Studies in Second Language Learning and Teaching
Department of English Studies, Faculty of Pedagogy and Fine Arts, Adam Mickiewicz University, Kalisz
SSLLT 0 (0). 2026. 1-26. Published online: 08.06.2026
https://doi.org/10.14746/ssllt.52990
http://pressto.amu.edu.pl/index.php/ssllt

Examining the role of linguistic characteristics of task performance in integrated multimodal viewing-to-write tasks

Judit Kormos ✉
Lancaster University, United Kingdom
University of Ljubljana, Slovenia
https://orcid.org/0000-0002-2643-7222
j.kormos@lancaster.ac.uk

Tineke Brunfaut
Lancaster University, United Kingdom
https://orcid.org/0000-0001-8018-8004
t.brunfaut@lancaster.ac.uk

Borbála Bánfalvi
Queen Mary University of London, United Kingdom
https://orcid.org/0000-0001-8961-8787
b.banfalvi@qmul.ac.uk

Abstract

This study investigates how second language (L2) writers' textual features vary across multimodal integrated tasks and how these features relate to raters' evaluations of performance. Reflecting the growing emphasis on multimodality in language use, our research focused on viewing-to-write tasks that combine auditory, visual, and written input. A total of 133 intermediate to advanced L2 English learners completed two task types: (1) viewing-to-describe, where participants watched a recording explaining and visually depicting a production process before writing a description; and (2) viewing-to-compare-

and-contrast, where learners viewed a videocast of two experts discussing a topic, supported by visuals, and then produced a comparative report. Performances were evaluated using a purpose-developed rating scale, and the written texts were analyzed for linguistic features, including accuracy, cohesion, lexical diversity and sophistication, and syntactic and lexical complexity. Results showed that these linguistic indices explained a substantial proportion of variance in performance scores. Although the overall use of linguistic features did not differ significantly between task types, their relative contribution to scores varied. The findings have important implications for using multimodal integrated tasks in task-based language teaching and assessment.

Keywords: multimodality; second language writing; viewing; task-based assessment; integrated language assessment

1. Introduction

Within the area of second language (L2) learning, teaching, and assessment, a strong tradition exists of the use of, and research on, independent writing tasks. Such writing-only tasks typically comprise short instructions prompting L2 learners to produce a piece of writing for a specified topic, audience, and genre. However, a sizable amount of real-life writing, especially in educational and professional contexts, involves the integration of various language skills. Consequently, integrated writing tasks, such as reading-to-write or listening-to-write, have increasingly been adopted in L2 learning, teaching, and assessment, complementary to independent writing tasks (Brunfaut & Kormos, 2025). Additionally, the fast-accelerating pervasiveness of technology means that communication now goes beyond linguistic modes and relies on visual, aural, gestural, and/or spatial modes to convey meaning; thus, communication is multimodal.

In terms of L2 research, a solid body of work on integrated writing has developed over the past two decades. Substantial insights have been generated into issues such as L2 learners' integrated writing task processes (e.g., Barkaoui, 2014), source use in integrated writing (e.g., Plakans & Gebril, 2012), approaches to scoring integrated writing performances (e.g., Chan & May, 2023; Kormos & Suzuki, 2024), and the predictive power of both receptive and productive skills, and their integration, for integrated writing scores (e.g., meta-analysis by Chan & Yamashita, 2022). A consistent conclusion of such research is that, even though during task administration test-takers use their receptive skills (listening and/or reading) prior to the production stage (writing), integrated writing tasks cannot be viewed simply as a sequence of receptive tasks followed by a productive task. Furthermore, building on the extensive body of research on linguistic characteristics of independent writing

performances, the discourse features of integrated writing performances have also been explored (see below), although comparatively much less so than the complexity, accuracy, and fluency (CAF) of independent writing.

Given the relative novelty of integrated *multimodal* tasks, which involve modes beyond narrowly linguistic ones, even less research is currently available on the linguistic features these tasks elicit and how they relate to raters' perceptions of successful task performance. Therefore, the research reported here aimed to provide insights into language use in multimodal tasks, through a focus on viewing-to-write tasks in which L2 learners watch a video with aural, visual, and gestural input and then write a text based on it. More specifically, this study aimed to expand the investigations of linguistic characteristics of integrated multimodal writing performances and to examine how these textual features contribute to ratings of multimodal performance.

2. Literature review

2.1. The construct and cognitive processes involved in multimodal integrated writing tasks

Integrated writing tasks, as their name suggests, involve the integration of various language skills. In integrated tasks that use source texts either in written, auditory, or multimodal format, L2 learners strategically allocate their attentional resources and select information that is relevant for the writing task and take notes to support their subsequent writing processes. Thus, they bridge and interconnect the various skills and modes involved in and required for successful integrated task completion. For example, in Brunfaut and Kormos' (2025) research on multimodal integrated viewing-to-write tasks, teenage L2 learners of English reported that they frequently engaged in task-oriented viewing and that they used their notes to help them recall relevant content, to organize and attribute ideas from the source text, and to transition into the writing stage.

Multimodal integrated writing tasks have also been shown to elicit a wide range of writing processes. For example, Brunfaut and Kormos (2025) found evidence that lower-order writing processes were operating effectively and that sufficient cognitive resources were available for linguistic formulation, editing, and revision in integrated viewing-to-write tasks. Similarly, their study showed that multimodal integrated writing tasks can elicit higher level writing processes such as summarization, paraphrasing strategies, source attribution, as well as coherent and well-organized content representation. This resonates with findings on integrated writing tasks more generally (not just multimodal ones) such as those of Plakans and Gebril (2012).

Abrams (2019) argued that content provision for a subsequent writing activity, like the source input in integrated tasks, might reduce the complexity of the task (Robinson, 2005; Skehan & Foster, 2012) and might make additional attentional resources available for the grammatically accurate formulation of ideas and for varied lexical choice in integrated task performances. In a classroom-based assessment study, Abrams (2019) observed that in multimodal viewing-to-write tasks students used a variety of lexical expansion and syntactic transformation strategies, probably due to the availability of attentional resources for the linguistic formulation stage of writing. Furthermore, these types of tasks can also facilitate alignment and prime learners to use lexical items and phrases as well as syntactic constructions from the input, as highlighted by Peng et al.'s (2018) research with story continuation tasks. However, predetermined content might also add to cognitive demands by requiring L2 learners to use specific lexical and syntactic constructions to convey the information presented in the input.

2.2. Previous research on linguistic predictors of L2 writing quality

One of the key determinants of L2 writing quality is writers' underlying L2 knowledge, which comprises linguistic, discourse, and sociolinguistic knowledge (Bachman & Palmer, 1996) and which writers can use as a resource for expressing their ideas in written form (Schoonen et al., 2011). Our study focuses on linguistic knowledge, which entails knowledge of syntax, morphology, orthography and lexis, and the organizational aspect of discourse knowledge, namely, knowledge about how sentences are connected to form a cohesive text.

