RESEARCH

Open Access

The effect of using online language-support resources on L2 writing performance

Dongkwang Shin¹, Suh Keong Kwon^{2*} and Yongsang Lee³

* Correspondence: sokekwon@cue.ac.kr

²Chinju National University of Education, Jinju, Republic of Korea
Full list of author information is available at the end of the article

Abstract

Language learners today have access to a much wider choice of resources and language-support tools that aid their writing processes. While research on the convenience and flexibility of the writing resources that learners use in real-life writing activities have argued for a re-conceptualization of writing ability, little is known about the effect of using them in a writing assessment setting. Accordingly, the present study aimed to examine whether L2 learners' use of writing resources in a writing test might have an effect on their writing performance. Furthermore, we also investigated how the effects of the writing resources might vary by test takers' writing proficiency and scoring domains. Results showed that the group who had access to writing resources outperformed the group who were not given the online resource, but both groups' scores were within the same scoring descriptor. The significant score improvement was more evident in the low- and intermediate-level learners. In a questionnaire, test takers reported that they normally use the writing resources in real-life writing situations and also found the online resource features helpful in improving their writing performance and agreed upon introducing such features in future writing tests.

Keywords: Second language writing assessment, Authentic writing task, Language-support resources, Online dictionary, Concordancer

Introduction

With the development of technology, a substantial amount of writing activities are now completed on computers, and this is more evident among learners in this generation regardless of their backgrounds. In fact, a recent report from *The Wall Street Journal* showed that 85% of elementary school students (grades 3–5) in the USA were comfortable with the computer-based writing mode, and approximately 80% of them also found computer-based tests easy (Porter, 2015). The ability to use digital devices as a communication means is considered an important component of digital literacy, which is a required competence in the fourth industrial revolution. The American Library Association (ALA, 2019) defines digital literacy as “the ability to use information and communication technologies to find, evaluate, create, and communicate

information, requiring both cognitive and technical skills". However, digital literacy can be defined differently depending on the digital means being used, and there is no single overarching definition of it (Meyers et al., 2013). Digital literacy for today's writing activities include skills to use digital technology to search and evaluate given resources to find appropriate, relevant, and valid information to create and communicate through the writing process.

Recent studies have shown that the underlying construct of foreign language assessment is to communicate and exchange information with others (Bachman, 2007; Bachman & Palmer, 2010). Bachman and Palmer's (1996) communicative language ability model indicates that language users' use of topical knowledge and language knowledge is associated with their strategic competence, which is a required ability in L2 communication. With regard to this, Bachman and Palmer (1996) introduced the test usefulness model and argued that the language tests should examine the learner's communicative language ability required in the target language use domain (*TLU domain*, hereafter) (Bachman & Palmer, 2010; Banerjee, 2019). It is, however, difficult for a test to fully reflect real-life language use tasks in a variety of contexts because it is also difficult to meet the individual variability of learners in terms of their background knowledge or experience. Therefore, test developers need to identify what language use tasks are essentially required in our daily lives, not only in terms of the content of the tasks but also in terms of the means of completing the tasks.

Given that potential test takers, particularly the current generation, are more accustomed to using a word processor on a computer for writing, it is reasonable to provide the answer mode of a language test on a computer rather than the traditional paper-based mode. Studies have constantly been arguing for a shift of the writing test mode from paper to computer, and many conventional language tests that include a writing section have already shifted the mode of test delivery to computer-based. Studies also have shown that a change in the writing medium has an effect on test takers' writing process and also the quality of writing outcomes (Brunfaut et al., 2018; Van Waes & Schellens, 2003). Winke and Isbell (2017) argue that computer-assisted language tests "facilitate, contextualize, and enhance" assessment of candidates' linguistic abilities (p. 313). Therefore, more studies are required to examine the extent to which changes in test mode affect candidates' test-taking processes and performance. In particular, if we are interested in developing authentic writing tasks that emulate those required in a real-life writing situation, it is necessary to better understand what specific features in computer-based writing activities should be included in a writing test. Simply shifting the delivery mode and input method into the computer-based model is not sufficient, as real-life computer-based writing involves much more complex and flexible processes. We believe using an online writing resource is one of the processes, and this study seeks to investigate the extent to which providing such a feature in a writing task has an effect on L2 learners' writing performance.

Writing, itself, is an asynchronous activity and is not restricted by a particular time and space. Real-time writing activities such as writing an e-mail, keeping a diary, or creating documents for work are good examples of the asynchronous nature of writing (Joo & Shin, 2012). However, the assessment of writing competence so far has been normally conducted in a controlled environment in which no external resources are available to test takers in a timed task. In reality, however, writing can be done at any

time, and referring to writing resources such as dictionaries or web-based corpora is a prevailing strategy to complete writing tasks more effectively. In fact, we hypothesize that the use of these references makes it possible to achieve qualitatively better writing output than writers' actual writing competence ($i + 1$) (Krashen, 1981). By using writing resources such as a dictionary, a term bank, corpora, or other language-support tools, the writer is likely to produce fewer errors, especially in grammar, spelling, and vocabulary selection. Therefore, to help ensure the validity of the assessment of writing, actual writing competence definitely needs to be measured in a more natural environment.

Literature review

Writing is a complex mechanism, and the process of writing is an important indicator that distinguishes skilled writers from unskilled writers (Weigle, 2005). While speaking is a spontaneous process, writing is a recursive process which allows writers to refer to the produced output and modify it (Hayes & Flower, 1981). In this regard, the cognitive development of writing skill is largely categorized into three stages, "knowledge telling" (creating and generating what the writer wants to say), "knowledge transforming" (changing what the writer wants to say), and "knowledge crafting" (shaping what to say and how to say it) (Bereiter & Scardamalia, 1987; Kellogg, 2008, pp. 6-7). Among these, knowledge transforming is particularly related to the topic of the present study, using online resources in L2 writing, because such feature triggers writers to carry out additional planning, translating, and reviewing processes. We speculate that the presence of online resources aids writers to use the knowledge transforming process by providing an access to external source of knowledge as well as language reviewing features.

Using writing resources in teaching and testing writing

The idea of using references when writing has often been raised. However, the main reason for using reference materials argued by related studies (e.g., Gaskell & Cobb, 2004; Lee et al., 2009; O'Sullivan & Chambers, 2006; Park et al., 2015; Yoon, 2016; Yoon & Hirvela, 2004) is to improve the quality of writing output, not to redefine fundamental writing competence. Early research on the reference consultation during writing focused on examining the impact of the using dictionary or corpus on L2 learners' writing output or the affective domain (Barnes et al., 1999; East, 2007). The use of reference materials in writing has naturally increased demand as online resources become more accessible and available today. However, little is known about whether having access to these resources might provide an accurate measure of test takers' writing ability in an assessment setting.

A number of existing studies have specifically looked at the nature of writing-resource use in writing practices. Frankenberg-Garcia (2005) categorized five strategies of using writing resources as (1) finding an L2 equivalent, (2) confirming a hunch, (3) finding a suitable collocate, (4) choosing the best alternative, and (5) checking spelling (p. 341). In a follow-up study, Frankenberg-Garcia (2011) also investigated how Portuguese learners of English used writing resources when translating and finding errors in the given texts and found that the most frequently used features were term banks, dictionaries, and collocation checkers. Participants in her study also used other features

such as language corpora or Internet search engines. However, the tasks used in Frankenberg-Garcia's (2011) study were not given in an assessment setting.

Yoon and Hirvela (2004) investigated students' perceptions of the advantages and disadvantages of using a corpus as an aid for writing activities in ESL academic writing courses at a large Midwestern American university. Approximately 82% of the participants in the courses were from Asian countries, and all participants were asked to consult the Collins COBUILD Corpus sampler, particularly about the actual use of the target vocabulary and grammar in their writing drafts. The results showed that most students had a positive attitude towards using a corpus. They thought the corpus search helped the students check and acquire authentic lexical or grammatical patterns although there was little change in their actual grammar score.

