Impact on Second Language Writing via an Intelligent Writing Assistant and Metacognitive Training

John Maurice Gayed
School of Environment and Society
Tokyo Institute of Technology
Meguro-ku, Tokyo, Japan
johnmaurice.gayed@gmail.com

May Kristine Jonson Carlon
School of Environment and Society
Tokyo Institute of Technology
Meguro-ku, Tokyo, Japan
maykristine.jonson@gmail.com

Jeffrey Scott Cross
School of Environment and Society
Tokyo Institute of Technology
Meguro-ku, Tokyo, Japan
cross.j.aa@m.titech.ac.jp

***Abstract*—This Research to Practice Full Paper 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 English as a foreign language (EFL) learners. This study employs a randomized control trial with university-level participants using an internally developed writing aid with next-word prediction, reverse translation support, and metacognitive prompts. 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's implications can be applied not only to EFL learners but also to engineering-related fields using English as a lingua franca.**

***Index Terms*—Writing, Metacognition, Educational software, Experimental research**

I. 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 [1]. Notably, Zhu [2] 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 [3] 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 [4], idea organization, and appropriate vocabulary use [5]. Examples such as Xiao and Chen's [6] 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 [7]. 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 [8]. 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 available 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.

II. Theoretical Framework

Writing ability is commonly seen as an indicator of language acquisition progression and proficiency [9]. Gaining writing proficiency has numerous benefits. As Basturkmen

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 JP20H01719.

and Lewis [10] 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 [11].

EFL learners often struggle in the writing process [12] and lack some of the formulating strategies [13] 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 [14]. 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; thus, it is more likely for the person to struggle again when they need to use the same word in the future [15]. This is an unproductive cognitive load for EFL learners practicing to write 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 [16]. 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 [17] and thus not succumb to over-reliance on assistance.

The digital writing assistant (AI KAKU) developed for this study [18] 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) users but do little to assist the L2 users struggling with language production. The researchers, therefore, attempt 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 researchers predict 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:

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

III. Literature Review

A. Intelligent writing assistants

Intelligent writing agents within the field of Computer Assisted Language Learning (CALL) are not new (see Bowerman's [19] 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's [20] scoping review of intelligent CALL applications identified 19 systems that aim to support L2 writers. The systems identified in the researcher's 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. [21] 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. [22] use NLP techniques to develop an application called "FLOW," which is intended to assist ELL 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 [23] 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: automatic writing feedback, text prediction, and post-writing evaluation, all supportive technologies for the L2 user.

B. Metacognition in engineering and language acquisition

Metacognition is concretely seen as the knowledge and regulation of ones' cognitive abilities [24]. In many ways, metacognition is considered to be domain-independent [25]; 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 [26]. 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 [27]. 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 [28] 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 [29] builds on the concept of "cognitive regulation" from Dimmitt and McCormick's work [30] 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 and Qin [31], 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.

IV. Methodology

A. Research design

Figure 1 shows the experimental design used for this research. This study employed a pretest-posttest experimental research design with control and treatment groups. 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.

While participants under the control condition received the TOEFL iBT and metacognition training, they had to complete the three writing tasks (pre-writing, post-writing 1, post-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.

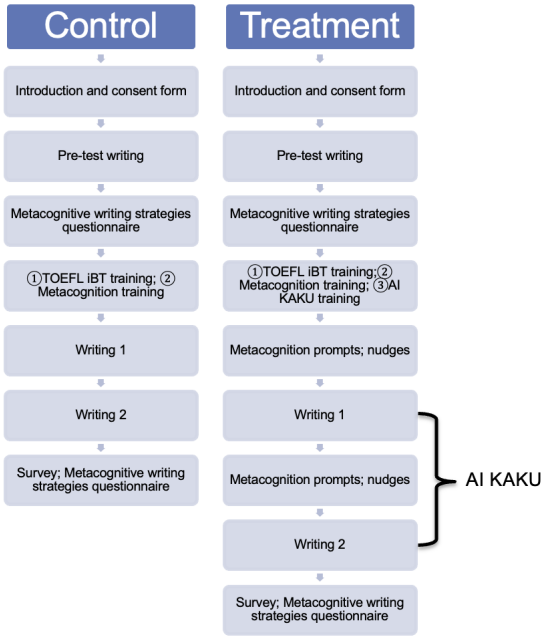

Fig. 1. Experimental design flow. "AI KAKU" is the intelligent writing assistant.