Several studies have examined how the textual features of L2 writing, such as its lexical and syntactic complexity and sophistication, accuracy, and cohesion, relate to either learners' L2 proficiency or writing scores. The underlying assumption is that insights from this line of research help us understand how efficiently learners can draw on their underlying linguistic and discourse competence during the writing process. In a recent meta-analysis of the relationship between linguistic measures and holistic ratings of L2 writing quality in *independent* writing (mostly narrative and argumentative tasks), Kojima and Kaneta (2022) found that productivity (quantity/length of L2 learners' output) had the strongest link (r = .57) with L2 writing scores. They also established a moderately strong correlation between L2 writing scores and accuracy (r = .47), which was substantially stronger than the association between L2 writing scores and lexical complexity (r = .29), syntactic complexity (r = .27), and cohesion (r = .20). Neither L2 proficiency nor the type of independent task moderated the relationship for any of the linguistic measures. For *integrated* writing, Chan and Yamashita's (2022) meta-analysis of

the relationship between textual features and integrated writing, mostly the integrated reading-writing and reading-listening-writing tasks of the Test of English as a Foreign Language internet-Based Test (TOEFL iBT), revealed that, just as in independent writing tasks, productivity was the strongest predictor of scores (r = .53), and syntactic complexity (r = .32) was also an important contributor. However, compared to the independent writing findings in Kojima and Kaneta's (2022) meta-analysis, organizational features (r = .34) showed stronger links, and lexical complexity (r = .14) weaker ones, with scores on integrated writing (as opposed to independent writing). The relatively small number of studies in Chan and Yamashita's (2022) meta-analysis did not allow for examining the moderating effect of L2 proficiency or integrated task type.

Zooming in on individual linguistic characteristics, in the context of both *independent* and *integrated* writing tasks, Kim and Crossley's (2018) more robust structural equation modeling approach revealed that the mean length of clauses was not a direct predictor of scores on TOEFL iBT writing-only and reading-listening-writing tasks; instead, it had an indirect effect through the mediation of the factor score labelled *lexical decision time* (more frequent use of words that take longer to retrieve). Thus, in both task types, raters might have given higher scores for longer sentences only when the test-takers used more sophisticated words. Plakans et al. (2019), who also investigated the TOEFL iBT reading-listening-writing task, found a significant, but small correlation between the mean length of clauses and the integrated task scores in a regression model that additionally included productivity and accuracy measures. Zhang and Ouyang's (2023) and Xu's (2019) studies of *integrated* reading-to-write-a-summary complemented these significant findings by showing that noun phrase and fine-grained syntactic sophistication indices were significantly related to raters' scores.

Regarding accuracy, Kojima and Kaneta's (2022) meta-analysis of *independent* writing tasks confirmed earlier findings by Brodkey and Young (1981) that more highly rated independent writing task performances contained fewer errors. In the context of *integrated* writing tasks, similar results were obtained by Cumming et al. (2005), and Gebril and Plakans (2013). Plakans et al. (2019) specifically showed that morphological error frequency was an important predictor of holistic integrated writing ratings. They argued that the input from the source text might reduce the number of lexical errors and that in L2 English writing more opportunities for making morphological errors might exist.

As for lexical sophistication, there is consistent evidence across *independent* and *integrated* task performances that highly rated L2 written outputs contain words and word combinations that occur with lower frequency in various corpora (e.g., Crossley & McNamara, 2012; Garner et al., 2020; Guo et al., 2013; Zhang & Ouyang, 2023). Jung et al. (2015) additionally found that in high-scoring

independent essays there are more words that have higher age of acquisition ratings (i.e., words acquired later by children speaking their first language). A similar significant correlation was found for *independent* and *integrated* writing tasks by Kim and Crossley (2018). Another indicator of lexical sophistication is the use of words that take longer to retrieve from the mental lexicon (Berger et al., 2019). Kim and Crossley (2018), in fact, found that these lexical decision-based time measures had the largest explanatory power for both *independent* and *integrated* writing tasks.

Regarding lexical diversity, Gebril and Plakans (2016) found significant correlations with TOEFL iBT *integrated* reading-listening-writing scores. In contrast, lexical diversity measures did not have significant predictive power in Zhang and Ouyang's (2023) study, which might be due to the short length of the required written summaries. Also, as mentioned, Chan and Yamashita's (2022) meta-analysis found a relatively weak relationship between lexical features and *integrated* writing task performance, but it might be because lexical diversity and sophistication measures were considered jointly, and their study pool contained texts written by both first language (L1) and L2 writers.

Finally, concerning cohesion measures, research indicates that these play a considerably smaller role in evaluations of both *independent* (Kojima & Kaneta, 2022) and *integrated* L2 writing quality (Chan & Yamashita, 2022) compared to accuracy, syntactic, or lexical measures. This may be for several reasons. On the one hand, the coherence of a text is a mental representation of the relationships of idea/meaning units in the text in readers' minds (Halliday & Hasan, 1976) and can only partially be described by examining linguistic expressions of cohesion. On the other hand, the use of cohesive devices at different levels of the text, such as sentence-, paragraph-, and text-level cohesion, might vary depending on L2 writing ability. Indeed, evidence indicates that lower rated *independent* essays tend to use more frequent lexical cohesive devices at local (sentence) levels and fewer cohesive ties at global (paragraph) levels (e.g., Crossley & McNamara, 2012). Similar negative relationships between the use of connectives and writing scores were observed by Phillips Galloway et al. (2020) for a mixed group of English L1 and L2 children in an *integrated* reading-into-writing task. Plakans and Gebril (2017) showed that, for the TOEFL iBT reading-listening-writing tasks, the frequency of cohesive tools did not differ across total task score levels. However, higher ratings on the same TOEFL iBT task type were significantly associated with stronger semantic similarity between sentences in the study by Guo et al. (2013), who argued that more explicit markers of cohesion are needed to effectively integrate the information from written sources in a summary.

As this review demonstrates, research on the linguistic predictors of L2 writing has concentrated on independent writing and conventional integrated

tasks (oral and written language modes). No previous English L2 research has examined how the linguistic characteristics of integrated *multimodal* viewing-to-write tasks predict human raters' scores. Additionally, tasks can vary in the language functions they target, which might influence linguistic characteristics of performances. Thus, it seems meaningful to also investigate whether the textual features of L2 writers' output vary in different types of multimodal viewing-to-write tasks and how the linguistic characteristics of students' written text predict evaluations of multimodal task performance. Specifically, we focus on tasks that require a process description (viewing-to-describe) and tasks that require the comparison and contrast of two speakers' views (viewing-to-compare-and-contrast; for a rationale for these language functions, see Brunfaut & Kormos, 2025). We investigate this in an understudied group of teenage L2 learners. Therefore, our research questions were:

RQ1: How do linguistic characteristics of task performance differ in integrated multimodal viewing-to-describe and viewing-to-compare-and-contrast tasks?

