Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis

Jaeho Jeon

Published online: 15 Oct 2021.

Submit your article to this journal

View related articles

View Crossmark data

Routledge
Taylor & Francis Group

Check for updates

Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis

Jaeho Jeon (iD)

Department of English Education, Seoul National University of Education, Seochojungang-ro, Seocho-gu, Seoul, Republic of Korea

ABSTRACT

This study investigated the effect of Chatbot-Assisted Dynamic Assessment (CA-DA) on vocabulary learning and provided insights into learner abilities drawn from its implementation. Through the use of mediating chatbots, this study implemented DA to multiple learners simultaneously and provided each learner with human-like interaction. The chatbots were created using Google's Dialogflow. Fifty-three Korean EFL primary school learners who were confirmed to demonstrate the same range in vocabulary size participated in this study. They were randomly assigned to three groups: CA-DA, Chatbot-Assisted Non-Dynamic Assessment (CA-NDA), and a control. For two treatment sessions, the learners were asked to read texts and identify the meaning of underlined target words. The chatbots provided graduated assistance to learners in the CA-DA group and only target word definitions to learners in the CA-NDA group. The control group did not utilize chatbots. Two posttests (receptive and productive) were administered both immediately after and two weeks after the second treatment session. Interaction records between the chatbots and learners across the two treatment sessions were also collected. Posttest results showed that vocabulary gains in the CA-DA group were significantly higher than in the other groups. Analysis of interactions between the chatbots and learners for the CA-DA sessions provided detailed evidence of learner development. The findings suggest that CA-DA could not only promote vocabulary acquisition but could also offer diagnostic information about individual learners concerning vocabulary learning. This study also demonstrated the potential of chatbot technology to support language learners.

I. Introduction

Given the limited amount of class time that can be used exclusively for vocabulary in EFL classroom settings such as South Korea, it is necessary to use effective methods that can increase opportunities for English

exposure and vocabulary learning at the same time. In this vein, glosses are useful because they work as both aids for English comprehension and tools for vocabulary learning simultaneously when learners read a text (Hulstijn, Hollander, & Greidanus, 1996; Ko, 2005). To address some concerns regarding text-only glosses, such as when a learner does not pay attention to glosses and considers them simply as aids for comprehension (Yanguas, 2009), different types of glosses have been introduced in SLA, such as a multimodal gloss or a multiple-question gloss, to make glossing more effective for vocabulary learning. Furthermore, having been developed using new technology, these glosses have been confirmed to be more effective for vocabulary learning than text-only glosses (Abraham, 2008; Yanguas, 2009). However, glosses used in previous studies were static in terms of presenting meanings and explanations for glossed words. In other words, learners passively received the given definition without any further assistance regarding identifying the meaning independently.

In order to compensate for this passive nature, glossing can be designed by drawing on Dynamic Assessment (DA), a framework that integrates assessment with instruction (Rassaei, 2020). DA intends not only to diagnose learners' language abilities but also to promote development as a mediated process of individual transition from other-regulation toward self-regulation (Davin, 2016; Lantolf & Poehner, 2011; Lantolf & Thorne, 2006; Poehner, 2008; Poehner & Leontjev, 2020). To this end, DA utilizes different types of verbal mediation that begin as implicit forms of prompt and become increasingly explicit if required by the learner (Aljaafreh & Lantolf, 1994). Interaction between learners and a more advanced partner who provides graduated mediation, usually a teacher in the classroom setting, can play a crucial role in promoting learners' development as well as in providing insights into learner abilities (Davin, Herazo, & Sagre, 2017; Herazo, Davin, & Sagre, 2019; Lantolf & Poehner, 2011).

Meanwhile, different challenges have been reported when implementing DA in the classroom setting. As Davin (2013) stated, "This form of DA administration in dyads is time-consuming, limiting the number of participants with whom a mediator can work" (p. 307). In an attempt to address this practical concern, Poehner and van Compernolle (2013) and Poehner, Zhang, & Lu (2015) introduced the concept of Computerized Dynamic Assessment (C-DA), in which a software program acted as a mediator that could provide automated and graduated mediation via a multiple-question format. By continuing this line of research, Ebadi, Weisi, Monkaresi, and Bahramlou (2018) also revealed how C-DA could be applied in the area of vocabulary learning. These projects have successfully demonstrated the potential of C-DA by overcoming the practical

difficulties that have been pointed out in the literature. For example, they showed that C-DA could be applied to multiple learners simultaneously, maximizing the number of participants within a given timeframe. In addition, C-DA software programs can automatically document learners' performance as data that teachers can then use for more finely-tuned subsequent teaching (Mehri Kamrood, Davoudi, Ghaniabadi, & Amirian, 2019; Qin & Van Compernolle, 2021).

Despite the practical benefits of C-DA, there are still some limitations regarding the technology used in the studies. Employing computer software as a mediator, C-DA primarily focused on L2 recognition and used a multiple-choice format where participants were limited to a few given options (Ai, 2017). These limitations deserve more attention regarding teaching and learning vocabulary because productive vocabulary knowledge is also one crucial aspect of knowing vocabulary (Nation, 2013). Learners need to receive opportunities not only to recognize but also to produce vocabulary during DA. Further, multiple-choice mediation is not a desirable type of interaction in the L2 classroom because learners are limited to only a few options. In response to these issues, the present study introduced the concept of Chatbot-Assisted Dynamic Assessment (CA-DA) as a form of glossing in which a chatbot acted as a mediator regarding unfamiliar words when learners read a text. During CA-DA, learners engaged in more open-ended dialogues with chatbots and gained opportunities to recognize and produce unfamiliar words by receiving automated mediation while not constrained to a multiple-choice format. The following research questions guided this study.

1. Is CA-DA effective for receptive and productive vocabulary learning in L2 learners?
2. How can CA-DA be used for diagnosing L2 learners' vocabulary knowledge?

II. Background literature

1. Glossing for vocabulary learning

A gloss is a brief definition or synonym, either in L1 or L2, which is provided with the text (Nation, 2013). Glosses can work as both aids for text comprehension and tools for vocabulary learning simultaneously (Hulstijn et al., 1996; Ko, 2005). However, it has been argued that learners do not pay attention to text-only glossing and see the glosses simply as aids to comprehension rather than as sources for vocabulary learning (Bowles, 2004; Yanguas, 2009). Different types of glosses have been introduced to facilitate vocabulary learning in response to this concern, including multimodal glosses and multiple-choice glosses.