B. Participants

Convenience sampling was used, and Japanese university students were invited to participate in the study from three institutions related to the researchers. The participants' consent was collected via the experiment's website, and they were able to retract their consent at any time during the experiment. 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.

C. Participant training

The metacognitive training consists of a ten-minute video recorded by the researchers largely inspired by Boston University's Teaching Writing Metacognition flipped classroom module [32]. 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 [33] but can be improved with training interventions [34].

Learnability is essential for software to be usable [35]. 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 [36]. 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.

D. Treatment software

This study employs a digital writing assistant called "AI KAKU" that has been previously used by the researchers in a case study [18] that showed potential in improving the writing performance of the EFL participants. This study expands upon that initial study 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 [37].

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 at a glance 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.

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 [38].

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. This measure of lexical diversity has been shown in a study by Treffers-Daller, Parslow, and Williams [39] to equate to the CEFR B1 (IELTS level 4, TOEFL iBT 42-71) level when a score of 70.14 is achieved.

E. Writing quality factors

1) Tools used: The researchers 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 [40]. 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 [41], Mizumoto's [42] web-based R interface (see also Mizumoto and Plonsky [43]) 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 [44] and accessible via https://aihaiyang.com/software/lca/ and the web-based L2 Syntactic Complexity Analyzer also developed by Lu, [45] and accessible via https://aihaiyang.com/software/l2sca/.

2) Writing quality dimensions: In order 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 [46].

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 D [47] 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 [48].

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 [49] 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 [50]. Measures of lexical sophistication are correlated to L2 development [51] as Crossley and McNamara [52] 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 [53].

The last writing quality dimension employed four university EFL educators to serve as raters using ETS's publicly available 6-point holistic assessment rubric. 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 [54]. In other words, the written output 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 [55]. Table I 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. Inter-rater agreement and rating normalization, scaling techniques will be discussed in the Results and Discussion section of the article.

TABLE I
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

F. Metacognitive measures

The Metacognitive Writing Strategies Questionnaire (MWSQ) [56] was modified from 18 questions to just ten questions after the review or course staff to reduce the likelihood of participant dropout. 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 [57]. 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.

V. RESULTS AND DISCUSSION

A. 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 researchers 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 researchers 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 II 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 [58].

TABLE II
SPEARMAN'S INTER-RATER CORRELATION VALUES FOR COMMONLY RATED CHUNK.

	Rater3	Rater4	Rater1	Rater2
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 [59]. This gave the researchers 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 researchers proceeded to standardize the independently scored samples (see Table I) by calculating a z-score using Equation (1).

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

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 researchers 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).

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

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.

B. 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. In order 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 III show that in the pre-test phase, there was no statistical difference along all writing quality measures used, giving the researchers confidence that the control and treatment groups started with similar writing abilities before any treatments were applied.

TABLE III
CONTROL AND TREATMENT PRE-TEST SCORES. MEAN, SD(), AND t-TEST RESULTS.