RQ2: How do linguistic characteristics of task performance predict ratings of integrated multimodal viewing-to-describe and viewing-to-compare-and-contrast tasks?

3. Method

3.1. Participants

The participant group comprised 133 upper-secondary pupils from six schools in Hungary (49% Year 10; 20% Year 11; 31% Year 12). They were English as a foreign language (EFL) learners, with their proficiency ranging from A2 to C levels on the *Common European Framework of Reference* (CEFR; Council of Europe, 2001, 2020) (listening: A2 = 1%, B1 = 2%, B2 = 19%, C1 and C2 = 78%; writing: B1 = 5%, B2 = 46%, C1 and C2 = 49%), as measured by the Aptis General test (https://www.britishcouncil.org/exam/english/aptis; note that Aptis General does not separate out scores at the C levels). Their home languages were Hungarian (97.8% of pupils), Chinese (1.5%), and Arabic (0.7%). 40% of our participants were female, and 60% were male.

3.2. Instruments: Viewing-to-write describe and compare-and-contrast tasks

We developed two types of integrated multimodal viewing-to-write tasks, a description- and compare-and-contrast task, targeting the CEFR B and C levels. We

chose these task types because the language functions of description and comparison-contrast represent authentic and widely used communicative purposes in both educational and professional contexts. These language functions are characterized by distinct linguistic features, such as specific conjunctions and syntactic structures, and they are commonly incorporated into language learning and assessment tasks across a range of proficiency levels. In the viewing-to-describe tasks, learners first watch a video in which they hear a speaker explain how something is made and see pictures, animations, numbers, and a couple of terms/phrases displayed when relevant to the content. The video is played twice (for a rationale, see Holzknecht & Harding, 2024) and the learner is allowed to take notes while watching. Then, the learner has to write a short magazine article (150-200 words) describing the production process. The topics of the tasks were chosen so that they concerned everyday items whose production processes few people know exactly about. This was to avoid learners writing their text based on background knowledge rather than their comprehension of the video. Similarly, while the visuals in the videos illustrated and enriched the spoken input, they were not so detailed that learners could write a satisfactory article without understanding the spoken input. The videos (task instructions + single play) were 4.5 minutes long, and learners had 18 minutes to write their text (time limits were established through extensive piloting; for details, see Brunfaut & Kormos, 2025). We developed three examples of viewing-to-describe tasks: one on how instant coffee is produced, one on how a train carriage is constructed, and one on how contact lenses are made.

The second task type was viewing-to-compare-and-contrast tasks, in which learners first watch a video in which they see and hear two experts who provide their insights on a specific topic. The expert discussion is accompanied by pictures, animations, graphs, and a couple of terms, as relevant to what the experts are talking about. The video is played twice, and the learner is allowed to take notes. Then, the learner writes a short report (200-250 words) in which they need to compare and contrast the two experts' views. Background knowledge is not expected to be a confounding factor, as the task emphasizes the content of each expert's contributions and the extent to which the two experts' perspectives on the topic align or diverge from one another, rather than requiring the learner's personal opinions or general advantages and disadvantages of the topic. Likewise, the visual elements of the videos do not provide sufficient information for learners to produce an adequate report without comprehending the accompanying audio input. The videos (task instructions + single play) were 4.5 minutes long, and learners had 20 minutes to write their text. We developed three versions of the viewing-to-compare-and-contrast task format: one in which the experts discuss space tourism, one on whether zoos should

exist, and one on car usage. An overview of the task types is shown in Figure 1. A more detailed description of the task development processes, including piloting, and the nature of the tasks is provided in Brunfaut and Kormos (2025). The tasks are freely available at https://osf.io/xude7/.

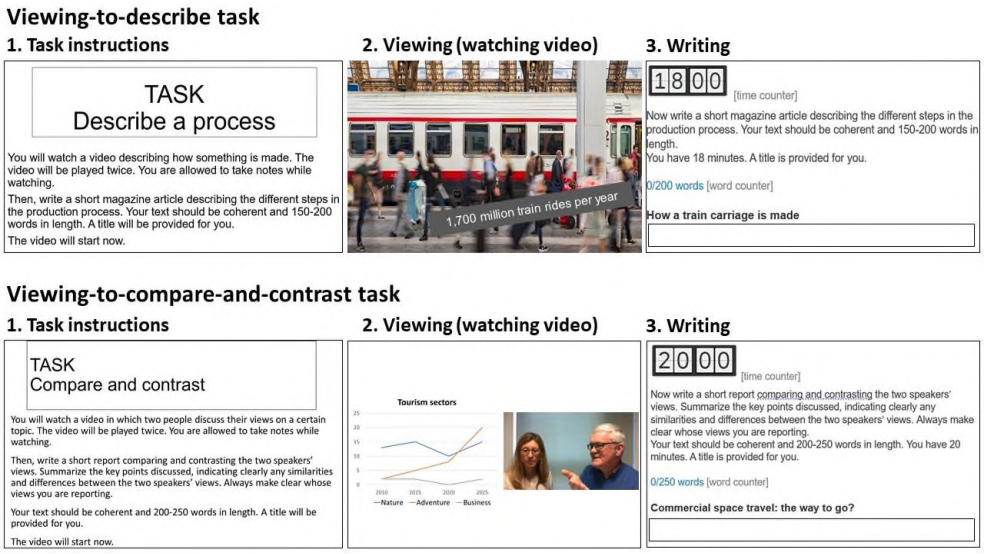

Figure 1 Viewing-to-write task examples (reproduced from Brunfaut & Kormos, 2025, p. 437; train image by benfuenfundachtzig @Pixabay.com)

3.3. Data collection procedures

First, permission was sought from the six schools and their English teachers. Then, pupils and parents/legal guardians were informed about the project via information sheets in Hungarian, and written consent was obtained from the pupil and their parent/guardian. Participation was voluntary. Overall ethics approval for the project was granted by Lancaster University's FASS-LUMS Research Ethics Committee.

Given the cognitive demands of the tasks and the time needed to complete each individual task (approx. 30 min), four of the six tasks were retained for the main study, that is, two viewing-to-describe tasks (contact lenses, train carriage) and two viewing-to-compare-and-contrast tasks (zoos, space tourism). The tasks were administered in pupils' classrooms during school hours, on PCs/laptops and using earphones/headphones. As slots of 1.5 hours were made available to the researchers, pupils were asked to complete three tasks, including at least one of each task type. A counterbalanced design that controlled for order and task type was adopted. The first task type was always a viewing-to-

describe task, as piloting had indicated that this was a more familiar language function for the pupils. Participants were randomly allocated to a group (see Table 1).