To examine the effectiveness of different types of glossing on vocabulary learning, Cognitive Load Theory can be considered. According to this theory, learning tasks, such as glossing tasks in language learning settings, become more effective when designed considering learners' cognitive architecture (Pass, Renkl, & Sweller, 2003). This theory suggests that working memory is limited in its capacity to hold information; thus, it is crucial for an instructional designer to consider a learner's limited cognitive ability and to effectively manage three types of cognitive load: intrinsic, extraneous, and germane. *Intrinsic cognitive load* is imposed by the complexity of the information to be processed; *extraneous cognitive load* is generated when tasks require unnecessary efforts from learners; *germane cognitive load* is one that enhances learning and contributes to schema acquisition and automation.

van Merriënboer, Kirschner, and Kester (2003) elaborated on how to manage cognitive overload by providing a scaffolding approach. Specifically, they suggested that when teachers present a complex task, teachers need to first present a simple version of the whole task and then progress to a more complex version to decrease cognitive load (van Merriënboer et al., 2003). For example, teachers can utilize varying levels of prompt (e.g., hints, leading questions, or examples) to adjust the task complexity and guide learners in successfully completing the task. In the same vein, Mayer's (2009) multimedia learning theory states that multimedia material, such as multimodal glossing, should be designed in a way that prevents learners' cognitive overload. The literature on multimodal glossing in the field of language learning has provided empirical evidence that glossing becomes more advantageous for vocabulary learning when it effectively manages cognitive load (Ramezanali & Faez, 2019; Teng, 2020; Türk & Erçetin, 2014).

Meanwhile, some scholars, noting that there might also exist the potential to make glossing interactive, have attempted to create an interactive gloss that involved learners in the decision-making process when identifying the meanings of glossed words (Hulstijn, 1992; Rott, Williams, & Cameron, 2002). In this type of gloss, known as multiple-choice glossing, learners have to choose the correct definition from reasonably similar choices in meaning to each other when reading a text. Hulstijn (1992) argued that multiple-choice glossing is more advantageous in terms of vocabulary learning because it fosters more thoughtful processing of glossed words than traditional glosses that simply provide the meanings of glossed words. One concern regarding this type of glossing arose when learners made incorrect choices and did not receive immediate feedback. However, technology-enhanced glossing overcame this obstacle by presenting immediate confirmation in response to learner choice (Nation, 2013).

Although learners are allowed to identify definitions and receive immediate confirmation about their choices with technology-enhanced multiple-choice glossing, there still exist two concerns. First, learners still passively accept the provided definition with no opportunity to learn how to identify unfamiliar words by themselves. Second, previous literature has indicated that some language learners, specifically L2 learners with limited language proficiency, might be cognitively overloaded if required to identify the meanings of unknown words without systematic assistance (Teng, 2020; Türk & Erçetin, 2014). Therefore, it is worth investigating how glossing could systematically guide learners to identify meanings independently in an interactive environment and how the interactive gloss could cognitively assist learners in identifying the definitions of glossed words; thus, leading to effective vocabulary learning. To manage cognitive load in interactive glossing, the current study utilizes mediation that consists of hints, leading questions, and examples and explores the effects on vocabulary learning.

2. Dynamic assessment for vocabulary learning

DA can effectively be applied to vocabulary learning and diagnosis. DA, as a form of alternative assessment, is a move away from static assessments that are primarily aimed at confirming learners' past learning outcomes. More specifically, DA incorporates learning opportunities with assessment. That is, it provides diagnostic information about learner abilities, and at the same time, promotes learner development as a mediated process of individual transition from other-regulation toward self-regulation (Lantolf & Thorne, 2006). DA is based on the Zone of Proximal Development theory (ZPD), defined as "the distance between the actual development level as determined by independent problem solving and the level of potential development as determined through problem-solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978, p. 86). Some research has applied DA to classrooms and explored its effect on L2 development (Davin, 2013, 2016; Herazo et al., 2019; Lantolf & Poehner, 2011; Poehner & Leontjev, 2020). In most classroom-based DA literature, given that teachers acting as mediators should take care of multiple learners simultaneously, it was pre-scripted and graduated mediation that was used to make it possible for mediators to conduct a systematic diagnosis of learning and promote development in large-sized classes. In this approach which is called interventionist DA (Lantolf & Poehner, 2004), a teacher analyzes past learner performance and predicts difficulties that learners may encounter in subsequent learning. Based on the analysis, the teacher prepares mediation which gradually becomes more explicit from implicit. When

the teacher observes a learner struggling in an area that was predicted to be difficult during prompt preparation, the teacher starts by providing the most implicit prompt as planned (e.g., Are you sure?). If the learner cannot correct the answer based on the first prompt, the second prompt, which will be more explicit than the first, is provided (e.g., Can you pay attention to the next part?). One by one, the teacher provides the planned prompts depending on the learner's needs until the learner gains an understanding regarding the problem area. Using this method, the teacher can assist the learner in a contingent, graduated, and dialogic manner by effectively taking advantage of the learner's ZPD (Aljaafreh & Lantolf, 1994). Therefore, the teacher can first promote learner development by engaging in this type of dialogue, and at the same time, the teacher can implement assessment based on learner responses to each prompt and how many prompts learners required across different learning situations.

Meanwhile, there is a body of literature that has utilized DA for L2 vocabulary learning. These studies have indicated that DA can be used as an effective means for vocabulary learning (Andujar, 2020; van der Veen, Dobber, & van Oers, 2016). In some studies, DA was used to assist learners in gaining information about unknown words while reading a text (Ebadi et al., 2018; Rassaei, 2020). According to the definition of gloss (Nation, 2013), this use of DA for vocabulary learning can be viewed as a type of gloss. This dynamic approach to glossing is different from a traditional gloss in that glossing based on DA utilizes graduated mediation to help learners identify the meanings of unfamiliar words. In contrast, traditional glossing simply presents the meaning of unknown words to learners. Furthermore, some have utilized technology in DA for vocabulary learning. For example, Ebadi et al. (2018) implemented DA in a computerized environment where multiple learners were simultaneously provided with automated mediation in a gradual way. When participants encountered difficult words while reading a text, a software program designed for the study provided learners with mediation in a multiple-choice format. This type of DA differs from traditional multiple-choice glossing in that the software guided learners in identifying the glossed word by themselves. Despite the practical benefits of the automated DA system, it should be noted that mediation using a multiple-choice format is not a desirable type of interaction in the L2 classroom because it limits learners to only a few options.