Gaskell and Cobb (2004) examined the effectiveness of corpus consultation in reducing learners' writing errors with 20 adult Chinese lower-intermediate ESL learners at a university in Canada. The participants tried to correct grammatical errors in their first writing drafts by referring to concordances from a corpus. The results reported that 50% of the participants showed some improvement in grammar, which was attributed to their better understanding of the grammatical structure of sentences through various exposures to concordance samples from the corpus.

O'Sullivan and Chambers (2006) conducted an experiment with English native students at the University of Limerick. The participants were asked to correct errors by referring to concordance samples during the process of writing in French. The researchers compared the results of this concordance referencing with those of referring to dictionaries and grammar materials while writing to examine the effect of corpus consultation. The participants also reported that referring to concordance samples was helpful in terms of the use of prepositions and word choice.

An original attempt was made by Park et al. (2015), who developed an error corpus by putting error tags into a Korean learner corpus consisting of essays by 300 Korean high school students (100 high-level, 100 mid-level, and 100 low-level). This error corpus identified grammatical errors that Korean learners frequently make along with corrected samples of those errors. Learners then chose an error scheme from the error category, and the tool showed sample sentences containing the target error type and the revised sentences corrected by a native speaker of English.

In a recent study, Oh (2020) investigated the nature of L2 learners' use of writing resources and examined the difference in their writing performance with and without an online writing resource feature. The study found that the spell-check function was used the most by the test takers, followed by a dictionary, a search engine, and a web-based translator. The use of these functions led to the improvement of their writing test scores across all scoring domains (content, organization, vocabulary, grammar, and appropriateness), which suggests that the use of such writing resources aid test takers to produce better quality writing. It is important to note that the test-takers in this study were instructed to use the writing resources that they would use in real-life writing tasks, and the majority of participants found the resources useful and were satisfied with the experience of using them.

Yoon (2016) looked into what L2 learners naturally do with the online reference resources while completing independent authentic writing tasks. With two ESL graduate students, Yoon (2016) conducted case studies and identified that there are wide

individual differences in the use of the reference resources. Findings of the study suggested that such wide differences are attributed by interactions among various factors associated with personal character, text, and context. However, the study is limited in that the case study findings are hard to generalize, and the writing tasks used in the study were independent free writing which were not given in a test setting.

A number of recent studies have also tapped into the similar issue but on different aspects of the computer-based writing instructions. Hajimaghsoodi and Maftoon (2020) provided a computer-assisted writing platform to 67 EFL learners and found that the students made a significant improvement in their writing through interactions and collaborations on the online platform. Rashtchi and Porkar (2020) investigated the extent to which integrating technology and brainstorming might affect EFL learners' writing skills. They showed that an online brainstorming tool, 'Wordle' promoted the participants' writing performance, largely because the word clouds stimulated the students' background knowledge, which led to useful ideas required for composing essays. However, more studies are required that specifically investigate whether exposure to other kinds of external sources, for example the internet search engines, might enrich the 'content' of the writing. Sun and Hu (2020) examined the effect of data-driven learning (DDL) on the use of lexicogrammatical resource, hedges, in writing and showed that the participants used hedges more frequently after the DDL treatment. In particular, the participants of their study reported that they used online dictionaries the most, and also found them helpful in learning to use the hedges more appropriately in their writings. However, Sun and Hu (2020) only examined the effect of DDL approach on the learners' language use aspects, specifically on the use of hedges. Little is yet known about what other elements of language use, such as collocations or grammar uses, can also be improved through the DDL approach.

Frankenberg-Garcia (2020) recently introduced a lexicographic tool, ColloCaid, which provides 'real-time help' to writers through automated features such as autocorrect, synonym suggestions, and finding collocations. She also showed that the new writing assistant tool, ColloCaid, provides necessary information on collocation to writers as they write without distracting their writing processes. Emphasizing the innovative features of the ColloCaid, Frankenberg-Garcia (2020) argued that bringing the dictionary to the writers instead of teaching them how to use one is an ideal way to create a digital writing environment. While this new project is still a work-in-progress, the present study presents relevant evidence on how such 'real-time' writing support resource, such as the Concord Writer which is used in this study, may affect L2 learners' writing processes and performance.

While the usefulness of writing resources and corpus consultation has frequently been studied, gaps remain in establishing a consensus on whether the use of writing resources should be a part of writing test constructs. In this regard, Gruba (2014) claimed that the Internet resources (e.g., online encyclopedias, search engines, blogs, etc.) are widely used in language classrooms, but their use in language assessment remains tentative large because of test fairness issue.

Reconceptualization of computer-integrated writing assessment

In the past, studies have used various terms when referring to computer-based test (e.g., computer-adaptive, computer-assisted, computer-based, computer-supported,

etc.), and these terms generally imply that a computer is being used as a supporting device that aids or enhances demonstrations of test takers' language ability (Chapelle & Douglas, 2006; Douglas, 2013). However, as we reconceptualize the role of computers in language testing as suggested by the studies mentioned above (Chalhoub-Deville, 2010; Chapelle, 2010; Chapelle & Douglas, 2006; Douglas, 2013; Yu & Zhang, 2017), it is essential that we emphasize the integrative function of computer technology. The importance of integrating computer technology to reconceptualize language ability is well documented; for example, Chapelle and Douglas (2006, p. 107) claimed that a new concept of language skill is required that represents "the ability to select and deploy appropriate language through the technologies that are appropriate for a situation". Douglas (2013, p. 2) also argue that we need to redefine "language construct to include appropriate technology in light of the target situation and test purpose". This new concept of computer integrative function should offer more flexible processes for accessing a variety of resources when completing a writing task. Online writing resources, such as web-based language-support tools, are ideal examples to apply this new concept in future writing tests when test takers' digital literacy is considered as part of their language ability. Therefore, as recommended by Yu and Zhang (2017), this study will use the term *computer-integrated* instead of *computer-based*.

Computer technology in language assessment has been treated as a tool or method for test delivery, scoring, or reporting test results (Sawaki, 2012). However, there has also been a growing concern about introducing computer technology in high-stakes language tests, mostly due to potential threats to the validity of computerized tests; existing studies (Huff & Sireci, 2001; Jin, 2012; Li, 2006) have argued that test familiarity or test takers' perceptions of the computerized test could be confounding variables that may cause a test to measure unwanted construct-irrelevant skills.

On the other hand, a number of studies have also argued for a reconceptualization of computerized language testing (Chalhoub-Deville, 2010; Chapelle, 2010; Chapelle & Douglas, 2006; Douglas, 2013; Jin & Yan, 2017; Yu & Zhang, 2017). Chapelle (2010) introduced three motives for using computer technology in language testing: (a) efficiency, (b) equivalence, and (c) innovation. In this regard, Chapelle (2010) claimed that computer technology can improve efficiency in test development and administration and has better comparability in test performances between computer-based and paper-based tests. Chapelle and Douglas (2006) argued that reconceptualized computer technology in language testing should define a learner's ability as "the ability to select and deploy appropriate language through the technologies that are appropriate for a situation" (p. 107). Douglas (2013) added that the newly conceptualized language construct should include technology-related skills that are required in the TLU domain and meet the purpose of a test. Douglas (2000) also warned that future transformations and innovations in computerized language tests should consider "technology being employed in the services of language testing" rather than a test being driven by technology (p. 275).