	Control	Treatment	t-test
Tokens	244.4 (80)	256.8 (75)	$t(119) = -0.86$ $p = 0.38$
MTLD	65.4 (18.4)	63.5 (15.1)	$t(119) = 0.62$ $p = 0.53$
LFP	0.10 (0.03)	0.10 (0.04)	$t(119) = -0.16$ $p = 0.87$
ETS	3.5 (1)	3.5 (0.94)	$t(119) = 0.32$ $p = 0.74$
Ldensity	0.52 (0.04)	0.53 (0.04)	$t(119) = -1.75$ $p = 0.08$
MLTunit	14.1 (3.1)	14.5 (3.4)	$t(119) = -0.69$ $p = 0.49$
Clause/Tunit	1.6 (0.2)	1.7 (0.3)	$t(119) = -1.13$ $p = 0.25$

C. 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 IV. Several

TABLE IV
TWO-WAY ANOVA ANALYSIS OF WRITING QUALITY FACTORS. F-RATIOS, P-VALUES DISPLAYED IN ().

	A	p.eta^2	B	p.eta^2	A - B Interaction	p.eta^2
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;
A = (Control - Treatment), B = (Pretest - Post-test 1 & 2)

dimensions indicate statistical significance between control and treatment conditions and pre-test, post-test writing tasks. According to Cohen's effect size interpretations [60], the ETS dimension exhibited the largest difference with a large effect between control, treatment (factor A), a small effect between pre and post-tests (factor B), and a large effect between the interaction of factor A and B. Considering the ETS dimension demonstrated the strongest directional significance, it is visually represented in Fig. 2. For all the dimensions that exhibited statistical significance, a Holm Bonferroni post-hoc analysis (see Table V) was conducted to determine the direction and strength of the relationships.

The post-hoc analysis reveals that the LFP dimension did not show a statistical difference between the control and treatment groups. However, post-tests 1 and 2 outperform the pre-test writing condition. This indicates to the researchers 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 post-test 1 and post-test 2. This might indicate potential fatigue that some participants reported from doing three writing tasks consecutively.

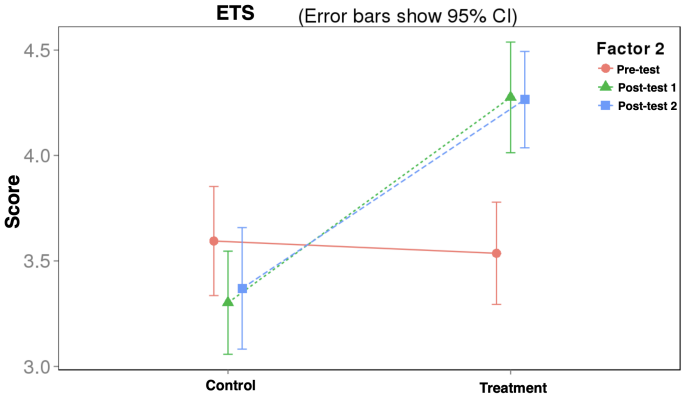

Fig. 2. ANOVA results for ETS human assessment scores.

TABLE V
HOLM BONFERRONI POST-HOC ANALYSIS OF SIGNIFICANT FACTORS. T-VALUES, p-VALUES ARE IN ().

	pre - post 1	pre - post 2	post 1 - post 2
LFP	6.8 (0.0000)*<	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: >, < indicates direction of significance.

The ETS dimension demonstrates significance between the control, treatment conditions, and pre and post-test writing tasks. Both post-test writing tasks perform better than the pre-test while under the treatment condition, while there is no significant difference in performance between pre and post-tests in the control condition. The researchers 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 and post-test writing tasks, albeit with less consistent directionality. Lexical density is improved in the pre-test compared to post-test 1 in the control condition and improved compared to post-test 1 and 2 in the treatment condition, while also showing post-test 2 outperforming post-test 1 in the control condition.

D. 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 Fig. 3 and Table VI) shows statistical significance between the treatment and

control conditions in addition to significance between pre and post-test.

TABLE VI
TWO-WAY ANOVA ANALYSIS OF MWSQ RESULTS.