Rassaei (2020) introduced the concept of dynamic glossing as "the process of offering mediation to learners to help them identify the correct definition for unfamiliar words in a text" (p. 289). He showed that a smartphone could be used as an effective glossing medium for

carrying out DA of vocabulary. In his study, when a participant encountered unfamiliar words in a text, a human mediator provided graduated feedback through the use of a smartphone chat application, more similarly replicating person-to-person conversation than the multiple-choice format because the participants were not limited to given options. However, this does not apply to large-sized classes because it is infeasible for a mediator to perform each individual interaction with every learner. Therefore, it is worth investigating other technology to effectively implement DA of vocabulary in order to overcome this practical constraint while not remaining limited to a few given options. This study introduces chatbot technology as a DA medium to provide each individual with opportunities for interaction where a learner is not limited to a multiple-choice format when attempting to identify unfamiliar words. To this end, this study employs an AI chatbot as an automated mediator that provides graduated mediation. The chatbot technology used for this study is discussed in further detail in the following section.

3. Artificial intelligence for dynamic assessment

With the assistance of an AI agent that is able to respond to a learner's errors in an automated manner, scholars in the field of language assessment have attempted to take advantage of AI technology when implementing assessment (e.g., Ai, 2017; Heift, 2017; Shute & Ventura, 2013). Ai (2017) created an AI agent that was equipped with the ability to identify grammatical errors when a learner performed Chinese to English translation tasks. In his study, the agent identified a learner's mistakes and provided a set of graduated feedback which became increasingly more explicit until the learner successfully corrected the error. Although the study also confirmed that room existed for AI technology improvement (e.g., technical errors), the AI agent was effective in the role of mediator and created opportunities for L2 grammar learning based on the theory of DA. AI agents hold potential in the implementation of C-DA in that they provide individual, graduated, and automated mediation to which a learner is able to respond through open-ended interaction (Ai, 2017). Despite this advantage, it may seem infeasible for teachers to use AI technology in the classroom because they would need to be equipped with programming knowledge in order to create an AI agent and apply AI-based DA. In this regard, it remains unanswered how teachers in the classroom could use AI technology when implementing DA because little is known about accessible AI technologies which can be applied to DA by teachers.

To address this concern, AI chatbots can be utilized to implement DA. Specifically, open-source chatbot builders such as Google's *Dialogflow*,

which is "a natural language understanding platform used to design and integrate a conversational user interface into mobile apps, web applications, devices, bots, interactive voice response systems" (Google, 2020), can be used with ease by teachers, given that the platform is designed to provide a simplified programming interface for general users (Lee, Yang, Shin, & Kim, 2020). Although detailed instruction for the use of the platform is beyond the scope of this research (for more information, visit https://cloud.google.com/dialogflow/docs), some important functions of the platform should be addressed regarding DA. Given that chatbots created using Dialogflow can be uploaded to digital devices, including smartphones and tablet PCs, it can be considered a ubiquitous technology that is free from time and space; therefore, it can be said that it provides a mobile-mediated setting for DA (Andujar, 2020; Ebadi & Bashir, 2021; Rassaei, 2021). In addition, Dialogflow offers an automatic transcript function that can be used as evidence of learner performance (Fryer, Coniam, Carpenter, & Lăpușneanu, 2020). This is more meaningful in DA because a mediator could easily access learning information regarding the number of prompts a learner required and how the learner responded to each prompt. However, despite the advantages of chatbot technology, research employing AI chatbots for DA, specifically, chatbots that are created with an open-source chatbot builder, does not exist. To fill this gap in the literature, the present study explores the effect of chatbots used for DA on vocabulary learning and demonstrates effective ways of using chatbots for vocabulary diagnosis.

III. Methodology

1. Participants

This research was carried out at a public primary school in South Korea. After receiving formal permission for this study from the school and from each learner and their parents, a brief demographic survey and the Vocabulary Size Test (Nation & Beglar, 2007) were administered to 81 twelve-year-old learners studying at the school to ensure group homogeneity. Regarding prior vocabulary knowledge, 58 learners who demonstrated vocabulary sizes of 1200–1500 words were chosen. Twenty-three learners were excluded because 18 learners tested above the selected range, and five learners tested below the range. Finally, three groups were created: two experimental groups of 18 learners (8 females and 10 males) each for CA-DA and CA-NDA and one control group of 17 learners (7 females and 10 males). The remaining five learners were selected for a pilot phase.

According to the demographic survey, no participants had study-abroad experience in English-speaking countries, and they all had around four

years of EFL learning experience at the time of the research. Furthermore, the participants were confirmed not to have previously used chatbots for language learning. Regarding their overall English proficiency, an English teacher who had instructed the participants for more than one year at the time of the research provided information; the participants belonged to the novice level according to ACTFL guidelines for reading and other skills. Given that all participants came from the same school and the same neighborhood, it can be assumed that all three groups were reasonably homogeneous not only in their vocabulary size, year of EFL learning, gender, overall English proficiency, and chatbot use experience for language learning but also in other aspects such as their socioeconomic background.

2. Procedures

Figure 1 shows the procedures for this study. Three tests (pretest, posttest, and delayed posttest) were conducted before and following two experimental sessions that were carried out over two successive days, and the results were analyzed to reveal the effect of the use of chatbots for DA on vocabulary acquisition. Each group took the three tests using the same format. First, a pretest implemented two weeks before the experimental class revealed words that were unknown to the learners, and these words were selected as target words for the experiment. Next, during the two experimental sessions implemented using learners' normal classroom time, participants were asked to read texts and to identify the meaning of underlined target words. The CA-DA group received graduated chatbot assistance, while the CA-NDA group only received target word definitions from the chatbots. The control group did not utilize chatbots. The three groups were allotted the same amount of time, 25 minutes, for each treatment session. Last, to examine vocabulary acquisition, a posttest was implemented immediately after the second experimental session on the same day, and a delayed posttest was conducted two weeks after the second session to examine vocabulary retention. The results of the tests were statistically analyzed for each group. Additionally, interactions between learners and the chatbots for the CA-DA sessions were transcribed verbatim and both quantitatively and qualitatively analyzed to reveal implications for vocabulary learning and diagnosis of learner abilities.