Careful considerations are also warranted since studies have also argued that the cognitive processes involved in writing on a computer are different from writing on paper, which is a major threat to the construct validity of a computer-integrated writing test. In Chen et al. (2011), for example, it was found that computer-based writing tests disadvantaged a particular group of economically disadvantaged test takers. In this regard, Winke and Isbell (2017) warned test developers to understand that the construct of a

writing test can be affected when the test mode is changed from paper to computer. To avoid such a threat, it is required to match the way the writing ability is tested with the way the test takers are instructed in language classrooms.

Taken together, these studies suggest that little is yet known about the impact of on-line writing resources on L2 learners' writing process and performance although the use of these functions prevails in today's real-life writing activities. A number of studies have examined the effect of using online resources to improve writing performance, but they focused on somewhat narrow aspects such as brain storming (Al-Shehab, 2020; Hajimaghsoodi & Maftoon, 2020; Rashtchi & Porkar, 2020) or use of hedges in writing (Sun & Hu, 2020). Only few studies looked into the nature of EFL learners' use of writing resources and its effect on their writing performance by specific scoring domains and proficiency (Oh, 2020). Therefore, decisions on whether or not to include such functions in an L2 writing test still need more evidence to establish a consensus among language test developers and associated researchers. With this in mind, the present study addresses the following research questions:

Research questions

1. To what extent do L2 test takers perform differently in a writing test when access to writing resources is provided to them?
2. To what extent does the effect of writing resource use in a writing test vary by scoring domain and by test-taker proficiency?
3. How do test takers perceive the use of writing resources in real writing activities and in a writing assessment setting? Is this preference in the use of writing resources associated with their writing performance?

Method

Participants

A total of 50 students from a teacher's college in Korea participated in this study. Of these students, 10 were male, and 40 were female, and they were all 3rd-year pre-service elementary teachers who were majoring in English education at the time they participated in this study. The participants were sampled for practical reasons because they were relatively proficient in English and took the academic writing course as part of the teacher's college curriculum. All participants were Korean, but 36% of them had experience of living or visiting English-speaking countries, while 64% of them had no experience. Essays were collected from three intact classes of English writing course at the university. Their estimated English proficiency level corresponds to approximately level B1 or B2 of the CEFR on average.

Instruments

Writing prompts

The participants were given two essay-writing tasks on different topics for each test. While the first test, non-referenced writing (NRW), was given without an access to writing resources, the second test, referenced writing (RW), was provided with writing resources accessible to the test takers. The following table (Table 1) presents the two essay topics given to the participant and a sample prompt of one of the topics.

Table 1 Writing topics and a sample prompt

- Topic 1: Advantages or disadvantages of eating breakfast				
- Topic 2: Advantages or disadvantages of using the Internet				
Read the following statements and reasons for the advantages and disadvantages of eating breakfast/using the Internet. Choose one of the two positions, and complete an essay that contains all three parts: introduction, body, and conclusion. In addition to the two reasons provided under each position, you need to add one more reason of own in the body paragraph. In total, provide three reasons for your argument with appropriate supporting ideas. (Word limit: 400 words)				
Introduction	Advantage	1) wake up brain in the morning	2) healthy diet plan	Conclusion
	Disadvantage	1) take more time	2) gain weight	

The participants were provided with the two reasons for each statement to reduce the cognitive burden of the participants and also to robustly focus more on linguistic features. An initial baseline writing test was also conducted before the two main tests, and an equivalent essay topic was given to all participants: “advantages and disadvantages of installing CCTVs in school”. These micro-genre writing tasks are frequently used essay topics in many EFL countries including Korea and Japan (Watanabe, 2016). According to the classification of the micro-genres introduced by Martin and Rose (2005), all tasks used “expositions” genre which asks test takers to persuade readers.

Online writing resources

In this study, we provided Concord Writer (Cobb, 2019) as a writing resource during the referenced writing test (RW test) (See Fig. 1). Concord Writer is an online tool consisting of three main features: a writing pad, a concordancer, and a dictionary. The writing pad is a collocation/usage dictionary where users type in any words or sentences and check appropriate use of the expression by double-clicking the target word. The concordancer searches for and lists up to a maximum of 100 usages of the selected keyword. If the concordancer fails to find an appropriate sample usage of the selected word, users can look up the keyword in one of the nine online dictionaries or one thesaurus that are built into Concord Writer. The dictionary search function provided by Concord Writer, a writing tool applied to the experimental writing activity of this study, provides eight bilingual dictionaries including English–Korean (L2–L1), and one monolingual (e.g. English–English) dictionary. However, it does not provide a Korean–English (L1–L2) dictionary. Although employing the Internet search engines (e.g., Google, Wikipedia) was encouraged in a recent relevant study (Oh, 2020), we did not provide this function in the present study in order to concentrate on examining the effects on linguistic features only.

Scoring rubric

For scoring the test takers’ essays, this study used Shin et al.’s (2012) scoring rubric, which was developed for classroom-based English writing practice and assessment in

Fig. 1 Concord Writer screenshot

South Korea. The rubric consists of the four rating domains of Task Completion, Content, Organization, and Language Use (Appendix). This scoring rubric (Shin et al., 2012) was originally designed to rate high school students' writing competence by non-native raters in Korea, and was developed by Korea Institute for Curriculum and Evaluation, a government-funded research institution. The ranges applied in the scoring rubrics are flexible and can be applied to various proficiency target test takers. Thus, the scoring rubric is the most suitable one to assess the participants of this study, the university students, because it is also the one that is most familiar to the participants of this study. The participants of the present study were also instructed that the writing tasks they perform would be rated under the four-scoring rubric.

Task Completion on the scoring rubric was originally benchmarked from Task 1 of the IELTS writing scoring rubric's Task Performance. Most of the existing writing scoring rubrics (e.g., Jacobs et al., 1981) only have Content as a domain related to the target task topic. Task Completion is different from Content, which deals more with the supporting ideas of the writer's argument. If a test taker's supporting ideas are very robust but the details digress from the target issue, the test taker would receive a low score in Task Completion. For example, a student can receive a low score on Task Completion if the essay is slightly off-topic, regardless of how robust the content is. Therefore, Task Completion is considered an umbrella-scoring domain; that is, the score of the other three domains apart from Task Completion could not exceed the score of Task Completion. This was intended to prevent the test takers from cheating by memorizing a full script on an irrelevant topic and receiving high scores in the other three domains. For each scoring domain, 5 scales were applied with intervals of 5 points each (maximum score: 25), and the total score was the sum of the four domains (maximum score: 100).

Questionnaire

A short questionnaire was used to collect the participants' demographic information (e.g., gender, experience of living in foreign countries, etc.) and perspectives on using

writing resources in a writing test. Questions included the participants' past experiences using writing resources in real-life writing activities, preference in the use of writing resources, and opinions about using the functions in a writing test. Apart from the questions for demographic information, there were a total of five items in the questionnaire. The participants answered these questions immediately after they finished the RW test.

Data collection procedure

A total of 50 participants were sampled from two intact classes, and all of them were given two writing tasks for this study. Prior to the first test, a diagnostic test was administered as a baseline test to distinguish the participants' writing proficiency into three different levels (i.e., low, intermediate, high) at the initial stage. Following the baseline test, two essay-writing tests were given in 1 week's time to examine the effects of using writing resources on writing performance. The first test was given in a referenced writing condition (RW), and the other test was given in a non-referenced writing condition (NRW). To minimize the possibility of test takers' topic familiarity or group characteristics (proficiency, topic, practice effect) playing confounding roles, we counterbalanced the essay topics given to each group.