Table 1 Counterbalanced design for viewing-to-write data collection (reproduced from Brunfaut & Kormos, 2025, p. 438)

Group		First task	Second task	Third task
Group 1	Group 1a	Describe-1	Compare & Contrast-1	Compare & Contrast-2
	Group 1b	Describe-1	Compare & Contrast-2	Compare & Contrast-1
Group 2	Group 2a	Describe-2	Compare & Contrast-1	Compare & Contrast-2
	Group 2b	Describe-2	Compare & Contrast-2	Compare & Contrast-1
Group 3		Describe-1	Describe-2	Compare & Contrast-1
Group 4		Describe-2	Describe-1	Compare & Contrast-2

The design in Table 1 was followed with 113 participants. The remaining 20 participants also took part in post-task recall interviews, reported on in Brunfaut and Kormos (2025). Given the additional time required for those recall interviews, these 20 pupils only completed two tasks each. These tasks were also administered in a counterbalanced design controlling for task and order, always starting with one of the two viewing-to-describe tasks and then completing one of the two viewing-to-compare-and-contrast tasks.

3.4. Data analyses

3.4.1. Scoring

In order to establish the extent to which participants had successfully completed the viewing-to-write tasks, and thus score their task performances, we developed two analytic rating scales – one for each task type (viewing-to-describe; viewing-to-compare-and-contrast). Both scales comprised four rating criteria:

1. *Viewing for writing*: This criterion focused on comprehension of the video input, that is, the relevant, comprehensive, and accurate representation of its content. Given the different language functions targeted by the two task types, the rating descriptors of this criterion were specific to, and thus different for each task type. Additionally, crib sheets stipulated for each individual task which content points from the video input needed to be reflected in the written performance.
2. *Organization and structure*: This criterion evaluated the written text's organization, flow, and use of signposting. Again, given the different target language functions of the two task types, the rating descriptors were

specific to each task type. Importantly, as this concerns an integrated task, the descriptors also referred back to the video (e.g., evidence of correct references to signpost which expert conveyed which opinion in the viewing-to-compare-and-contrast tasks).

3. *Language use*: This criterion assessed the accuracy, range, and complexity/sophistication of the grammar and vocabulary used in the written performance, and any effects of linguistic errors on a reader's comprehension of the text. It also referred back to the video input, urging raters to consider these aspects in relation to the language use in the video. The criterion was shared between the two task types.
4. *Mechanics*: This criterion was used to score spelling and punctuation control aspects of the written performances. The descriptors were the same for both task types.

The first three criteria (viewing for writing; organization and structure; language use) were scored on a scale of 0-4. The last criterion (mechanics) was scored on a scale of 0-2. Detailed information on the rating scale development process, including piloting, is provided in Brunfaut and Kormos (2025). The rating scales are freely available at https://osf.io/xude7/.

Task performances were scored by two experienced raters using the two scales. Training on 30 performances showed good inter-rater reliability (*viewing-to-describe* Spearman's *rho* = .94, $p < .01$; *viewing-to-compare-and-contrast Spearman's rho* = .77, $p < .01$). For our research purposes, the two raters' scores on each performance were therefore averaged.

3.4.2. Linguistic measures

In selecting the measures, we considered previous studies by Kim and Crossley (2018), Zenker and Kyle (2021), Mizumoto and Eguchi (2023) and Kormos and Suzuki (2024) as well as Kojima and Kaneta's (2022) meta-analysis. To reduce the risk of Type I error resulting from performing numerous statistical tests, we applied principal component analysis to establish whether linguistic variables used to tap into the hypothesized underlying language construct formed a common factor. If a factor was successfully identified, we calculated a regression score, which was then used for further analyses. We checked normal distribution for all linguistic measures by examining skewness and kurtosis levels. The values for skewness and kurtosis were between −2 and +2, indicating univariate normality (George & Mallery, 2010). The only exception to this was the variable of dependent clauses per clause, which had a kurtosis value of 2.69 in the viewing-to-compare-

and-contrast task type, but we decided to keep this variable in the analysis because it was the only one that tapped into syntactic complexity, and was normally distributed in the viewing-to-describe task type. We also examined multicollinearity between indices, and the inter-correlation of the selected variables did not exceed $r > .70$ (see Appendix).

We operationalized accuracy using the number of grammatical, spelling, and punctuation errors per 100 words, which was established with the help of the GPT-4 Turbo model using the OpenAI API interface (see Table 2 for an overview of the linguistic of measures). Using the procedures and prompts proposed by Mizumoto et al. (2024), we processed the text of the essays using Google Collab in Python (for discussion of the reliability of this procedure, see Mizumoto, 2025). The factor analysis confirmed that the three accuracy measures formed a single factor, which we called the error factor. This factor had an eigenvalue of 1.92, explaining 64.09% of the variance (KMO = .620; Bartlett's test of sphericity $p < .001$).

To tap into the lexical retrieval speed aspect of lexical sophistication, we followed Kim and Crossley (2018) and Mizumoto and Eguchi (2023), and used the variables of lexical decision time *Z* score and *SD* for content words and word naming response time Z score for content words to create a factor score. This lexical retrieval speed factor had an eigenvalue of 1.58 and explained 61.92% of the variance (KMO = .651; Bartlett's test of sphericity $p < .001$). Lexical complexity was assessed by the single variable of the MRC familiarity index for content words, and lexical frequency by the two variables of Corpus of American English (COCA) Academic Frequency Log Content Words and COCA Academic Bigram Frequency Log. We originally aimed to create a factor score based on the three key variables applied in Kormos and Suzuki (2024), and Mizumoto and Eguchi (2023) that would have included COCA Academic Frequency Log Function Words, but these variables failed to converge on a single factor. Another aspect of lexical sophistication was the lexical diversity factor (eigenvalue = 2.85, variance explained = 71.30%; KMO = .700; Bartlett's test of sphericity $p < .001$), which was formed using the variables of MTLD for content and function words, HDD42-index for all word types, and HDD42-index for function words.

For measuring syntactic complexity, we used the single variable of dependents per clause because the mean length of T-units showed many outliers and was strongly intercorrelated with this measure. Based on Mizumoto and Eguchi (2023), and Kormos and Suzuki (2024), cohesion was operationalized as the frequency of all connectives and the latent semantic analysis (sentence) index. We decided not to include word count as a measure because the text length (number of words) had been specified for the writing tasks, and it showed a relatively narrow distribution.

Table 2 Description of linguistic performance measures used in the study

Linguistic aspect	Measure	Tool
Error factor	Frequency of grammatical errors	GPT-4
	Frequency of spelling errors	Turbo
	Frequency of punctuation errors	OpenAI API
Lexical sophistication		
Lexical retrieval speed	Lexical decision time Z score for content words	TAALES
(one factor)	Lexical decision time SD for content words	
	Word naming response time Z score for content words	
Lexical complexity	MRC familiarity for content words	TAALES
Lexical frequency	COCA Academic Frequency Log Content Words;	TAALES
(single variables)	COCA Academic Bigram Frequency Log	
Lexical diversity	MTLD for content words	TAALED
(one factor)	MTLD for function words	
	HDD42-index for all word types	
	HDD42-index for function words	
Syntactic complexity	Dependent clause per clause	L2SCA
Cohesion	Frequency of all connectives	TAACO
(single variables)	Latent Semantic Analysis (sentence)	

Note. TAALED = Tool for the Automatic Analysis of Lexical Diversity; TAALES = Tool for the Automatic Analysis of Lexical Sophistication; TAACO = Tool for the Automatic Analysis of Cohesion; L2SCA = L2 Syntactic Complexity Analyzer

3.4.3. Statistical analyses

To answer RQ1, generalized linear mixed-effects modeling (GLMM) was employed. The GLMMs were estimated through the *lmer* function in the *lme4* package (Bates et al., 2015), using R statistical software 4.0.2 (R Core Team, 2021). The GLMMs were constructed separately for each linguistic measure, using task type as a categorical fixed-effect predictor variable with task content and individual participants as random-effects predictors. The maximal model with random slopes for *participant* and *task content* failed to converge, and therefore the random slopes were removed.