3. Target vocabulary and materials

Considering that this research aimed to explore the effect of CA-DA on vocabulary learning, it was necessary to identify words that were
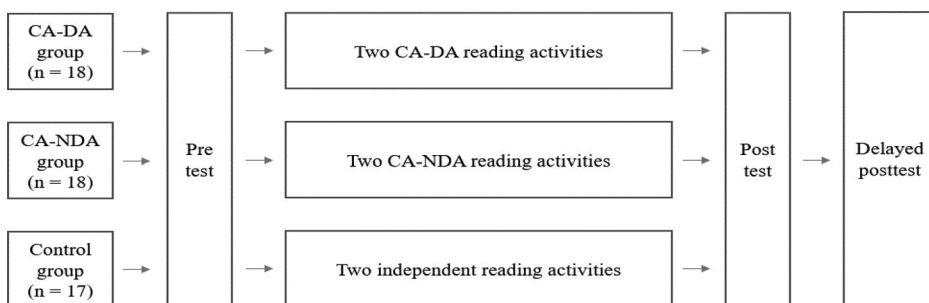
Figure 1. Study design.

unknown to the learners before creating the chatbots. To this end, a pretest was administered which contained 32 possible target items derived from reading material selected for the treatment sessions. The learners were asked to provide any written knowledge they had of the selected items; that is, they were required to either recognize or produce the written forms of the words. Twenty-two items that the learners were deemed to know were excluded, and the 10 remaining items identified as unknown to the learners were chosen as target items for the treatment sessions. The target items were as follows: *blizzard, demon, jar, jewelry, kernel, moisture, pot, refrigerator, sand,* and *torch.*

Two sections of a book designed for young EFL learners, The Popcorn Book (de Paola, 1978), were selected and revised by the researcher into reading material for the two treatment sessions. To keep the frequency of each target word consistent, the researcher adapted the original text and revised complex structures that were deemed to be beyond the learners' level. The final modified materials consisted of 209 and 221 words, respectively, including the 10 target words that were manipulated to appear at least once in each text.

4. Chatbot technology in this study and mediating chatbot design

The chatbots in this research were built with Google's open-source chatbot builder, Dialogflow, and were uploaded to tablet PCs. With the chatbots running on tablet PCs, learners could interact with the chatbots by typing their answers and reading chatbot utterances, which means that the interaction was carried out in the written form while learners were reading materials. Initially, the researcher planned to use an audio-based interaction feature that is also available in Dialogflow. While preparing the pilot study, it was discovered that the participants needed to receive word definitions in their L1 due to their limited English language proficiency. However, chatbots created using the platform did not work naturally in the audio mode when using two languages; they

only worked naturally with two languages in the written mode. Therefore, the researcher decided to exclusively use written interaction in which learners and chatbots could interact using both languages, Korean and English.

To develop prompts that would be fed into the chatbots, the researcher followed the process that previous literature utilized (Poehner, Zhang, & Lu, 2015). First, the researcher created an initial set of prompts which gradually became more explicit based on literature that had used DA for vocabulary (Camilleri & Botting, 2013; Rassaei, 2020) and Nation's (1990) five-step procedure designed to help learners identify the meanings of unknown words. Afterward, for the pilot study, the initial prompts were tested with five learners who were randomly selected from among those who were not involved in the experimental sessions. The pilot study was performed through a chat application on tablet PCs to provide as similar an environment as possible to the experimental sessions. During the pilot study in which interaction also occurred in the written form, the researcher took the role of the chatbots and asked the learners to perform the reading activities while mediating the learners based on the initial version of the prompts. This process was used to compensate for the inflexibility of pre-scripted mediation and to reflect learner needs (Poehner & van Compernolle, 2013). The whole process was recorded and used as data for building the chatbots. Considering the learner responses to each prompt, a final version of the prompts that were ultimately fed into the chatbot was created, as detailed in Table 1 and Figure 2.

The prompts were arranged from most implicit to most explicit. Under CA-DA conditions, after the initial question (What's the meaning of *target word*?), the learners received prompts one by one from a Level 1 prompt (most implicit) to a Level 4 prompt (most explicit), depending on the learners' requirements. In contrast, under CA-NDA conditions, the learners were first required to identify a target item in response to the initial question, and if they failed, the chatbots were designed to provide the correct answer. That is, the CA-DA group dynamically used prompts, but the CA-NDA group used non-dynamic prompts.

Table 1. Prompt plan.

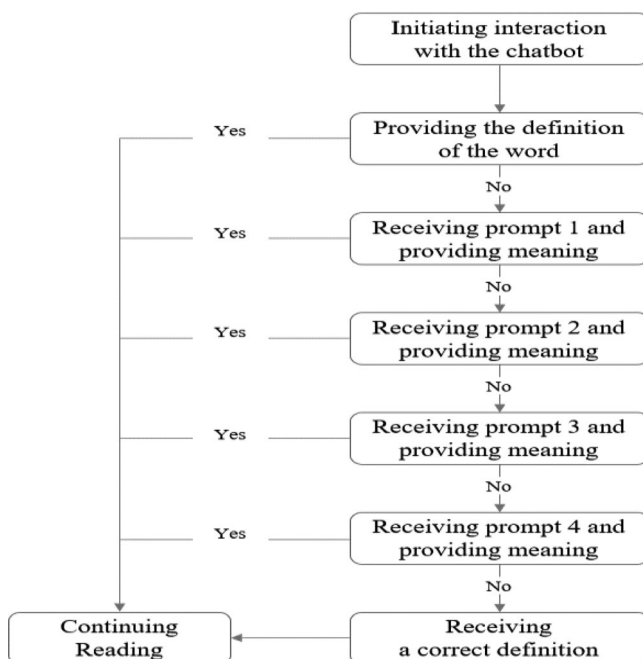| Level	Prompt	Point
1	Asking a learner to read the sentence again and to guess the meaning	4
2	Highlighting a portion of the text and asking a learner to consider it and reattempt	3
3	Providing a leading question/clue about the relationship between the text portion and the word	2
4	Providing another sentence that includes the word in a new, more specific context	1

Figure 2. Prompt sequence from the learner perspective.

5. Measures

Considering the mode of interaction carried out, the written mode, the researcher developed two vocabulary tests to assess two written aspects of vocabulary learning: receptive and productive. The formats of each test were adapted from Fuente's study (2003), with new target words used to meet the purpose of the current study. The productive test was implemented before the receptive test for both the posttest and delayed posttest to avoid a test effect (Fuente, 2003). First, for the productive vocabulary test, learners received a sheet with images of the 10 target vocabulary words and were asked to write the corresponding word in English. With each image, the first letter of the corresponding target word was supplied along with the number of total letters as cues (see Appendix 1). Spelling errors were ignored as long as the word was intelligible. For scoring, the number of words for which a learner provided a correct answer was counted and reported as their score. Next, for the receptive vocabulary test, the learners were presented with the target words and were asked to provide either Korean or English definitions (see Appendix 2). For scoring, the number of words for which a learner provided correct Korean or English definition was calculated as their score. Another SLA professional was asked to participate in scoring the tests, and 98% interrater reliability was achieved. The test items were also checked for internal consistency utilizing Cronbach

alpha. The researcher applied binominal scoring of 1 (correct) and 0 (incorrect) for tests. Alpha levels were confirmed to exceed the widely accepted standard of 0.70 (Lance, Butts, & Michels, 2006).