An equivalent essay topic, the installation of CCTVs in schools, was given to all participants for the initial baseline test, and access to writing resources was not allowed this time. In 1 week's time, the NRW and the RW tests were carried out on the same day (in the morning and afternoon classes), consecutively. In the NRW test, an essay topic on eating breakfast was given to the participants in Class A ($n = 35$), and a topic on using the Internet was given to participants in Class B ($n = 15$); accessing the writing resources was not allowed. In the RW test, the topic on using the Internet ($n = 35$) was given to Class A, and the topic on eating breakfast was given to Class B ($n = 15$). Each writing test was conducted for an hour, but most of the participants spent 30–40 min to complete the task, and only few (2 or 3 of them) spent the entire hour. A timer was not used during the test to prevent the test takers from paying too much attention to it, more than the writing resources. This policy was also intended to encourage the test takers to complete the writing tasks as naturally as they can without any pressure. We also presumed that 1 h was far enough to complete the given tasks. An access to the writing resources on Concord Writer was fully available in the RW test. At the end of the RW test, a questionnaire on perceptions and usage of the writing resources was given to all participants.

The following figure (Fig. 2) briefly illustrates the procedure for administering the writing exams.

Data analysis

For scoring, two raters, the first and corresponding authors of the present study, were involved. Both raters are language testing specialists with sufficient experience in developing various language tests and scoring rubrics for similar target test takers as the participants of this study. Both raters used the same scoring rubric, and sample scoring was conducted on a few answers first for calibration. Based on the sample scoring results, both raters met and discussed their scoring methods and agreed upon a unified rule for each scoring domain in the rubric. Following this process, all 100 essays

produced by the 50 participants from the two tests (NRW and RW) were scored by the raters independently. Inter-rater reliability was measured in each scoring domain and for each test, and a high degree of reliability was found. The average intraclass correlation coefficient (ICC) was .93 for the NRW test and .89 for the RW test.

First, the scores on the NRW and the RW tests were compared both as total scores and as sub-scores in each scoring domain. A paired *t* test was conducted to examine the significance level of the score differences followed by the Shapiro–Wilk normality test. The scores on the writing tests were further analyzed by test takers’ proficiency level. For this, the initial test results were used as a baseline to allocate three equal-sized proficiency groups. Furthermore, we used baseline ability estimated from the Rasch model to identify students’ proficiency level. This way, we were able to precisely examine the score difference between the NRW and the RW tests by proficiency without an intervention of the rater effect. The scores on the NRW and the RW tests (both the total and the sub-scores) were compared for each proficiency level separately using a paired *t* test. Test takers’ responses to the questionnaire were analyzed by computing frequency and proportion for each question to demonstrate an overall structure of the data. To examine the relationships between the test takers’ perceptions of using the writing resources and their writing performance in the RW test, the Pearson correlation coefficient was computed.

Results

Test score results

To examine the extent to which allowing access to language-support resources in a writing test might have had an effect on the test taker’s writing performance, the test scores on the two tests were first compared. Descriptive statistics of the scores on each rating domain and the total mean scores on each test are presented in Table 2 below.

The findings in Table 2 show that scores on the four scoring domains and the total mean scores on the RW test were higher than the scores on the NRW test. This finding indicates that the test takers performed better when the writing resources were provided than when they were not. To examine whether the differences in the total mean scores were statistically significant, paired *t* tests were conducted on all sub-scores and the total mean scores between the NRW test and the RW test. Shapiro–Wilk normality test results confirmed a normal distribution of all scores, and thus the parametric paired *t* test was used to examine the significance of the score differences. The results of the paired *t* tests are as follows.

The findings in Table 3 suggest that the scores of the RW test, total mean score, and all four domains, were significantly higher than the scores of the NRW test. However, the scores of all four domains that showed significant difference are still within the

Table 2 Descriptive statistics of test scores

		<i>N</i>	<i>Min</i>	<i>Max</i>	<i>Mean</i>	<i>S.D.</i>	<i>Skewness</i>	<i>Kurtosis</i>
NRW test	Task completion	50	12.00	24.00	17.50	2.64	0.28	– 0.13
	Content	50	11.00	24.00	16.07	2.89	0.53	0.16
	Organization	50	10.50	23.00	16.12	2.85	0.47	– 0.37
	Language use	50	10.00	22.00	14.93	2.97	0.48	0.02
	Total	50	47.00	93.00	64.62	10.57	0.63	0.03
RW test	Task completion	50	13.50	25.00	18.72	2.76	0.15	– 0.79
	Content	50	12.00	25.00	17.57	2.92	0.33	– 0.37
	Organization	50	12.00	24.00	17.38	2.97	0.19	– 0.51
	Language use	50	9.50	22.50	16.11	3.08	0.05	– 0.32
	Total	50	50.50	96.50	69.78	11.24	0.31	– 0.58

range of the same descriptor. For example, the score for the Content domain showed the largest improvement in test scores (by 1.50), but they were still within the same range of the descriptor (16–20). Therefore, it can be argued that having access to writing resources only had a small positive impact on test takers' writing test performance.

Effects of online writing resources by test-taker proficiency level

To examine whether the effects of the online writing resources might vary by test-taker proficiency, further analysis of the test scores was carried out for each proficiency level separately. The initial baseline test results were used to allocate three equal-sized groups: low, intermediate, and advanced. Descriptive statistics showed that the low- and intermediate-level test takers demonstrated noticeable improvements in their writing test performance, but the advanced-level test takers' score gap was comparably smaller. This small gap found in the advanced-level test takers was also not statistically significant [$t(16) = -.57, p = .58$] (see Table 4). The results of paired t tests on the other two proficiency groups showed that the total mean score gaps between the NRW and the RW tests were statistically significant : low [$t(16) = -2.52, p < .05$]; intermediate [$t(15) = -2.77, p < .05$].

Descriptive statistics of the scores for the four scoring domains were also computed for each proficiency group separately. Table 5 indicates that low-proficiency test takers' RW test scores were higher than their NRW test scores across all four domains (see Table 5). When paired t tests were conducted following the Shapiro–Wilk normality test, statistically significant score differences were found in Content [$t(16) = -2.52, p < .05$] and Organization [$t(16) = -2.52, p < .05$]. The improved RW scores for both the Content and Organization were in the higher range (16–20) than the scores from

Table 3 Paired t test results: NRW vs. RW

	<i>Mean dif.</i>	<i>S.D.</i>	<i>S.E.</i>	<i>t</i>	<i>df</i>	<i>p value</i>
Task completion	– 1.22	2.82	0.40	– 3.06	49	0.004
Content	– 1.50	3.27	0.46	– 3.24	49	0.002
Organization	– 1.26	3.03	0.43	– 2.94	49	0.005
Language use	– 1.18	3.49	0.49	– 2.39	49	0.021
Total	– 5.16	11.64	1.65	– 3.14	49	0.003

Table 4 Descriptive statistics of test takers' test scores by proficiency

		<i>N</i>	Min	Max	Mean	S.D.	Skewness	Kurtosis
Low	NRW	17	47.00	72.50	59.59	6.90	− 0.11	− 0.42
	RW	17	50.50	86.50	65.88	10.50	0.56	− 0.67
Intermediate	NRW	16	52.50	77.50	61.09	6.72	0.83	0.88
	RW	16	53.00	82.50	68.66	8.71	0.13	− 0.87
Advanced	NRW	17	48.00	93.00	72.97	11.77	− 0.50	0.15
	RW	17	51.50	96.50	74.74	12.73	− 0.16	− 0.47

the NRW condition (11–15) in the scoring rubric. No significant score difference was found in Task Completion [$t(16) = -1.63, p = .12$] or Language Use [$t(16) = -1.46, p = .17$]. These findings suggest that the significant positive improvements were found in the low-level test takers' writing performance when an access to the online resources are provided, particularly in the Content and Organization domains.

