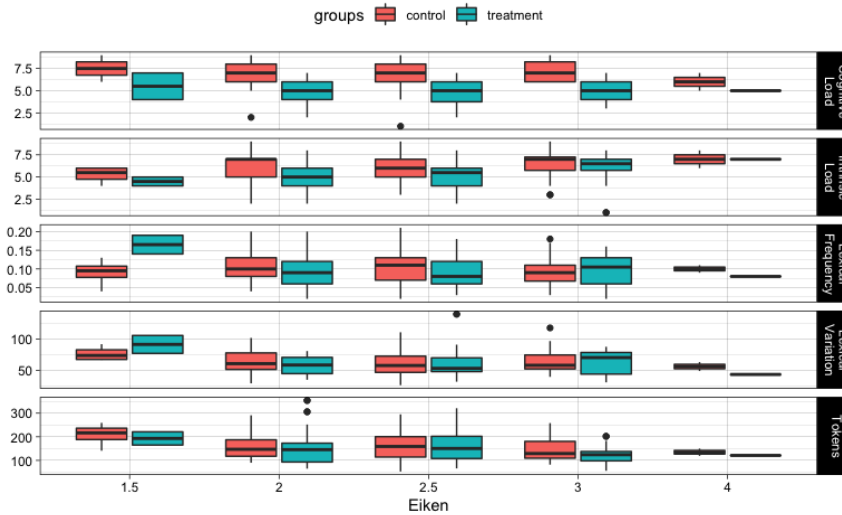

Figure 6. Performance across all EIKEN levels

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 researchers believe 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 researchers could see some effects regarding the cognitive load, lexical frequency, and lexical density. Lower-level users' 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 researchers recommend 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.

Acknowledgements

This work is supported by the Japan Society for the Promotion of Science (JSPS) via the Grants-in-Aid for Scientific Research (Kakenhi) Grant Number 22K00718 and 20H01719.

References

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.

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

Ashton, T. M. (1999). Spell checking: Making writing meaningful in the inclusive classroom. *Teaching Exceptional Children*, *32*(2), 24–27.

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

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.

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.

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

Collins, A., & Grignetti, M. C. (1975). *Intelligent CAI.* Bolt Beranek and Newman Inc, Cambridge, MA.

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.

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.

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.

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.

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

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.
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.
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.
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.
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.
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.
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.
Nawal, A. F. (2018). Cognitive load theory in the context of second language academic writing. *Higher Education Pedagogies*, *3*(1), 385–402.
Ngiam, L. C. W., & See, S. L. (2017). Language e-Learning and music appreciation. In J. I. Kantola, T. Barath, S. Nazir, & T. Andre (Eds.), *Advances in human factors, business management, training and education* (pp. 865–877). Springer.
Oh, S. (2020). Second language learners' use of writing resources in writing assessment. *Language Assessment Quarterly*, *17*(1), 60–84.
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.
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.
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.
Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912.
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.
Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, *26*(2), 582–599.
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.

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.
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.
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.
Tan, B.-H. (2011). Innovating writing centers and online writing labs outside North America. *The Asian EFL Journal Quarterly*, *13*(2), 390–417.
Tanil, C. T., & Yong, M. H. (2020). Mobile phones: The effect of its presence on learning and memory. *PloS One*, *15*(8), e0219233.
Wang, Y. (2019). Effects of L1/L2 captioned TV programs on students' vocabulary learning and comprehension. *CALICO Journal*, *36*, 204–224.
Zhang, Z. V. (2020). Engaging with automated writing evaluation (AWE) feedback on L2 writing: Student perceptions and revisions. *Assessing Writing*, *43*, 100439.