	F-ratio (p)	p.eta^2
(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

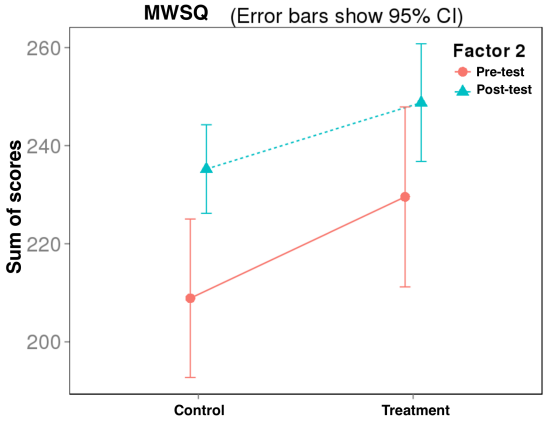

Fig. 3. ANOVA results for pre and post MWSQ survey results.

Following Cohen's guidelines [60], the difference between both control and treatment and pre-test / post-test show large effect sizes, indicating to the researchers that participants in both the control and treatment groups were able to improve their metacognitive awareness of the writing task after metacognitive training.

E. Relative Importance Analysis

Taking the results from the writing quality dimensions and the MWSQ inventory, the researchers conducted a relative importance analysis to gain insight into which factors contribute to a higher ETS score. Using the ETS score as the dependant variable, a Random Forest Boruta analysis identifies the variables that are important to the dependant variable. As seen in Figure 4, 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 important factors that led the human raters to assess the writing sample at a higher level.

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 researchers to fine-tune treatment tools to be used in future studies.

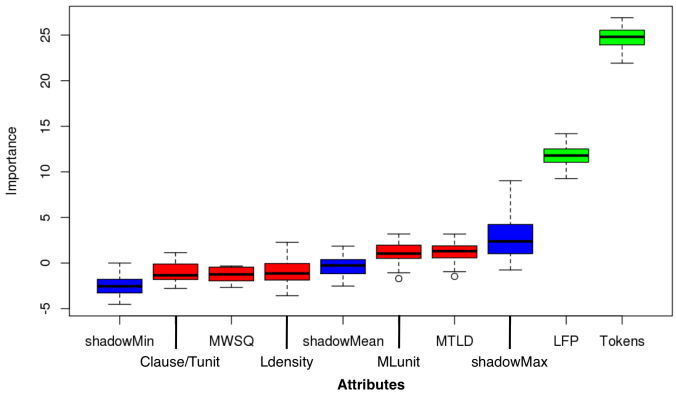

Fig. 4. Boruta (Random Forest) analysis of variable importance to ETS score.

F. Participant feedback

In order 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 = 37) from all the participants were collected. The two survey questions 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 5, 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."

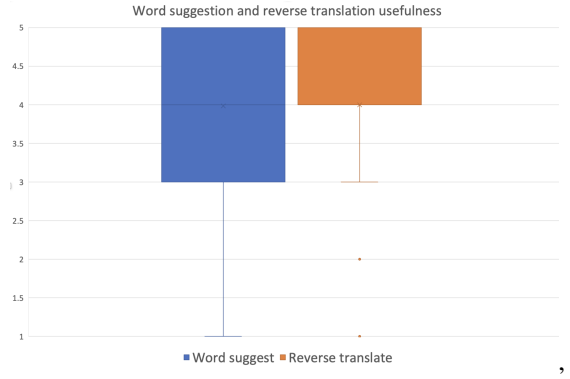

Fig. 5. Post-experiment feedback on AI KAKU. 1 = "Strongly Disagree, 5 = "Strongly Agree".

Looking further into the qualitative feedback received, the researchers analyzed the comments received and subjectively keyed each comment as positive, neutral, or negative. Of the 37 comments, 16 (43%) were tagged as positive, 5 (15%) neutral, and 16 (43%) were tagged as having negative sentiment. Figure 6 visualizes the top frequency words. The word cloud reveals some common themes among the participants, including task difficulty, a desire for improvement, more granular feedback, and pre-task planning. Lastly, selected comments

with both positive and negative sentiments are summarized in Table VII.