The scores of the intermediate group showed a similar trend, where the sub-scores of all four scoring domains were higher in the RW test than in the NRW test (see Table 6). The results of the paired t test showed that the differences were statistically significant in Task Completion [$t(15) = -3.02, p < .01$], Content [$t(15) = -2.75, p < .05$], and Language Use [$t(15) = -3.09, p < .01$], but not in Organization [$t(15) = -1.65, p = .12$]. In summary, the intermediate-level test takers performed slightly better in the RW test mode than in the NRW mode, but only the scores from the Language Use domain showed a significant improvement in the score range (from 11–15 to 16–20). The scores of other three scoring domains were within the same score range in both conditions. In short, the intermediate-level test takers significantly benefited in the language use aspect from the online resources.

According to Table 7, the advanced-level test takers' scores on the RW test were higher than on the NRW test in all four domains as well, but the score gaps were not as large as they were for the other two proficiency levels. When paired t tests were conducted, no significant results were found in any of the four domains: Task Completion [$t(16) = -.84, p = .41$], Content [$t(16) = -.66, p = .52$], Organization [$t(16) = -.71, p = .49$], and Language Use [$t(16) = -.04, p = .97$]. In summary, the inclusion of writing resources did not have any significant effect on the advanced-level test takers' writing performance in terms of either the total score or the score in each domain.

Table 5 Sub-score comparison of the low-proficiency group

			<i>N</i>	Mean	S.D.	Skewness	Kurtosis
Low	Task completion	NRW	17	16.41	2.29	− 0.05	− 0.01
		RW	17	17.94	3.07	0.27	− 0.58
	Content	NRW	17	14.59	1.94	− 0.44	− 0.49
		RW	17	16.65	3.26	0.20	0.26
	Organization	NRW	17	14.53	2.18	1.10	1.04
		RW	17	16.71	2.89	− 0.32	− 0.37
	Language use	NRW	17	13.59	2.96	0.28	− 0.45
		RW	17	14.88	3.59	0.54	0.30

Table 6 Sub-score comparison of the intermediate-proficiency group

			N	Mean	S.D.	Skewness	Kurtosis
Intermediate	Task completion	NRW	15	16.73	2.15	0.79	1.24
		RW	15	18.53	2.53	− 0.61	− 0.78
	Content	NRW	15	15.47	2.67	0.77	1.38
		RW	15	17.40	2.56	− 0.49	− 0.36
	Organization	NRW	15	15.87	2.64	− 0.34	0.53
		RW	15	16.60	2.87	− 0.52	− 0.37
	Language use	NRW	15	14.53	2.36	0.12	0.13
		RW	15	16.60	2.41	− 1.28	3.42

In summary, the use of online resources generally improved L2 test takers' writing test performance, and this effect was found in all four scoring domains. In particular, low-proficiency test takers' scores for the Content and the Organization scores and intermediate-proficiency test takers' Language Use scores were improved to a higher level range of the scoring descriptor in the RW condition. It can be speculated that the advanced-level learners did not rely on the use of writing resources as much as the other two proficiency groups did when completing the RW test. Taken together, these results suggest the inclusion of writing resources in a writing test partially supports test takers' writing performance, and the effect varies by test-taker proficiency. An overview of these test score comparisons by proficiency is presented in Fig. 3 below.

Multidimensional multi-facet model analysis

To further investigate the extent to which the effect of online writing resources might vary by test takers' proficiency, we used a multidimensional multi-facet Rasch model to compare the NRW test and RW test on the same scale. The Rasch model technique estimates students' abilities in the two tests while controlling the rater effect. For this, we used the baseline ability estimated in the Rasch model to precisely divide the test takers' proficiency into three levels. Then, we compared the scores of the NRW and the RW tests, which were treated as separate dimensions, in the multidimensional multi-facet Rasch model. The results computed through the Rasch model analysis are presented in Table 8 below.

In Table 8, the scale unit that indicate students' writing ability is logit. Hence, the higher logit value indicates higher writing ability. The results show that the

Table 7 Sub-score comparison of the high-proficiency group

			N	Mean	S.D.	Skewness	Kurtosis
Advanced	Task completion	NRW	18	19.06	3.32	− 1.08	2.03
		RW	18	19.72	3.06	− 0.24	− 0.07
	Content	NRW	18	17.83	3.81	− 0.35	− 0.56
		RW	18	18.44	3.88	0.20	− 0.73
	Organization	NRW	18	18.44	2.91	− 0.26	− 0.89
		RW	18	18.83	3.00	0.40	− 0.73
	Language use	NRW	18	16.83	3.40	0.47	− 1.14
		RW	18	16.56	3.47	0.15	− 1.06

intermediate-level test takers showed the largest performance improvement in the RW test by 1.28 logit, while the low- and high-proficiency test takers only showed small improvements in the RW test—by 0.24 and 0.42 logits, respectively. Findings together suggest that the effect of online references in writing tests vary by the proficiency level when the rater effect is removed. More specifically, we also found that the intermediate-level test takers benefitted the most from the online resource feature compared with other two proficiency groups.

Questionnaire results

To answer the third research question, which asks the degree to which test takers use writing resources in real-life and on the writing test, test takers were asked to respond to a questionnaire immediately after the second test was finished. Of the 50 test takers, 42 of them (84%) reported that they used writing resources in real-life writing activities. The descriptive statistics of the dichotomous item on whether writing resources should be allowed or not in conventional foreign language writing tests showed an ambivalent result: 54% agreed, 36% disagreed, and 10% did not respond. In terms of the usefulness of the online writing aids, 98% of the total test takers found the writing resources provided in the RW test helpful in improving their writing (see Table 9).

Table 8 Test performance scale logit by proficiency (rater effect removed)

		N	Min	Max	Mean	S.D.
Low	NRW	17	− 7.30	2.79	− 2.40	3.27
	RW	17	− 10.28	2.63	− 2.16	3.61
Intermediate	NRW	16	− 10.52	2.91	− 2.80	3.59
	RW	16	− 10.15	7.35	− 1.52	4.86
Advanced	NRW	17	− 7.01	9.16	− 0.92	4.05
	RW	17	− 5.03	8.16	− 0.50	4.19

Table 9 Usefulness and helpfulness of the online writing resources

	N	Frequency (ratio %)	
		Yes	No
Do you normally use online resources in a real-life writing tasks?	50	42 (84%)	8 (16%)
Do you think we should allow using writing resources in conventional foreign language writing tests?	45	27 (54%)	18 (36%)
Did you find the writing resources helpful in improving the quality of your writing?	50	49 (98%)	1 (2%)

The relationship between test takers' positions (pro or con on introducing writing resources in a test setting) and their actual RW test scores was examined, and the results of Pearson correlation indicated that there was no statistically significant correlation between the two [$r(48) = 0.93$, $p = .545$]. In summary, we found that test takers, in general, preferred using writing resources in their writing process, but their perceptions toward using the resources was not significantly associated with their actual writing performance.

On the question about the use of writing resources during Test 2, 37 test takers (80%) responded that they used the online dictionary, and the rest nine test takers (20%) said they used the concordance. Also, on the question that asked what writing resources should be included in the future writing tests, 46 test takers (97.8%) agreed that the online dictionary should be included while only eleven test takers (23.4%) believed the Internet search engine can be included in the test (Table 10).