To examine the predictive role of linguistic measures for task performance ratings (RQ2), GLMM models were constructed separately for the two tasks. The independent variables were the above linguistic measures except for the two variables aimed at assessing cohesion and the Medical Research Council (MRC) database familiarity content word index. The two cohesion measures showed no significant correlation (see Appendix for correlations) with the task ratings, while the MRC familiarity value showed moderately strong inter-collinearity with the lexical diversity and sophistication measures. Therefore, to preserve statistical power a decision was made to exclude these variables. We used the *lmer* function of the lme4 package (version 1.1.27.1; Bates et al., 2015) in R (version 4.1.2; R Core Team, 2021). The significance of fixed effects was assessed with the Satterthwaite approximation

for degrees of freedom using the lmerTest package (Kuznetsova et al., 2017). The models included random intercepts of participants and task content only because the maximal models including random slopes failed to converge. For each task type, we estimated a model in which the dependent variable was the total viewing-to-write score (accurate representation of the listening input and the three language-related criteria).

4. Results

For RQ1, we examined whether the linguistic characteristics of integrated multimodal task performance differed across the two tasks. Table 3 reports descriptive statistics for the linguistic performance measures. These results suggest that, in the viewing-to-compare-and-contrast task, participants made somewhat more errors and produced slightly less complex sentences, but they used more diverse and sophisticated vocabulary than in the viewing-to-describe task. Performances on the describe tasks contained words that had lower familiarity ratings than those in the compare-and-contrast task. However, none of these differences were statistically significant (see Table 4). Similarly, the mean values for the cohesion measures (frequency of connectives and latent semantic analysis) were almost identical in the two tasks and were not statistically significantly different.

Table 3 Descriptive statistics for linguistic performance measures

Linguistic measures	*Min*	*Max*	*Mean*	*SD*	*Skew*	*Kurtosis*
D Error factor	-2.03	4.40	-.08	1.01	.98	1.65
C&C Error factor	-1.66	2.92	.08	.98	.68	.36
D Lexical diversity factor	-3.94	2.96	-.31	1.00	-.11	.88
C&C Lexical diversity factor	-2.19	3.08	.31	.89	-.26	.64
D Lexical retrieval speed factor	-2.17	1.94	-.22	.85	.27	-.04
C&C Lexical retrieval speed factor	-3.03	2.50	.23	1.09	-.20	-.39
D Dependent clause per clause	.00	.61	.37	.13	-.73	.53
C&C Dependent clause per clause	.00	.73	.43	.12	-1.06	2.69
D MRC familiarity for content words	560.09	594.23	575.52	5.37	.12	.68
C&C MRC familiarity for content words	562.55	599.82	579.04	6.20	.18	.01
D COCA Academic frequency log content words	1.74	2.47	2.11	.13	-.13	.02
C&C COCA Academic frequency log content words	1.92	2.53	2.19	.11	.25	-.13
D COCA Academic bigram frequency log	1.00	1.59	1.31	.10	-.07	-.03
C&C COCA Academic bigram frequency log	1.01	1.62	1.30	.12	.01	-.04
D Latent semantic analysis (sentence)	00.	.75	.32	.13	.68	1.14
C&C Latent semantic analysis (sentence)	.00	.65	.30	.12	.19	.64
D Frequency of all connectives	.03	.14	.08	.02	.24	.43
C&C Frequency of all connectives C&C	.01	.14	.08	.02	.13	-.20

Note. D = viewing-to-describe task, C&C = viewing-to-compare-and-contrast task

Table 4 Comparison of linguistic measures between the two task types

Fixed effects	Estimate	*SE*	*t*	*p*
Error factor				
Intercept	0.04	0.17	0.26	.815
Task type Describe	-0.18	0.22	-0.80	.506
Lexical diversity factor				
Intercept	0.31	0.38	0.82	.492
Task type Describe	-0.58	0.54	-1.09	.388
Lexical retrieval speed factor				
Intercept	0.25	0.20	1.21	.327
Task type Describe	-0.44	0.28	-1.58	.249
Dependent clause per clause				
Intercept	0.43	0.03	16.27	.003
Task type Describe	-0.06	0.04	-1.74	.225
COCA Academic frequency log content words				
Intercept	2.19	0.03	68.00	< .001
Task type Describe	-0.08	0.05	-1.81	.210
COCA Academic bigram frequency log				
Intercept	1.30	0.02	64.82	< .001
Task type Describe	0.01	0.03	0.46	.691
MRC familiarity content words				
Intercept	580.92	1.46	397.57	< 0.001
Task type Describe	-3.20	1.91	-1.68	.100
Frequency of all connectives				
Intercept	0.08	0.00	48.68	<0.001
Task type Describe	0.00	0.00	-1.03	.306
Latent semantic analysis (sentence)				
Intercept	0.30	0.01	31.63	< 0.001
Task type Describe	0.02	0.01	1.51	.133

To examine how linguistic characteristics predict ratings (RQ2), we used mixed effects modeling separately for the two task types. The linear mixed effects model for the total viewing-to -write score for the viewing-to-describe task (see Table 5) captured 46% proportion of variance (marginal R^2 = .46 for fixed effects, conditional R^2 = .67 with random effects). The use of higher frequency COCA Academic content words was associated with significantly lower total viewing-to-write scores (b = -2.81, p = .024). Similarly, higher error factor values were related to lower scores *(*b = -1.26, p < .001). In contrast, a greater proportion of dependent clause per clause (b = 3.77, p = .001) and higher lexical diversity (b = 0.49, p = .005) were both positively related to the total viewing-to-write score. COCA Academic bigram frequency (b = -1.27, p = .445) and lexical retrieval speed factor (b = -0.29, p = .132) were not significant predictors in the model.

For the viewing-to-compare-and-contrast task (Table 6), the linear mixed-effects model for the total viewing-to-write score accounted for a somewhat higher proportion of the variance than the model for the describe task (marginal

R^2 = .60 for fixed effects, conditional R^2 = .85 with random effects). The modeling showed significant predictors for lexical diversity (*b* = 0.43, *p* = .009), the use of words with higher lexical retrieval speed (*b* = 0.67, *p* < .001) and error frequency (*b* = -1.03. *p* < .001), and non-significant predictors for the frequency of academic content words (*b* = -0.03, *p* = .981), bigram frequency (*b* = -1.29, *p* = .240), and the proportion of dependent clause per clause (*b* = 0.26, *p* = .794). There was evidence of between-participant variability in both linear mixed-effects models, but no substantial variance attributable to task content.