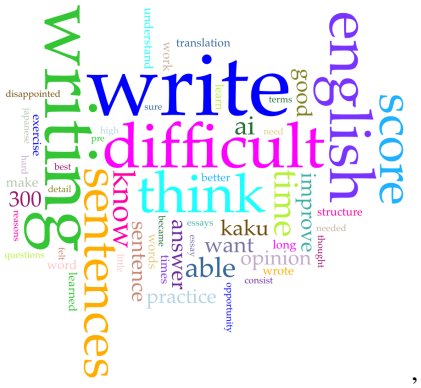

Fig. 6. Word cloud of top frequency words.

TABLE VII
SELECTION OF FEEDBACK FROM PARTICIPANTS.

Sentiment	Comments marked as either positive (+) 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 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.

VI. Conclusion

In this study, the researchers have combined 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 post-tests, as shown by the LFP dimension. Lexical density and the mean length of T-unit were not in line with the researchers' 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 are not important factors.

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 researchers intend 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.

VII. Acknowledgements

We would like to extend our 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 at Tokyo Tech who supported this research.

References

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

[2] W. Zhu, "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*, vol. 13, no. 1, pp. 29–48, 2004.

[3] E. Wheeler and R. L. McDonald, "Writing in engineering courses," *Journal of Engineering Education*, vol. 89, no. 4, pp. 481–486, 2000.

[4] Y. R. Dong, "Non-native graduate students' thesis/dissertation writing in science: Self-reports by students and their advisors from two us institutions," *English for Specific Purposes*, vol. 17, no. 4, pp. 369–390, 1998.

[5] J. Bitchener and H. Basturkmen, "Perceptions of the difficulties of postgraduate l2 thesis students writing the discussion section," *Journal of English for Academic Purposes*, vol. 5, no. 1, pp. 4–18, 2006.

[6] G. Xiao and X. Chen, "English academic writing difficulties of engineering students at the tertiary level in china," *World Transactions on Engineering and Technology Education*, vol. 13, no. 3, pp. 259–263, 2015.

[7] E. Koenig, K. Guertler, D. Żarnowska, and J. Horbačauskienė, "Developing english language competence for global engineers," in *2020 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2020, pp. 242–249.

[8] I. O. for Standardization, "Technical product documentation." [Online]. Available: https://www.iso.org/ics/01.110/x/p/1/u/0/w/0/d/0

[9] H. Aydoğan and A. A. Akbarov, "The four basic language skills, whole language & integrated skill approach in mainstream university classrooms in Turkey," *Mediterranean Journal of Social Sciences*, vol. 5, no. 9, pp. 672–672, 2014.

[10] H. Basturkmen and M. Lewis, "Learner perspectives of success in an EAP writing course," *Assessing writing*, vol. 8, no. 1, pp. 31–46, 2002.

[11] C. M. Tardy and P. K. Matsuda, "The construction of author voice by editorial board members," pp. 32–52, 2009.

[12] D. Nunan and R. Carter, *The Cambridge guide to teaching English to speakers of other languages*. Ernst Klett Sprachen, 2001.

[13] N. O. Ceylan, "Student perceptions of difficulties in second language writing," *Journal of Language and Linguistic Studies*, vol. 15, no. 1, pp. 151–157, 2019.

[14] S.-C. Tsai, "Using google translate in efl drafts: a preliminary investigation," *Computer Assisted Language Learning*, vol. 32, no. 5-6, pp. 510–526, 2019.

[15] L. Abrams and D. K. Davis, "The tip-of-the-tongue phenomenon: Who, what, and why," in *Cognition, Language and Aging*. John Benjamins Publishing Company, 2016, pp. 13–54.