Discussion

Summary of findings

This paper presented findings on the extent to which having access to online writing resources in a computer-integrated writing test had an effect on L2 test takers' writing test scores. A total of 50 learners of English in Korea were involved in two short essay-writing tests, the first one (NRW) in a non-referenced mode and the second one (RW) with access to online writing resources such as a concordancer and dictionaries. The results showed that test takers' writing test scores were significantly improved in the RW test condition in all four writing scoring domains: Task Completion, Content, Organization, and Language Use. However, the improved scores of the RW test was still within the

Table 10 Type of online resource features favored by test takers

	N	Frequency (ratio %)		
		Online dictionary	Concordancer	Search engine
What writing resources did you use during the Test 2?	46	37 (80.43%)	9 (19.57%)	0
What writing resources do you think should be included in a real writing test?	47	46 (97.87%)	0	11 (23.4%)

same scoring descriptor, which indicates that the online resources only had a weak positive effect on writing performance. When the effect of the online resources were examined by proficiency, the low- and intermediate-level test takers showed significant improvement in the RW test result while the advanced-level test takers did not show any significant improvement. Most test takers also reported that they used writing-support resources in real-life writing activities and believed that use of such aids improved the quality of their writing. On the idea of introducing writing-resource features in writing tests, test takers showed a mixed opinion, and their position was not necessarily correlated with their writing performance in the RW condition. However, most participants responded that they used online dictionary the most in the RW test, and almost all of them (98%) argued that such feature should be included in conventional writing tests.

The nature of using language-support resources in writing tests

We hypothesized that the test takers would show a substantial improvement in the Language Use score of the RW test because using the writing resources was expected to compensate for the L2 learners' lack of linguistic ability. Hence, we expected linguistic tools, such as the concordance checker or dictionary, would improve their test scores, particularly their score in the Language Use domain. In general, the results of this study met our hypothesis in which the test takers' scores including in the Language Use domain significantly increased. Exceptions were found in the low- and advanced-level learners, whose Language Use scores did not improve or remained in the same scoring range in the RW mode. It can be speculated that the low-level test takers' inappropriate use of vocabulary or incorrect grammar was not noticeably improved by the use of writing resources. This might be due to their lack of strategic knowledge in finding and choosing appropriate words or phrases in their writings. Similar to this finding, Tian and Zhou's (2020) study also showed that the low-proficiency-level Chinese learners of English benefitted more from the teacher feedback than the automated feedback feature. The advanced-level test takers of the present study have not used the linguistic tools as much as the other two proficiency groups, possibly because they were more confident about their language ability. In fact, several spelling errors were found in the advanced-level test takers' writings, which implies that they did not actively use the writing resources that offered a spell-checking function. In a similar vein, East (2007) also showed that L2 test takers' German writing test scores varied for different proficiency levels when access to a dictionary was given, and the advanced-level learners' writing performance was negatively affected by it. The present study also found that the intermediate-level test takers benefitted the most from the online writing resources. Their writing scores were significantly improved in the RW condition, with an exception only in Organization. Taking this into consideration, it seems reasonable to speculate that the score for Organization was not largely affected by the Concord Writer because the tool is technically designed to support language-use aspects only (Cobb, 2019).

It was also found that the scores for Content were also improved by introducing the writing resources to L2 learners. This means that the test takers provided richer contents in the RW test by presenting a greater variety of supporting ideas or examples. This echoes the findings of similar existing studies, which showed how the external resources facilitated EFL writers' brainstorming phase and ultimately improved the content of the writing output (Al-Shehab, 2020; Hajimaghsoodi & Maftoon, 2020; Rashtchi & Porkar, 2020).

Test takers in this study used the dictionary function the most, which allowed them to use a greater variety of vocabulary, including words that they had not acquired. In fact, 80% of the test takers said they used the dictionary, and we believe that the use of the dictionary functions helped them to present more superior ideas in the essay. Similarly, Sun and Hu (2020) also revealed that the Chinese learners of English resorted to online dictionaries the most as reference tools while completing the writing tasks. In this regard, Barnes et al. (1999) showed that teachers in the UK decided to allow students to use dictionaries in foreign language exams, believing that their use was a relevant and authentic skill. However, Frankenberg-Garcia (2020, p. 30) argued that a large number of dictionary users simply look up meanings or spellings of words without knowing that they could also "consult dictionaries" to address higher-level needs of writing. In summary, the results of the present study suggest that the use of online resources in writing tests could compensate for L2 test takers' lack of vocabulary knowledge.

Test fairness and authenticity issues in web-based writing assessments

Poe and Elliot (2019) discussed issues of test fairness based on systematic reviews on 73 articles and identified a total of five major trends. Among these trends, the introduction of online resources in a writing test is associated with "elimination of bias" and "fairness as the pursuit of validity" issues. We believe that providing online resources is a way of eliminating a bias in a test and is contextually a more valid method to measure test takers' true writing ability. Although the present study did not use psychometric techniques to examine test biases, we paid more attention to the test construct that should be measured through writing tasks instead. If test developers wish to introduce writing-resource features in their writing tests, they should first investigate whether providing such features would greatly benefit particular test takers but not others. For this, it is essential that test developers clearly understand and apply what the constructs are. If the construct of a writing assessment is solely measuring learners' language knowledge while referring to external resources is completely restricted, the inclusion of writing resources would be a construct-irrelevant variable and thus not fair. However, if the construct of writing ability is to measure learners' real-life writing ability as required in the TLU domain, we cannot deny that the ability to use external resources is also a construct-relevant skill. Rather, it would be a construct under-representation if we did not measure what test takers are truly capable of, particularly if the

ability is what is required in the TLU domain. Oh (2020) also argued that providing writing resources such as a dictionary is “a part of the variance that contributes in understanding the test taker’s writing ability” (p. 3). In other words, in the traditional paper-based writing-test mode, we might be underestimating candidates’ true writing ability. Therefore, we argue that better performance shown by the inclusion of online resources in a writing test should be encouraged rather than obstructed. However, a caution is also warranted in deciding the degree to which test developers provide such writing support tools to the test takers. It would not be a fair or valid test if test takers are allowed to search for ideas (via online search engines, etc.) and copy them into their writing draft. It should be also acknowledged that there could be some test takers’ anxiety and keyboarding skill-related effect on writing as well. Since the foundational conception and definition of fairness still needs further theorization (Poe & Elliot, 2019, p. 14), much work still remains particularly with regard to using this digital language reference resources that is relatively new in this field.

When we call for authenticity in the use of writing resources in a writing test, it is essential that we understand whether the use of such features in an assessment condition corresponds to real life. Prensky (2001) first introduced the term ‘digital natives’ and argued that the introduction and spreading of new [digital technology](#) in the recent decades had changed the way students think and process information and perform academic writing tasks. An authentic language test presents tasks that reflect a real-life writing activity and a clear relationship between the task and the writer. This present study found that test takers, particularly the low and intermediate learners, used the online features to compensate for their lack of linguistic ability while completing the writing tasks. Given that such online features are also available in the real-life computer-based or mobile-based writing setting, we believe providing them is a better representation of the real-life writing construct. As Frankenberg-Garcia (2005) argued, it is now much easier for language learners to find authentic examples and appropriate usage of language with the help of online language-support resources. Therefore, writing tasks with access to online resources are situationally authentic for today’s language learners.

When Bachman and Palmer (1996) introduced the term ‘situational authenticity’, they suggested that the degree of authenticity of a test is investigated by “comparing the characteristics of test tasks and target-language use tasks” (p. 45). With this in mind, we argue that a situationally authentic task does not only indicate the authenticity of the contents used in the task but also the ways to accomplish the given task in a real-life situation. In other words, being able to utilize tools or resources to compensate for a lack of knowledge or information to complete a given task should also be considered an authentic skill required in real-life communication situations. In terms of assessing writing skills, the utilization of external resources to correct mechanical errors and enrich the content of the writing is an authentic skill required in the modern digital society and is considered as part of digital literacy skills.