Table 5 Summary of mixed effects model predicting total viewing-to-write score for the viewing-to-describe task

Predictor	Estimate (*b*)	*SE*	*t*	*p*
Intercept	15.37	3.153	4.87	< .001
COCA Academic frequency log content words	-2.81	1.235	-2.28	.024
COCA Academic bigram frequency log	-1.27	1.657	-0.77	.445
Dependent clause per clause	3.77	1.126	3.35	.001
Lexical diversity factor	0.49	0.173	2.85	.005
Lexical retrieval speed factor	-0.29	0.191	-1.52	.132
Accuracy factor	-1.256	0.159	-7.89	< .001

Table 6 Summary of mixed effects model predicting total viewing-to write score for the compare-and-contrast task

Predictor	Estimate (*b*)	*SE*	*t*	*p*
Intercept	10.80	2.59	4.17	< .001
COCA Academic frequency log content words	-0.03	1.13	-0.02	.981
COCA Academic bigram frequency log	-1.29	1.10	-1.18	.240
Dependent clause per clause	0.26	1.01	0.26	.794
Lexical diversity factor	0.43	0.16	2.65	.009
Lexical retrieval speed factor	0.67	0.14	4.89	< .001
Accuracy factor	-1.03	0.15	-7.04	< .001

5. Discussion

Our first research question focused on the differences in the linguistic characteristics of two types of integrated multimodal task performances: a description task and a compare-and-contrast task. The results revealed that the accuracy, cohesion measures, lexical diversity and sophistication, and syntactic and lexical complexity indices did not differ across the two types of tasks, suggesting that both task types elicited similar linguistic performance. This finding is consistent with the results of Brunfaut and Kormos (2025) that showed no significant differences in any of the analytic criteria scores or the total viewing-to-write score awarded to the learners'

writing. This lack of differences in the linguistic measures and scores might be due to several factors (and their combination). First, in terms of task design features, both types of integrated multimodal tasks might have posed similar attentional and working memory demands for the learners. In designing and administering the tasks, we aimed to reduce any construct-irrelevant variance potentially caused by individual differences in attention regulation and working memory resources. We did this through allowing learners to listen twice, permitting notetaking, and providing sufficient time for writing. Thus, according to cognitive load theory (Sweller et al., 2019), we ensured that the extraneous cognitive load of the task would not interfere with performance. Secondly, it is possible that the intrinsic cognitive load of the two tasks was similar because the learners were quite proficient in the L2, most being at CEFR B2 and C levels, and they had relevant L2 knowledge and skills to execute the tasks equally well despite different genre and content demands. Thirdly, the intrinsic cognitive load of the two task types might not have differed much either because input for writing was provided through multimodal viewing.

From a task-based language teaching and assessment perspective, our findings seemingly differ from previous research, which demonstrated genre effects on the syntactic complexity and lexical characteristics of L2 learners' writing performance for independent writing tasks (e.g., Yoon & Polio, 2017). As mentioned earlier, our task design features and the integrated multimodal nature of our tasks might have neutralized the genre effects. It might also be possible that descriptions of processes and summarizing contrasting views elicited lexical and syntactic constructions that did not differ in complexity and sophistication for learners who were mainly at the B2 and C levels.

In our study, we also investigated how linguistic characteristics of task performance predict ratings of multimodal viewing-to-describe and viewing-to-compare-and-contrast tasks (RQ2). Our findings showed that the selected relevant measures of accuracy, cohesion, lexical diversity and sophistication, and syntactic and lexical complexity explained a substantial proportion of the variance in the total viewing-to-write score awarded to participants' performances. This suggests that our choice of performance measures and the decision to create factor scores were appropriate to tap into relevant characteristics of written texts produced in integrated multimodal viewing tasks and align well with evaluations of writing quality by human raters. Our results compare well with those of Plakans et al. (2019), who found that CAF measures predicted 46% of the variance of scores in TOEFL iBT reading-listening-writing tasks. In fact, our explained variance in the viewing-to-describe task was also 46%, although we did not include word count as a proxy for writing fluency, which accounted for 25% of the variance in Plakans et al.'s (2019) study. Furthermore, the variance explained by the fixed effects in the viewing-to-compare-and-contrast task in our study was even higher,

that is, 60%, which suggests that the selected linguistic measures are better indicators of writing quality in this task type than in viewing-to-describe tasks.

We note, however, that not all our originally selected measures were predictive of writing scores. Namely, neither the frequency of connectives nor latent semantic analysis (sentence) correlated with the total viewing-to-write score. This, in fact, reflects prior research; Plakans and Gebril (2017) also observed no statistically significant differences in the use of connectives across writing levels, and Chan and Yamashita's (2022) meta-analysis of how linguistic features are related to scores in integrated tasks similarly demonstrated limited predictive power of cohesion measures. As the viewing input in our tasks might have provided a coherent structure for the writing (as also suggested by the qualitative data in Brunfaut & Kormos, 2025) and as the notetaking enabled this further, it is possible that linguistic markers of cohesion contributed little to raters' perceptions of task performance.

Measures of lexical frequency that were established based on the COCA Academic corpus correlated weakly and significantly with the total viewing-to-write score in both task types but were not found to be predictive of performance in the presence of other lexical, syntactic, and accuracy variables in the mixed effects model. In independent writing tasks, higher-rated texts and more proficient writers tend to use lower-frequency words (e.g., Crossley & McNamara, 2012; Garner et al., 2020; Guo et al., 2013; Zhang & Ouyang, 2023), but in Chan and Yamashita's (2022) meta-analysis, the joint effect of lexical sophistication and diversity measures on scores was found to be weak. This might be because, unlike in independent writing tasks, in integrated reading/listening/viewing-to-write tasks, L2 writers need to express information predetermined by the input, and they therefore generally have less scope to tailor the content of their text to their existing lexical resources. Furthermore, in most integrated task types, a considerable proportion of the required lexical units is provided in the input and is required for successful task completion, which might explain why frequency-based measures of lexical sophistication might not show substantial variation across higher and lower rated task performances, especially when most test-takers are at upper-intermediate and advanced proficiency levels already.

In contrast to the lexical frequency measures, lexical diversity was a moderately strong predictor of viewing-to-write scores in the correlation analysis and an important variable in the mixed effects models for both our task types. This reflects Gebril and Plakans' (2016) results of significant correlations between lexical diversity and writing scores in integrated reading-listening-writing TOEFL iBT tasks. Also, in qualitative post-task interviews reported in Brunfaut and Kormos (2025), several participants mentioned having edited their text to avoid lexical repetition and to enhance the variety of lexical choices. Therefore, we might hypothesize that our higher ability L2 writers strategically paid attention to lexical

diversity, and this was also apparent in the significant relationship between raters' evaluations of overall task performance and lexical diversity.