[16] K. Ohtani and T. Hisasaka, "Beyond intelligence: a meta-analytic review of the relationship among metacognition, intelligence, and academic performance," *Metacognition and Learning*, vol. 13, no. 2, pp. 179–212, 2018.

[17] C. Gama, "Metacognition in interactive learning environments: The Reflection Assistant model," in *International Conference on Intelligent Tutoring Systems*. Springer, 2004, pp. 668–677.

[18] J. M. Gayed, M. K. J. Carlon, A. M. Oriola, and J. S. Cross, "Exploring an ai-based writing assistant's impact on english language learners," *Computers and Education: Artificial Intelligence*, p. 100055, 2022.

[19] C. Bowerman, "Writing and the computer: An intelligent tutoring systems solution," in *Computer Assisted Learning: Selected Contributions from the CAL'91 Symposium*. Elsevier, 1992, pp. 77–83.

[20] J. Gamper and J. Knapp, "A review of intelligent CALL systems," *Computer Assisted Language Learning*, vol. 15, no. 4, pp. 329–342, 2002.

[21] X. Dai, Y. Liu, X. Wang, and B. Liu, "Wings: writing with intelligent guidance and suggestions," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 25–30.

[22] M.-H. Chen, S.-T. Huang, H.-T. Hsieh, T.-H. Kao, and J. S. Chang, "FLOW: a first-language-oriented writing assistant system," in *Proceedings of the ACL 2012 System Demonstrations*, 2012, pp. 157–162.

[23] G. Dizon and J. M. Gayed, "Examining the impact of Grammarly on the quality of mobile l2 writing." *JALT CALL Journal*, vol. 17, no. 2, pp. 74–92, 2021.

[24] J. H. Flavell, "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry," *American Psychologist*, vol. 34, no. 10, p. 906, 1979.

[25] R. Azevedo, "Reflections on the field of metacognition: issues, challenges, and opportunities," *Metacognition and Learning*, 6 2020.

[26] A. Yadav, C. Ocak, and A. Oliver, "Computational thinking and metacognition," *TechTrends*, pp. 1–7, 2022.

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

[28] C. Dabarera, W. A. Renandya, and L. J. Zhang, "The impact of metacognitive scaffolding and monitoring on reading comprehension," *System*, vol. 42, pp. 462–473, 2014.

[29] Y. Knospe, "Metacognitive knowledge about writing in a foreign language: A case study," in *Metacognition in language learning and teaching*. Routledge, 2018, pp. 121–138.

[30] C. Dimmitt and C. B. McCormick, "Metacognition in education." in *APA educational psychology handbook, Vol 1: Theories, constructs, and critical issues.* American Psychological Association, 2012, pp. 157–187.

[31] L. J. Zhang and T. L. Qin, "Validating a questionnaire on efl writers' metacognitive awareness of writing strategies in multimedia environments," in *Metacognition in language learning and teaching*. Routledge, 2018, pp. 157–178.

[32] Boston University Teaching Writing, "Metacognition — teaching writing," 2020. [Online]. Available: https://www.bu.edu/teaching-writing/resources/metacognition/

[33] M. K. J. Carlon, J. M. Gayed, and J. S. Cross, "Development of open-response prompt-based metacognitive tutor for online classrooms," in *2021 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*. IEEE, 12 2021.

[34] M. Sato and C. Dussuel Lam, "Metacognitive instruction with young learners: A case of willingness to communicate, L2 use, and metacognition of oral communication," *Language Teaching Research*, p. 13621688211004639, 2021.

[35] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1990, pp. 249–256.

[36] B. Strahm, C. M. Gray, and M. Vorvoreanu, "Generating mobile application onboarding insights through minimalist instruction," in *Proceedings of the 2018 Designing Interactive Systems Conference*, 2018, pp. 361–372.