Conclusion

Limitations

This study is not without its limitations. In terms of the design of the study, counterbalancing was not precisely done since the number of test takers for each topic and each condition was not equal. The data collection phase was carried out with an intact class unit, and thus convenience sampling was inevitably carried out. Also, a possible fatigue effect may be found in the RW test because it was carried out after the NRW test on the same day. The small sample size is also a concern, and the results of the inferential statistics may lack power. In terms of the research findings, we acknowledge that the improved sub-score of 'content' criteria lacks clear evidence to explain the rationales for such improvement. Follow-up studies replicating this study need to investigate these specific issues using qualitative approaches. Lastly, the traditional "write an essay on a given topic" task we used in this study might not be an ideal item type to examine the effect of writing resources precisely because there is no flexibility to utilize a wide range of ideas. Instead, the new innovative scenario-based language tests (SBLA) might be a better model to approach in follow-up studies (see Banerjee, , 2019).

Suggestion for future studies

Based on the findings of this study, we believe the ability to use online writing resources is an authentic and construct-relevant skill in L2 writing assessment. The ability to use digital technology to search for and evaluate appropriate, relevant, and valid information and utilize it as a source of writing has become a core academic writing skill required in real life. The recent generation has rapidly become familiar with these technology-mediated writing activities, and a variety of different functions and tools offer various strategies that aid writers in completing given tasks more efficiently. Future studies need to explore further other types of writing resources that go beyond the simple dictionary function or corpus tools. For example, web-based machine translators, such as Google Translate, or automatic proof-reading tools, such as Grammarly, might be new types of writing resources for the new generation (Chon & Shin, 2020; Kirchhoff et al., 2011). Both types of tools use machine-learning artificial intelligence (A.I.) algorithms to offer mechanical support for language processing. Another relatively new A.I. program, chatbot, might also be an effective feature to be introduced in writing assessments if one uses it to ask for question and answer exchanges to check, clarify, and receive relevant information while completing a writing task. An empirical study has found that these question-and-answer exchanges with the chatbot were more effective than the Internet search engines because the learners showed a longer memory retention (Abbasi & Kazi, 2014). Looking into the effects of such A.I. technology as a writing assistant on the quality of writing, and whether the product of writing produced through such technology might be construct-relevant or not, are unexplored topics that should be investigated in future studies. Considering the growing interest in and reliance on A.I. technology, studies on these features would be a timely, relevant, and beneficial guideline for future directions of language testing.

Appendix

Table 11 Scoring rubric

Score	Scoring dimensions			
	Task completion	Content	Organization	Language use
	<ul style="list-style-type: none"> ✓ Completed an argumentative essay with 300–400 words ✓ Explained the writer's position with the given two reasons ✓ Explained the writer's additional reason ✓ Explained the writer's position including the three parts of Introduction, Body, and Conclusion 	<ul style="list-style-type: none"> ✓ Clear main idea ✓ Provided adequate supporting details in a logical manner ✓ In-depth discussion of the main idea 	<ul style="list-style-type: none"> ✓ Clear topic sentences ✓ Clear supporting details ✓ Clear concluding sentences ✓ Logical progression of the main idea 	<ul style="list-style-type: none"> ✓ Used only English and appropriate vocabulary ✓ Variety in sentence structures ✓ Few grammar errors ✓ Few errors in spelling ✓ Appropriate use of mechanics
21–25	The writer completely addresses the assigned writing task.	The writer provides relevant content that is complete, concrete, and thoroughly developed.	The writer develops an adequate organizational structure. Main ideas are complete and logically sequenced.	Few grammar or spelling errors are evident. Vocabulary usage is generally controlled and ideas are expressed clearly.
16–20	The writer makes a reasonable, mostly complete, attempt to address the writing task.	The writer provides relevant content that is mostly complete.	The writer develops a mostly complete organizational structure. The sequencing of main ideas is mostly complete and logical.	Some grammar and spelling errors are evident, but they do not distract from the writer's message. Vocabulary usage is mostly correct, although some words may be misused.
11–15	The writer makes a reasonable but incomplete attempt to address the writing task.	The writer provides some relevant content, but it may be incomplete or undeveloped.	The writer develops an incomplete organizational structure. The sequencing of main ideas may be incomplete and illogical.	Some grammar and spelling errors may affect the communication of the writer's message. Some control of vocabulary usage is evident, although errors may affect communication of the writer's message.
6–10	The writer makes a poor, incomplete attempt to address the writing task.	The writer attempts to provide relevant content, but it may be irrelevant, undeveloped, and incomplete.	The writer attempts to develop an organizational structure, but it is incomplete. Main ideas may not be evident or logical.	Numerous errors in grammar, spelling, and vocabulary usage negatively affect the communication of the writer's message.
1–5	The writer fails to address the writing task.	The writer fails to provide any content that is relevant or complete.	The writer fails to develop a meaningful organizational structure, or logical progression of main ideas.	Pervasive errors in grammar, spelling, and vocabulary usage significantly impair communication of the writer's message.
0	The writer fails to address the writing task. Answer not suitable for the given task (scribbling, use of obscene words, etc.); no answer provided; over 70% is written in any language other than English; or a memorized text on a different topic	Completely irrelevant.	No organizational structure or logical progression of main ideas.	Not comprehensible due to language errors.

Abbreviations

CEFR: Common European Framework of Reference for Languages; COBUILD: Collins Birmingham University International Language Database; CSAT: College Scholastic Ability Test; ESL: English as a Second Language; IELTS: International English Language Testing System

Acknowledgements

Not applicable

Authors' contributions

DS, the first author, managed the overall process of conducting this study including data collection, data analysis, and writing up the results. SKK, the corresponding author, participated in rating the essays, analyzing the data, reviewing the literature, and writing up the manuscript. YL, the co-author, contributed in computing quantitative analysis (particularly the Rasch model) and writing up the results. The author(s) read and approved the final manuscript.

Funding

Not applicable

Availability of data and materials

The data that support the findings of this study are available on request from the corresponding author, S. K. The data are not publicly available due to ethical reasons, as the containing information could compromise the privacy of research participants.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

¹Gwangju National University of Education, Gwangju, Republic of Korea. ²Chinju National University of Education, Jinju, Republic of Korea. ³Yeungnam University, Gyeongsan, Republic of Korea.

Received: 6 October 2020 Accepted: 28 February 2021

Published online: 25 March 2021

References

- Abbasi, S., & Kazi, H. (2014). Measuring effectiveness of learning chatbot systems on student's learning outcome and memory retention. *Asian Journal of Applied Science and Engineering*, 3, 57–66.
- Al-Shehab, M. (2020). The role of mobile-assisted language learning (MALL) in enhancing the writing skills of intermediate IEP students: Expectations vs reality. *Language Teaching Research Quarterly*, 20, 1–18.
- Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.
- Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered*, (pp. 41–71). University of Ottawa Press. <https://doi.org/10.2307/j.ctt1ckpcf9>.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.
- Banerjee, H. L. (2019). Investigating the construct of topical knowledge in second language assessment: A scenario-based assessment approach. *Language Assessment Quarterly*, 16(2), 133–160. <https://doi.org/10.1080/15434303.2019.1628237>.
- Barnes, A., Hunt, M., & Powell, B. (1999). Dictionary use in the teaching and examining of MFLs at GCSE. *The Language Learning Journal*, 19(1), 19–27. <https://doi.org/10.1080/09571739985200051>.
- Bereiter, C., & Scardamalia, M. (1987). An attainable version of high literacy: Approaches to teaching higher-order skills in reading and writing. *Curriculum inquiry*, 17(1), 9–30.
- Brunfaut, T., Harding, L., & Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing*, 36, 3–18. <https://doi.org/10.1016/j.asw.2018.02.003>.
- Chalhoub-Deville, M. (2010). Technology in standardized language assessments. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics*, (2nd ed., pp. 511–526). Oxford University Press.
- Chapelle, C. A. (2010). Technology in language testing. In G. Fulcher, & R. Trasher (Eds.), *Language testing videos* Retrieved from <http://languagetesting.info>.
- Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511733116>.
- Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing*, 16(1), 49–71. <https://doi.org/10.1016/j.asw.2010.11.001>.
- Chon, Y. V., & Shin, D. (2020). Direct writing, translated writing, and machine-translated writing: a text-level analysis with Coh-Metrix. *English Teaching*, 75(1), 25–48. <https://doi.org/10.15858/engtea.75.1.202003.25>.
- Cobb, T. (2019). *Concord writer* [A writing tool on complete lexical tutor]. Retrieved from <https://www.lexutor.ca/cgi-bin/conc/write/?lingo=English>.
- Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge University Press.
- Douglas, D. (2013). Technology and language testing. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Wiley-Blackwell.
- East, M. (2007). Bilingual dictionaries in tests of L2 writing proficiency: Do they make a difference? *Language Testing*, 24, 331–353.
- Frankenberg-García, A. (2005). A peek into what today's language learners as researchers actually do. *International Journal of Lexicography*, 18(3), 335–355. <https://doi.org/10.1093/ijl/eci015>.
- Frankenberg-García, A. (2011). Beyond L1-L2 equivalents: Where do users of English as a foreign language turn for help? *International Journal of Lexicography*, 24(1), 97–123. <https://doi.org/10.1093/ijl/ecq038>.