To diagnose vocabulary learning, interaction records between learners and chatbots obtained from the CA-DA group were analyzed, and a scoring system, shown in Table 1, was used to diagnose learner abilities (Poehner & Lantolf, 2013; Poehner et al., 2015). Two scores for each learner, an actual and a mediated score for each CA-DA session, were calculated and compared. To determine the actual score, only unmediated responses were considered; that is, learners received either 5 or 0 points for each word depending on whether they received mediation from the chatbots or not. For example, if a learner was able to provide a correct answer in response to the initial question that only asked for the definition of a word without requiring the prompts in Table 1, the learner was awarded 5 points for that word. If the learner required any prompts after not being able to answer the first question, 0 points were given for that word. On the other hand, to obtain the mediated score, the level of mediation required was taken into consideration. For example, if the learner did not require any prompts and successfully responded to the initial question that asked for the definition of the word, 5 points were given for that word, as was the case with the determination of the actual score. However, in the case of a successful response that required the use of prompts, a different scoring system was applied. If the learner required only a Level 1 prompt, which is the most implicit, the learner was given 4 points. For a level 2 prompt, 3 points were given, and finally, for a level 4 prompt, 1 point was given. If the learner could not provide the correct answer with the help of the chatbot, the chatbot provided the learner with the meaning of the target word, and the learner received 0 points. Thus, the total mediated score for one learner was calculated by adding up all points assigned for each item.

6. Data analysis

Both quantitative and qualitative data analysis methods were performed in this study. First, scores from two posttests were statistically analyzed to explore the effect of CA-DA on receptive and productive vocabulary acquisition in learners. To be more specific, one-way ANOVAs were conducted using the scores to identify the differences in the two post-tests among the three groups as a result of the experimental sessions. Scheffe's post hoc analysis was also performed to locate the differences among the groups. Second, interaction transcripts between learners and the chatbots for two CA-DA sessions were both quantitatively and qualitatively analyzed to reveal potential implications for vocabulary learning

Table 2. Descriptive statistics for the acquisition of receptive vocabulary knowledge.

Group	N	Posttest		Delayed posttest	
		M	SD	M	SD
CA-DA	18	7.94	1.95	7.17	2.23
CA-NDA	18	5.89	1.78	5.06	1.43
Control	17	1.88	1.11	1.41	1.00

Table 3. Descriptive statistics for the acquisition of productive vocabulary knowledge.

Group	N	Posttest		Delayed posttest	
		M	SD	M	SD
CA-DA	18	6.33	2.30	4.56	1.79
CA-NDA	18	4.28	1.96	3.17	1.86
Control	17	2.06	1.34	1.41	1.12

and diagnosis of learner abilities. Specifically, mediated and actual scores were calculated and compared between the two CA-DA sessions to diagnose learner abilities. Qualitative analysis of interaction data between learners and the chatbots was also conducted to explore learners' moment-to-moment changes and development across the two consecutive sessions (Lantolf, 2000).

IV. Results

1. The effect of CA-DA on vocabulary knowledge

Tables 2 and 3 present descriptive statistics for the acquisition of vocabulary knowledge by group. Figure 3 displays the groups' overall receptive and productive vocabulary test performances. Learners' scores for the pretest were 0 because words for which the learners did not possess any knowledge were selected as the target words.

As shown in Figure 3, in both the posttest and delayed posttest for receptive vocabulary knowledge, the CA-DA group outperformed the other two groups, and the CA-NDA group outperformed the control group in both tests. To identify whether statistical differences existed among the groups, a one-way ANOVA was carried out with the treatment conditions set as the independent variable and learners' scores as the dependent variable. It was confirmed that the distribution of the scores for each group was normal. The results revealed statistical differences among the three groups. That is, in the posttest and delayed posttest, the statistical results were $F (2, 50) = 59.696$, $p < .05$, $\eta^2 = 0.70$ and $F (2, 50) = 54.425$, $p < .05$, $\eta^2 = 0.69$, respectively. To locate the differences, Scheffe's post hoc analysis was run, and the results revealed that the two chatbot-assisted groups acquired receptive knowledge more effectively than the control group ($p < .05$), and more importantly, the CA-DA group statistically outperformed the CA-NDA group in both the posttest and delayed posttest ($p < .05$).
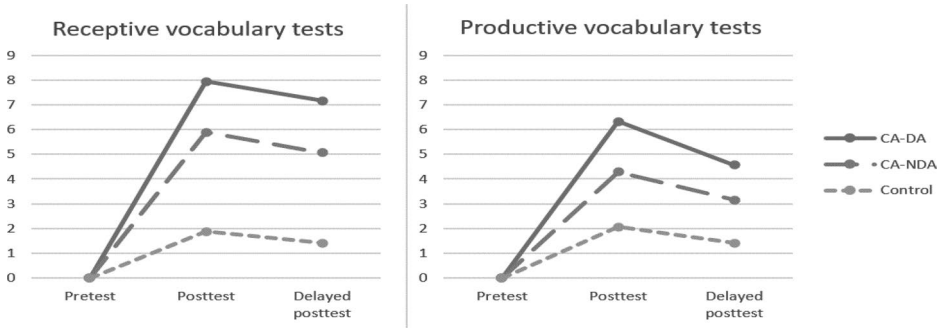
Figure 3. The groups' overall receptive and productive vocabulary test performances.

Regarding immediate productive vocabulary acquisition measured by the posttest, as shown in Figure 3, the CA-DA group outscored the other two groups, and the CA-NDA group outperformed the control group. Similarly, in regard to the delayed posttest implemented to examine vocabulary retention, the CA-DA group outperformed the other two groups, and the CA-NDA group outscored the control group. To examine the statistical differences regarding productive vocabulary knowledge among the groups, a one-way ANOVA was run on participants' scores. The results revealed statistical differences among the three groups, F (2, 50) = 21.640, $p < .05$, $\eta^2 = 0.46$. Scheffe's posttest comparisons also indicated an advantage in both chatbot-assisted groups over the control group ($p < .05$). In addition, the CA-DA group acquired productive vocabulary knowledge more effectively than the CA-NDA group ($p < .05$). The same results were obtained in the delayed posttest. That is, there were statistical differences among the three groups, F (2, 50) = 16.277, $p < .05$, $\eta^2 = 0.39$. More specifically, it was observed that both chatbot-assisted groups acquired productive vocabulary knowledge in the written form more effectively than the control group ($p < .05$), with the CA-DA group outperforming the CA-NDA group ($p < .05$).