[37] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[38] K. S. McCarthy, R. D. Roscoe, L. K. Allen, A. D. Likens, and D. S. McNamara, "Automated writing evaluation: Does spelling and grammar feedback support high-quality writing and revision?" *Assessing Writing*, vol. 52, p. 100608, 2022.

[39] J. Treffers-Daller, P. Parslow, and S. Williams, "Back to basics: How measures of lexical diversity can help discriminate between cefr levels," *Applied Linguistics*, vol. 39, no. 3, pp. 302–327, 2018.

[40] ETS, "TOEFL writing rubrics - educational testing service," 2019. [Online]. Available: https://www.ets.org/s/toefl/pdf/toefl-writing-rubrics.pdf

[41] P. M. McCarthy, "An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (mtld)," Ph.D. dissertation, The University of Memphis, 2005.

[42] A. Mizumoto, "Langtest (version 1.0)[web application]. kansai university," 2015.

[43] A. Mizumoto and L. Plonsky, "R as a lingua franca: Advantages of using R for quantitative research in applied linguistics," *Applied Linguistics*, vol. 37, no. 2, pp. 284–291, 2016.

[44] X. Lu, "The relationship of lexical richness to the quality of ESL learners' oral narratives," *The Modern Language Journal*, vol. 96, no. 2, pp. 190–208, 2012.

[45] H. Ai and X. Lu, "A corpus-based comparison of syntactic complexity in nns and ns university students' writing," *Automatic treatment and analysis of learner corpus data*, pp. 249–264, 2013.

[46] S. A. Crossley and D. S. McNamara, "Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication," *Journal of Research in Reading*, vol. 35, no. 2, pp. 115–135, 2012.

[47] P. M. McCarthy and S. Jarvis, "Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment," *Behavior research methods*, vol. 42, no. 2, pp. 381–392, 2010.

[48] R. Koizumi, "Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens," *Vocabulary Learning and Instruction*, vol. 1, no. 1, pp. 60–69, 2012.

[49] R. Breeze, "Researching simplicity and sophistication in student writing," *International Journal of English Studies*, vol. 8, no. 1, pp. 51–66, 2008.

[50] B. Laufer and P. Nation, "Vocabulary size and use: Lexical richness in l2 written production," *Applied linguistics*, vol. 16, no. 3, pp. 307–322, 1995.

[51] C. G. Polio, "Second language development in writing: Measures of fluency, accuracy, and complexity. pp. 187." *Studies in second language Acquisition*, vol. 23, no. 3, pp. 423–425, 2001.

[52] S. A. Crossley and D. S. McNamara, "Shared features of l2 writing: Intergroup homogeneity and text classification," *Journal of Second Language Writing*, vol. 20, no. 4, pp. 271–285, 2011.

[53] J. E. Casal and J. J. Lee, "Syntactic complexity and writing quality in assessed first-year l2 writing," *Journal of Second Language Writing*, vol. 44, pp. 51–62, 2019.

[54] C. S. Wiseman, "A comparison of the performance of analytic vs. holistic scoring rubrics to assess l2 writing," *International Journal of Language Testing*, vol. 2, no. 1, pp. 59–92, 2012.

[55] R. Schoonen, "Generalizability of writing scores: An application of structural equation modeling," *Language testing*, vol. 22, no. 1, pp. 1–30, 2005.

[56] C. G. Zhao and L. Liao, "Metacognitive strategy use in L2 writing assessment," *System*, vol. 98, p. 102472, 2021.

[57] M. A. Revilla, W. E. Saris, and J. A. Krosnick, "Choosing the number of categories in agree–disagree scales," *Sociological Methods & Research*, vol. 43, no. 1, pp. 73–97, 2014.

[58] H. Akoglu, "User's guide to correlation coefficients," *Turkish journal of emergency medicine*, vol. 18, no. 3, pp. 91–93, 2018.

[59] K. Krippendorff, *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[60] J. Cohen, *Statistical power analysis for the behavioral sciences*. Routledge, 2013.