Interestingly, the role of lexical retrieval speed, another aspect of lexical sophistication that we captured using factor scores, was only significant in the viewing-to-compare-and-contrast task. Our compare-and-contrast task type shares some genre and task characteristics with the TOEFL iBT integrated reading-listening-writing task examined by Kim and Crossley (2018), who also showed that lexical decision-based time measures are important predictors of essay ratings. Therefore, it is possible that the use of words that take longer to retrieve from the mental lexicon might play a different role in L2 writing quality depending on task type and genre demands, as well as the input for the writing task. These potential explanations for this distinct finding between the compare-and-contrast and describe task types also seem to align with the CEFR (Council of Europe, 2001; 2020), where the ability to describe already features prominently in B1 written production descriptors (e.g., "Can produce clear, detailed texts on a variety of subjects related to their field of interest, synthesizing and evaluating information and arguments from a number of sources;" "Can give straightforward, detailed descriptions on a range of familiar subjects within their field of interest"), whereas the ability to synthesize different viewpoints and arguments appears from the B2-level onwards (e.g., "Can produce clear, detailed texts on a variety of subjects related to their field of interest, synthesizing and evaluating information and arguments from a number of sources;" "Can synthesize information and arguments from a number of sources"). Combined with the fact that most of our participants were above B1, one might expect processing speed aspects to be more uniform within the participant group for the B1-associated function of describing. The B2-associated function of synthesizing, including comparing and contrasting, is instead more likely to involve different processing speed demands between lower versus higher B2 participants in the group.

The accuracy factor values, representing the frequency of grammatical, spelling, and punctuation errors, were one of the strongest predictors of ratings in both task types. This finding is in line with the results of Cumming et al. (2005), Gebril and Plakans (2013), and Plakans et al. (2019), who established that the frequency of errors had considerable predictive power in integrated writing tasks.

The contribution of syntactic complexity at the clause level was found to vary across the two types of tasks in our study. In the compare-and-contrast task, clausal complexity was not associated with the ratings, whereas in the describe task, it was a significant predictor. Previous meta-analyses by Chan and Yamashita (2022) for integrated tasks showed a small but significant effect of syntactic complexity on writing quality. In Kim and Crossley (2018), clause-based complexity measures were not directly associated with ratings of integrated reading-

listening-writing TOEFL iBT tasks, which are similar in genre to our compare-and-contrast task. Interestingly, in our study, the ratio of dependent clauses per clause did not differ across the two task types. Combined with the findings on syntactic complexity at the clause level, this suggests that in the compare-and-contrast task, the input and task demands elicited similar clausal complexity regardless of L2 writing ability, and variation in clausal complexity was not perceived to be a marker of better performance. In contrast, in the describe task, the use of more complex clauses might have been less determined by task design features and therefore might have been a more relevant predictor of higher-level L2 writing skills.

6. Conclusion

In our research, we examined the linguistic characteristics of task performance in two types of integrated multimodal viewing-to-write tasks and analyzed how these measures predict human ratings. Our study is one of the first projects to have investigated the textual features of integrated multimodal writing tasks and the contribution of these features to perceptions of writing quality. The results showed that the two types of tasks, one that required learners to describe a process presented in a video input and the other that required learners to compare and contrast the views of two speakers on a topic in a videocast, elicited similar output in terms of lexical and syntactic complexity, lexical sophistication and diversity, cohesion, and accuracy. These findings mirrored the results of our previous study (Brunfaut & Kormos, 2025) in which the scores human raters awarded did not differ significantly either. Thus, our results indicate that, in task-based assessment and teaching contexts, both types of tasks might be effectively used to evaluate and develop B2 to C1 level secondary school L2 learners' integrated viewing-to-write skills. Among the linguistic measures, accuracy and lexical diversity were important predictors of performance in both task types, whereas the role of lexical sophistication and syntactic complexity varied across the two task types.

From a pedagogical perspective, our findings highlight that even though overall linguistic characteristics of task performance are similar in the two task types, higher- and lower-performing learners use different lexical items and syntactic structures in them depending on task demands. Therefore, it is important to raise L2 learners' awareness of these linguistic features and apply genre-based pedagogies to teach L2 writers how to use genre-specific linguistic structures in their essays. The results also showed that raters assign higher scores to essays with increased levels of lexical diversity and accuracy regardless of task type. Therefore, teachers should develop L2 learners' revision strategies and encourage L2 writers to focus on these aspects of their writing during composing

and editing. Checklists and self-assessment guidelines can also be beneficial for improving these linguistic aspects of L2 writing. From the viewpoint of assessment, it is necessary to consider these results in rater training and direct raters' attention to the task-relevant features of language use when evaluating performance. The findings suggesting key differences in the contribution of lexical sophistication and syntactic complexity between the two task types are also relevant for developing automated rating systems using large language models. In line with the findings of our previous study (Brunfaut & Kormos, 2025), the results of this research confirm that the two multimodal task types elicit similar performance and, because of their similar cognitive load, can be effectively used for instructional and assessment purposes.

Our study is not without limitations. We only examined two types of integrated multimodal writing tasks, and the participants were recruited from one age group, L1 background, and context, and included L2 learners mostly from the B1 level and higher. Therefore, further research is needed to investigate other types of integrated multimodal writing tasks in different contexts, potentially involving lower proficiency and younger learners. Integrated multimodal language uses are also frequent in academic and professional contexts, and more research is required to understand their potential for supporting L2 skills development and using such tasks for formative and summative assessment in those contexts.

Acknowledgements

The work was supported by the British Council through an Assessment Research Grant 2022. The British Council does not discount or endorse the methodology, results, implications, or opinions presented by the researchers.

References

Abrams, Z. I. (2019). The effects of integrated writing on linguistic complexity in L2 writing and task-complexity. *System, 81*, 110-121. https://doi.org/10.1016/j.system.2019.01.009

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing, 31*(2), 241-259. https://doi.org/10.1177/0265532213509810

Berger, C. M., Crossley, S. A., & Kyle, K. (2019). Using native-speaker psycholinguistic norms to predict lexical proficiency and development in second-language production. *Applied Linguistics*, *40*(1), 22-42. https://doi.org/10.1093/applin/amx005

Brodkey, D., & Young, R. (1981). Composition correctness scores. *TESOL Quarterly, 15*(2), 159-167. https://doi.org/10.2307/3586407

Brunfaut, T., & Kormos, J. (2025). Assessing multimodal viewing-to-write constructs: Task design, performance, processing, and rating. *Language Assessment Quarterly*, *22*(4-5), 429-459. https://doi.org/10.1080/15434303.2025.2596374