2. Diagnosis of learner abilities based on CA-DA

To diagnose learner abilities that could be detected during the CA-DA treatment sessions, the actual and mediated scores of learners from the CA-DA group across both sessions were comparatively analyzed. Table 4 displays the actual and mediated scores of four selected learners for each session as well as the overall mean and standard deviation of the actual and mediated scores of all learners for each session.

Increases in the mean actual and mediated scores from session 1 to session 2 revealed that learners in the CA-DA group improved in terms of both independent and mediated performance. Regarding individual learner results, Learners 1 and 2 received the same actual scores in

Table 4. Actual and mediated scores for the CA-DA group.

Learner		Actual score		Mediated score	
		Session 1	Session 2	Session 1	Session 2
1		5	20	24	48
2		5	10	18	42
3		0	5	19	34
4		0	5	9	20
…..					
Total ($N = 18$)	M	1.11	8.61	15.78	28.89
	SD	2.14	5.89	5.46	8.73

session 1; therefore, if one takes a static perspective for assessment, it may be assumed that their levels of performance were the same. However, regarding their mediated scores in the same session, the two learners showed different levels of performance for the same word as can be observed in Excerpts 1 and 2 from session 1.

Excerpt 1

Session 1

1. Learner 1: kernel

2. Chatbot: What's the meaning of kernel? (initial question)

3. Learner 1: ??

4. Chatbot: Please, read the sentence again and try to guess the meaning of kernel. What is the meaning? (prompt 1)

5. Learner 1: More hint, please.

6. Chatbot: Kernels were found in Peru that could still "be popped." Can you guess the meaning of kernel? (prompt 2)

7. Learner 1: Seed?

8. Chatbot: That's correct. The kernel of a nut is the part inside the shell. Please, continue reading. kernel은 견과류 중심에 있는 씨앗이나 알맹이를 뜻합니다. 계속 읽기를 진행해주세요!

[The Korean part is a translation of the English explanation]

Excerpt 2

Session 1

1. Learner 2: kernel

2. Chatbot: What's the meaning of kernel? (initial question)

3. Learner 2: What is that?

4. Chatbot: Please, read the sentence again and try to guess the meaning of kernel. What is the meaning? (prompt 1)

5. Learner 2: I don't know.

6. Chatbot: Kernels were found in Peru that could still "be popped." Can you guess the meaning? (prompt 2)

7. Learner 2: Difficult.

8. Chatbot: Kernels can be "cooked" and popped. Can you tell me the meaning? (prompt 3)

9. Learner 2: What's that?

10. Chatbot: Read this sentence. "We bought some kernels to make popcorn at home" Can you guess the meaning? (prompt 4)

11. Learner 2: 씨앗 [Seed?]

12. Chatbot: That's correct. The kernel of a nut is the part inside the shell. Please, continue reading. kernel은 견과류 중심에 있는 씨앗이나 알맹이를 뜻합니다. 계속 읽기를 진행해주세요!
[The Korean part is a translation of the English explanation]

Learner 1 provided the correct definition of *kernel* after receiving a Level 2 prompt, while Learner 2 provided the correct definition of the same word after receiving a Level 4 prompt. Each learner received a different mediated score for each word depending on the amount of assistance they required from the chatbot.

Furthermore, qualitative analysis of mediated interaction between the learners and the chatbots provided evidence regarding vocabulary development within the ZPD across the two consecutive sessions. When compared to Excerpt 2, Excerpt 3, which was obtained from session 2, revealed how Learner 2 responded differently to the chatbot's prompts for the same word across both sessions.

Excerpt 3

Session 2

1. Learner 2: kernel

2. Chatbot: What's the meaning of kernel?

3. Learner 2: popcorn?

4. Chatbot: Please, read the sentence again and try to guess the meaning of kernel. What is the meaning? (prompt 1)

5. Learner 2: 모르겠음 [I don't know]

6. Chatbot: "When the kernel is heated, the core becomes bigger until it pops." Can you guess the meaning? (prompt 2)

7. Learner 2: popcorn's seed.

8. Chatbot: That's correct. The kernel of a nut is the part inside the shell. Please, continue reading. kernel은 견과류 중심에 있는 씨앗이나 알맹이를 뜻합니다. 계속 읽기를 진행해주세요!
[The Korean part is a translation of the English explanation]

In session 1, although Learner 2 could not identify *kernel* without prompts in response to the initial question, he could identify the correct definition of the word after receiving the fourth prompt that was the most explicit based on Table 1. In session 2, where the learner read a different text, the learner was also unable to provide the definition in response to the initial question. However, he was able to supply the correct definition after receiving only two prompts even though he encountered the word in a different context. This would not be detectable if a teacher were to only consider actual scores.

In addition, both Learners 3 and 4 scored 0 for their actual scores, which indicates that they were not able to offer the correct definitions for the target words without prompts from the chatbots in session 1. However, with the help of the chatbot, Learner 3 was able to perform the task successfully in session 2 with only 16 prompts, as could be confirmed with the learner's mediated score. Learner 4 was also able to provide the definitions of the words with the assistance of the chatbot with a total of 30 prompts. Based on their mediated scores, it can be argued that they possessed a different starting point in terms of vocabulary learning ability; therefore, different types of subsequent learning and teaching should be prepared for each learner. Subsequent learning can be further supported by the information regarding learner performance for each word, as shown in Figure 4. For example, after the teacher realized that Learner 4 had more difficulty than Learner 3 in identifying the definition of a word, for example, *pot*, the teacher could provide Learner 4 more specific examples related to the word while
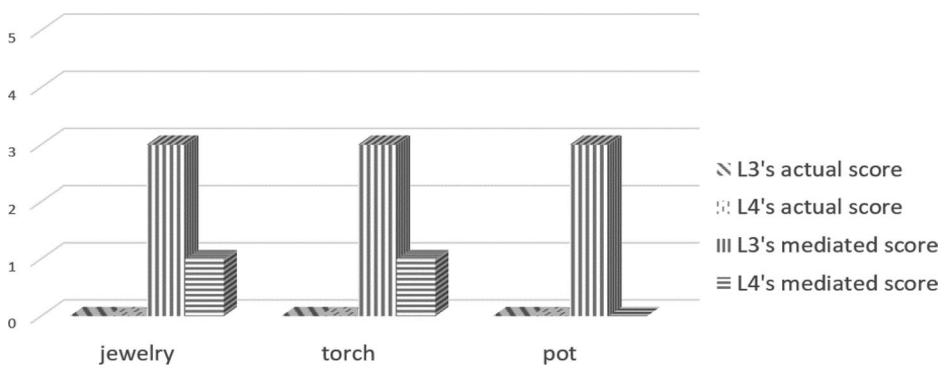
(L3: Learner 3, L4: Learner 4)

Figure 4. Actual and mediated scores for learners 3 and 4: jewelry, torch, and pot.

preparing more complex and challenging sentences that included the word for Learner 3 after the DA period.