- Frankenberg-Garcia, A. (2020). Combining user needs, lexicographic data and digital writing environments. *Language Teaching*, 53(1), 29–43. <https://doi.org/10.1017/S0261444818000277>.
- Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, 32(3), 301–319. <https://doi.org/10.1016/j.system.2004.04.001>.
- Gruba (2014). New media in language assessments. In A. J. Kunnan (Ed.), *The companion to language assessment*, (vol. II, Part 8, pp. 1–16). Wiley.
- Hajimaghsoodi, S., & Maftoon, P. (2020). The effect of activity theory-based computer-assisted language learning on EFL learners' writing achievement. *Language Teaching Research Quarterly*, 16, 1–21. <https://doi.org/10.32038/ltrq.2020.16.01>.
- Hayes, J. R., & Flower, L. (1981). Uncovering cognitive processes in writing: An introduction to protocol analysis. ERIC Clearinghouse.
- Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice*, 20(3), 16–25.
- Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Harfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: practical approach*. Rowley: Newbury House.
- Jin, Y. (2012). Validating computer-based language tests: computer literacy and the construct of a computer-based language test. *Computer-Assisted Foreign Language Education*, 1, 11–15.
- Jin, Y., & Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. *Language Assessment Quarterly*, 14(2), 101–119. <https://doi.org/10.1080/15434303.2016.1261293>.
- Joo, H., & Shin, D. (2012). *Effect of the use of the concordancer on student performance in English writing assessment*, Research Report (RRE-2012-19) (). Korea Institute for Curriculum and Evaluation.
- Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of writing research*, 1(1), 1–26.
- Kirchhoff, K., Turner, A. M., Axelrod, A., & Saavedra, F. (2011). Application of statistical machine translation to public health information: A feasibility study. *Journal of the American Medical Informatics Association*, 18(4), 473–478. <https://doi.org/10.1136/amiajnl-2011-000176>.
- Krashen, S. D. (1981). *Second language acquisition and second language learning*. University of Southern California.
- Lee, M., Shin, D., & Chon, Y. V. (2009). Online corpus consultation in L2 writing for in-service teachers of English. *English Teaching*, 64(2), 233–254.
- Li, J. (2006). The mediation of technology in ESL writing and its implications for writing assessment. *Assessing Writing*, 11(1), 5–21. <https://doi.org/10.1016/j.asw.2005.09.001>.
- Martin, J. R., & Rose, D. (2005). Designing literacy pedagogy: scaffolding democracy in the classroom. In R. Hasan, C. M. I. M. Matthiessen, & J. Webster (Eds.), *Continuing discourse on language: a functional perspective*, (pp. 251–280). Equinox.
- Meyers, E. M., Erickson, I., & Small, R. V. (2013). Digital literacy and informal learning environments: an introduction. *Learning, Media and Technology*, 38(4), 355–367. <https://doi.org/10.1080/17439884.2013.783597>.
- Oh, S. (2020). Second language learners' use of writing resources in writing assessment. *Language Assessment Quarterly*, 17(1), 60–84. <https://doi.org/10.1080/15434303.2019.1674854>.
- O'Sullivan, I., & Chambers, A. (2006). Learners' writing skills in French: corpus consultation and learner evaluation. *Journal of Second Language Writing*, 15, 46–68.
- Park, Y., Min, H., Kim, J., & Kim, Y. (2015). An investigation into the KICE writing assessment software for high schools: teacher and student perceptions. *Studies in English Education*, 20(3), 195–237.
- Poe, M., & Elliot, N. (2019). Evidence of fairness: twenty-five years of research in assessing writing. *Assessing Writing*, 42, 100418. <https://doi.org/10.1016/j.asw.2019.100418>.
- Porter, C. (2015). Common core-linked tests spur schools to teach typing. *The Wall Street Journal*. Retrieved from <https://www.wsj.com/articles/common-core-linked-tests-spur-schools-to-teach-typing-1423073700>.
- Prensky, M. (2001). Digital natives, digital immigrants Part 1. *On the Horizon*, 9(5), 1–6. <https://doi.org/10.1108/10748120110424816>.
- Rashtchi, M., & Porkar, R. (2020). Brainstorming revisited: does technology facilitate argumentative essay writing? *Language Teaching Research Quarterly*, 18, 1–20. <https://doi.org/10.32038/ltrq.2020.18.01>.
- Sawaki, Y. (2012). Technology in language testing. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing*, (pp. 426–437). Routledge.
- Shin, D., Min, H., & Joo, H. (2012). *The manual of English speaking and writing assessment for preparing NEAT, KICE Research Material (ORM 2012-67)* (). Seoul: Korea Institute for Curriculum and Evaluation.
- Sun, X., & Hu, G. (2020). Direct and indirect data-driven learning: an experimental study of hedging in an EFL writing class. *Language Teaching Research*. <https://doi.org/10.1177/1362168820954459>.
- The American Library Association. (2019). Digital literacy. Retrieved from <https://literacy.ala.org/digital-literacy/>
- Tian, L., & Zhou, Y. (2020). Learner engagement with automated feedback, peer feedback and teacher feedback in an online EFL writing context. *System*. <https://doi.org/10.1016/j.system.2020.102247>.
- Van Waes, L., & Schellens, P. J. (2003). Writing profiles: the effect of the writing mode on pausing and revision patterns of experienced writers. *Journal of Pragmatics*, 35(6), 829–853. [https://doi.org/10.1016/S0378-2166\(02\)00121-2](https://doi.org/10.1016/S0378-2166(02)00121-2).
- Watanabe, H. (2016). Genre analysis of writing tasks in Japanese university entrance examinations. *Language Testing in Asia*, 6(1), 1–14.
- Weigle, S. C. (2005). Second language writing expertise. In *Expertise in second language learning and teaching* (pp. 128–149). Palgrave Macmillan, London.
- Winke, P. M., & Isbell, D. R. (2017). Computer-assisted language assessment. In S. L. Thorne, & S. May (Eds.), *Language, education and technology*, (3rd ed., pp. 1–13). Springer. https://doi.org/10.1007/978-3-319-02237-6_25.
- Yoon, C. (2016). Individual differences in online reference resource consultation: case studies of Korean ESL graduate writers. *Journal of Second Language Writing*, 32, 67–80. <https://doi.org/10.1016/j.jslw.2016.04.002>.
- Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, 13(4), 257–283. <https://doi.org/10.1016/j.jslw.2004.06.002>.
- Yu, G., & Zhang, J. (2017). Computer-based English language testing in china: present and future. *Language Assessment Quarterly*, 14(2), 177–188. <https://doi.org/10.1080/15434303.2017.1303704>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.