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Chan, S., & May, L. (2023). Towards more valid scoring criteria for integrated reading-writing and listening-writing summary tasks. *Language Testing*, *40*(2), 410-439. https://doi.org/10.1177/026553222211350

Chan, S., & Yamashita, J. (2022). Integrated writing and its correlates: A meta-analysis. *Assessing Writing, 54*, 100662. https://doi.org/10.1016/j.asw.2022.100662

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment: Companion volume.* Council of Europe Publishing.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35*(2), 115-135. https://doi.org/10.1111/j.1467-9817.2010.01449.x

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*(1), 5-43. https://doi.org/10.1016/j.asw.2005.02.001

Garner, J., Crossley, S., & Kyle, K. (2020). Beginning and intermediate L2 writer's use of N-grams: An association measures study. *International Review of Applied Linguistics in Language Teaching*, *58*(1), 51-74. https://doi.org/10.1515/iral-2017-0089

Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly, 10*(1), 9-27. https://doi.org/10.1080/15434303.2011.642040

Gebril, A., & Plakans, L. (2016). Source-based tasks in academic writing assessment: Lexical diversity, textual borrowing and proficiency. *Journal of English for Academic Purposes, 24*, 78-88. https://doi.org/10.1016/j.jeap.2016.10.001

George, D., & Mallery, M. (2010). *SPSS for Windows step by step: A simple guide and reference, 17.0 update* (10th ed.). Pearson.

Guo, Q., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18*(3), 218-238. https://doi.org/10.1016/j.asw.2013.05.002

Halliday, M., & Hasan, R. (1976). *Cohesion in English*. Longman.

Holzknecht, F., & Harding, L. (2024). Repeating the listening text: Effects on listener performance, metacognitive strategy use, and anxiety. *TESOL Quarterly*, *58*(1), 451-478. https://doi.org/10.1002/tesq.3249

Jung, Y., Crossley, S. A., & McNamara, D. S. (2015). Linguistic features in MELAB writing performances (Working Paper No. 2015-05). http://www.cambridgemichigan.org/wp-content/uploads/2015/04/CWP-2015- 05.pdf.

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

Kojima, M., & Kaneta, T. (2022). L2 writing and its internal correlates: A meta-analysis. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (pp. 109-158). John Benjamins.

Kormos, J., & Suzuki, S. (2024). Creativity and the linguistic features of argumentative and narrative written task performance. *System*, *127*, 103531. https://doi.org/10.1016/j.system.2024.103531

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1-26. https://doi.org/10.18637/jss.v082.i13

Mizumoto, A. (2025). Automated analysis of common errors in L2 learner production: Prototype web application development. *Studies in Second Language Acquisition*, *47*(3), 867-884. https://doi.org/10.1017/S0272263125100934

Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics*, *3*(2), 100116. https://doi.org/10.1016/j.rmal.2024.100116

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Peng, J., Wang, C., & Lu, X. (2018). Effect of the linguistic complexity of the input text on alignment, writing fluency, and writing accuracy in the continuation task. *Language Teaching Research*, *24*(3), 364-381. https://doi.org/10.1177/1362168818783341

Phillips Galloway, E., Qin, W., Uccelli, P., & Barr, C. D. (2020). The role of cross-disciplinary academic language skills in disciplinary, source-based writing: Investigating the role of core academic language skills in science summarization for middle grade writers. *Reading and Writing, 33*(1), 13-44. https://doi.org/10.1007/s11145-019-09942-x

Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing, 17*(1), 18-34. https://doi.org/10.1016/j.asw.2011.09.002

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing*, 22(3), 217-230.

Plakans, L., & Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing Writing, 31*, 98-112. https://doi.org/10.1016/j.asw.2016.08.005

Plakans, L., Gebril, A., & Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing, 36*(2), 161-179. https://doi.org/10.1177/0265532216669537

R Core Team. (2021). *R: A language and environment for statistical computing* (4.0.2). R Foundation for Statistical Computing.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics, 43, 1-32.* https://doi.org/10.1515/iral.2005.43.1.1

Schoonen, R., Van Gelderen, A., Stoel, R. D., Hulstijn, J., & De Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning, 61*(1), 31-79. https://doi.org/10.1111/j.1467-9922.2010.00590.x

Skehan, P., & Foster, P. (2012). Complexity, accuracy fluency and lexis in task-based performance: A synthesis of the research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 199-220). John Benjamins.

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 Years later. *Educational Psychology Review*, *31*(2), 261-292. https://doi.org/10.1007/s10648-019-09465-5

Xu, L. (2019). Noun phrase complexity in integrated writing produced by advanced Chinese EFL learners. *Papers in Language Testing and Assessment, 8*(1), 31-51.

Yoon, H. J., & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*, *51*(2), 275-301. https://doi.org/10.1002/tesq.296

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47*, 100505. https://doi.org/10.1016/j.asw.2020.100505

Zhang, Y., & Ouyang, J. (2023). Linguistic complexity as the predictor of EFL independent and integrated writing quality. *Assessing Writing*, *56*, 100727. https://doi.org/10.1016/j.asw.2023.100727

APPENDIX

Correlations among variables

Correlations among the total viewing-to-write scores and linguistic variables in the describe task

Variable	2	3	4	5	6	7	8	9	10
1. Total V to W score	.283**	.136	-.604**	.093	.072	-.196*	-.208*	-.148	.196*
2. Lexical diversity factor		.170*	-.134	.031	.016	-.063	-.209**	-.043	.308**
3. Lexical retrieval speed factor			-.373**	-.030	.041	-.388**	.021	-.477**	.063
4. Accuracy factor				.044	.092	.207**	-.032	.306**	-.013
5. Frequency of all connectives					.114	-.114	-.079	-.033	-.057
6. Latent semantic analysis (sentence)						-.075	-.078	-.087	-.057
7. COCA Academic frequency							.149	.587**	.060
8. COCA Academic bigram frequency								.023	-.002
9. MRC familiarity									.071
10. Dependent clause per clause									

Note. * $p < .05$; ** $p < .01$ (2-tailed)

Correlations among the total viewing-to-write scores and linguistic variables in the compare-and-contrast tasks

Variable	2	3	4	5	6	7	8	9	10
1. Total V to W score	.561**	.630**	-.697**	.048	-.139	-.327**	-.361**	-.518**	.152
2. Lexical diversity factor		.432**	-.500**	-.006	-.130	-.165*	-.464**	-.405**	.224**
3. Lexical retrieval speed factor			-.518**	.016	-.060	-.505**	-.205**	-.693**	.105
4. Accuracy factor				-.018	.139	.195*	.326**	.419**	-.203*
5. Frequency of all connectives					-.015	-.135	.039	-.129	-.138
6. Latent semantic analysis (sentence)						.021	.104	.037	-.114
7. COCA Academic frequency							.170*	.593**	.153
8. COCA Academic bigram frequency								.167*	-.026
9. MRC Familiarity								1.00	-.028
10. Dependent clause per clause									

Note. * $p < .05$; ** $p < .01$ (2-tailed)