V. Discussion

Through the use of chatbots, the effect of CA-DA on vocabulary learning was investigated from two perspectives: 1) learning and 2) diagnosis. The CA-DA group outscored the CA-NDA and control groups in both the receptive and productive tests, with statistical differences being revealed. More importantly, qualitative analysis of the interaction between chatbots and learners in the CA-DA group provided evidence for learner development across two different tasks. The Sociocultural Theory's ZPD and Cognitive Load Theory discussed in the Background literature are supportive of these results.

First, the results can be explained in terms of the effect of ZPD-based mediation. According to previous DA studies which have adopted graduated prompt systems, graduated prompts were confirmed to be facilitative of learners' sensitive learning levels for development (i.e., ZPD) (e.g., Davin, 2016; Lantolf & Poehner, 2011; Poehner, 2008). Some literature also demonstrated the positive effects of ZPD-based mediation on language development through technology such as computer software or mobile application (Ebadi et al., 2018; Rassaei, 2021; Yang & Qian, 2020). Along the same vein, the present study was designed to provide graduated prompts to learners through the use of chatbots. The learners received prompts for unknown items depending on the extent to which they were needed to enable learners to identify the definition of a word. For example, Learner 2 initially required four prompts to identify the definition of *kernel*, but in session 2, which used a different text, the learner could supply the definition after receiving only two prompts that were less explicit than the ones needed in session 1. This change toward more self-regulation during negotiation with the chatbot can be interpreted as learner growth (Aljaafreh & Lantolf, 1994; Lantolf & Thorne, 2006). In addition, the significant increase in learners' mediated scores from sessions 1 to 2 can also be taken as an indication that effective vocabulary learning occurred in the CA-DA group (Ebadi & Rahimi, 2019; Rassaei, 2021). In sum, graduated prompts used in the CA-DA group successfully targeted the learners' ZPD and led to more effective acquisition and development of vocabulary than non-dynamic or no assistance in the other two groups. Therefore, this study demonstrated the positive effect of the use of chatbots for DA and supported the findings of previous studies that ZPD-based mediation can facilitate learner development by tailoring instruction to learners' specific needs.

Another explanation for more effective vocabulary learning in the CA-DA group compared to the CA-NDA group can be explained through Cognitive Load Theory. When identifying unknown words from context, learners have to process various elements, such as finding contextual clues or substituting their guess for the unknown word (Nation, 1990). Such a process could be cognitively challenging to L2 learners with limited language proficiency (Teng, 2020; Türk & Erçetin, 2014).

However, in this study, the graduated characteristic of the chatbot prompts helped learners to successfully identify unknown words and facilitated more vocabulary acquisition by preventing cognitive overload, which is consistent with previous studies that demonstrated the positive effect of controlling cognitive overload on vocabulary learning (deHaan, Michael Reed, & Kuwada, 2010; Ramezanali & Faez, 2019; Teng, 2020). The graduated prompts provided by the chatbots in the current study allowed the learners to begin work on a simpler inferential task and progress toward a more complex inferential task, which effectively managed cognitive load (van Merriënboer et al., 2003). To be more specific, the graduated prompts contributed to a decrease in intrinsic cognitive load because the inherent complexity of the inferential task decreased every time the prompt was provided for the learners; this created more effective conditions for vocabulary learning. For example, the two inferential tasks that Learner 3 experienced for the target word *torch* demonstrated how graduated prompts contributed to the control of task complexity. During session 1, the learner received four prompts that were necessary to identify the meaning of *torch*. While reading a different text in session 2, as shown in Figure 4, the learner identified the same word with only two prompts, which indicated that more information compared to session 1 was managed by the learner himself; but he successfully identified the meaning without cognitive overload. In other words, the graduated mediation from the chatbots effectively controlled intrinsic cognitive load, so more working memory could be allotted to the germane cognitive load (i.e., vocabulary acquisition). In contrast, the CA-NDA group was only given the definitions by the chatbots, so the intrinsic cognitive load was not adequately controlled. Therefore, it can be argued that more effective vocabulary learning conditions were created for the CA-DA group based on Cognitive Load Theory.

In addition to the effectiveness of CA-DA in terms of vocabulary learning, the process of CA-DA generated individual learner diagnostic information about language abilities, as demonstrated in previous DA studies in the field of CALL (Mehri Kamrood et al., 2019; Qin & Van Compernolle, 2021; Rassaei, 2020; Zhang & Lu, 2019). These studies utilized either computer software or mobile application to generate

individual learner diagnostic profiles. Unlike the literature, the current study employed chatbot technology to create diagnostic information. Specifically, it was confirmed that the CA-DA group could provide diagnostic information about learner abilities which was not available from the CA-NDA group. Even though two learners in the CA-DA group achieved the same test results, their different mediated scores revealed that one learner experienced more impressive gains than the other when external mediation was introduced (Vygotsky, 1978). Teachers can use this information to make pedagogical decisions for subsequent learning (Poehner et al., 2015). For example, Learners 1 and 2 might have been assessed at the same level if tested statically but graduated mediation revealed that Learner 2 had more difficulty providing the correct definition of *kernel* than Learner 1, as can be observed in Excerpts 1 and 2. This type of diagnostic information is meaningful because teachers can use the information to prepare subsequent learning finely attuned to an individual's problem areas (Qin & Van Compernolle, 2021; Zhang & Lu, 2019).

VI. Pedagogical implications

This study has important pedagogical implications for the practices of teachers, chatbot developers, and other practitioners. First, by utilizing chatbots, this study attempted to provide each learner with interactive vocabulary learning opportunities and attempted to overcome some of the practical constraints of DA that teachers may face in the classroom, such as large-class size. Furthermore, unlike previous DA literature, where a mediator physically maintained record sheets for keeping track of learner performance (e.g., Herazo et al., 2019), teachers can take advantage of automatically transcribed interaction records available through the open-source chatbot builder, Dialogflow. This use of chatbot technology will help minimize the labor intensity required of teachers when implementing and recording DA, and thus, will both make it more feasible for teachers to implement DA to multiple learners simultaneously and help them prepare subsequent teaching which is individually fine-tuned according to DA results.

Second, specific implications for chatbot developers regarding the design of chatbots for language learning can be drawn from this study. Unlike other research that only utilized chatbots equipped with limited response patterns, this study created chatbots that were able to provide ZPD-based mediation and demonstrated its positive effect on vocabulary learning; thus, introducing this type of chatbot as one more valuable tool for language learning. In addition, chatbots created using Dialogflow did not work naturally in the audio mode when using two languages;

they only worked naturally with two languages in the written mode. A chatbot that can communicate both in L1 and L2 using both the audio and written modes might serve as a better teaching tool for beginner-level learners such as young EFL students in this study.

Last, by introducing the concept of CA-DA, this research presents a new research avenue for the DA research community and, in a broader perspective, for the growing field of technology-enhanced language assessment. Scholars in the field of education technology have suggested that AI agents could be a powerful learning tool; some have already provided empirical evidence regarding their role and positive impact on language learning (e.g., Xu, Wang, Collins, Lee, & Warschauer, 2021). This research further advances our pedagogical understanding regarding the use of AI technology in that it employed AI chatbots not only as learning assistant tools but also as DA mediators that were able to provide graduated prompts focusing on vocabulary learning and diagnosis.

VII. Conclusion

This research demonstrated that chatbots could be employed as learning tools for facilitating SLA (Bibauw, François, & Desmet, 2019), specifically for vocabulary learning. By applying classroom DA through chatbot-assisted activities, this study confirmed that CA-DA could promote vocabulary acquisition and provide diagnostic information about learner abilities during the process of learning. In other words, it showed the potential that chatbot technology holds for the integration of assessment and learning. More specifically, it was confirmed that chatbots provided several advantages when implementing DA in the classroom setting. First, due to human-like features that allowed learners to interact with chatbots, automated interactions that occurred in the DA sessions approximated ones we could expect from interactions between human mediators and learners. These features created effective conditions for vocabulary learning through DA. One could expect the further potential for automated interaction in other areas of SLA beyond vocabulary learning. Next, one-to-one interaction was implemented for 18 learners simultaneously in this study through the use of chatbots, which took around 25 minutes in total. Although the chatbots built for this study were exclusively utilized by the researcher for 18 learners in the classroom, it is possible to apply the chatbots to a larger number of learners if the educational context is considered appropriate. Last, chatbots could be operated in a mobile-mediated setting with teachers sharing the software on the internet and learners uploading it to their own mediums such as smartphones. In this setting, learners can interact with chatbots ubiquitously, anytime, anywhere (Fryer et al., 2020). The use of chatbots

through different devices for other areas of language learning beyond DA seems to be worth investigating in future research.

Limitations in this study call for caution when interpreting the findings. First, given that the time interval between the pretest and posttest and the posttest and delayed posttest was two weeks each, a practice effect might have existed. Tests with a time interval of three to six weeks between tests would yield more valid results (Brown, Irving, & Keegan, 2008). Second, the images or instructions included in the tests might have influenced the test results. Future studies need to employ more systematic testing methods examined by experts in terms of content validity. Third, the results were derived from the implementation of only two reading activities during two experimental sessions. Expanding the number of activities and experimental sessions is required to generalize the results. Fourth, the participants were young EFL learners within a single context. This may not accurately reflect the learning of the larger learner population. Finally, chatbot interaction in this study was implemented using the written form in both L1 and L2. Future studies may include more varied types of chatbot interaction such as audio-based interactions or interactions using only L2.

Acknowledgments

Disclosure statement

Notes on contributor

Jaeho Jeon is a doctoral student of English Education at Seoul National University of Education and a primary school English teacher in South Korea. His research interests include dynamic assessment, teacher education, and computer-assisted language learning.

ORCID

Jaeho Jeon http://orcid.org/0000-0002-1161-3676

References

Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning*, *21*(3), 199–226. doi:10.1080/09588220802090246

Ai, H. (2017). Providing graduated corrective feedback in an intelligent computer-assisted language learning environment. *ReCALL*, *29*(3), 313–334. doi:10.1017/S095834401700012X

Aljaafreh, A., & Lantolf, J. P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *The Modern Language Journal*, *78*(4), 465–483. doi:10.2307/328585

Andujar, A. (2020). Mobile-mediated dynamic assessment: A new perspective for second language development. *ReCALL*, *32*(2), 178–194. doi:10.1017/S0958344019000247

Bibauw, S., François, T., & Desmet, P. (2019). Discussing with a computer to practice a foreign language: Research synthesis and conceptual framework of dialogue-based CALL. *Computer Assisted Language Learning*, *32*(8), 827–877. doi:10.1080/09588221.2018.1535508

Bowles, M. A. (2004). L2 glossing: To CALL or not to CALL. *Hispania*, *87*(3), 541–552. doi:10.2307/20063060

Brown, G. T. L., Irving, S. E., & Keegan, P. J. (2008). *An introduction to educational assessment, measurement, and evaluation: Improving the quality of teacher-based assessment* (2nd ed.). NZ: Pearson Education.

Camilleri, B., & Botting, N. (2013). Beyond static assessment of children's receptive vocabulary: The dynamic assessment of word learning (DAWL). *International Journal of Language & Communication Disorders*, *48*(5), 565–581. doi:10.1111/1460-6984.12033

Davin, K. J. (2013). Integration of dynamic assessment and instructional conversations to promote development and improve assessment in the language classroom. *Language Teaching Research*, *17*(3), 303–342. doi:10.1177/1362168813482934

Davin, K. J. (2016). Classroom dynamic assessment: A critical examination of constructs and practices. *The Modern Language Journal*, *100*(4), 813–829. doi:10.1111/modl.12352

Davin, K. J., Herazo, J. D., & Sagre, A. M. (2017). Learning to mediate: Teacher appropriation of dynamic assessment. *Language Teaching Research*, *21*(5), 632–651. doi:10.1177/1362168816654309

de Paola, T. (1978). *The popcorn book*. Holiday House.

deHaan, J., Michael Reed, W. M., & Kuwada, K. (2010). The effect of interactivity with a music video game on second language vocabulary recall. *Language Learning and Technology*, *14*(2), 74–94.

Ebadi, S., & Bashir, S. (2021). An exploration into EFL learners' writing skills via mobile-based dynamic assessment. *Education and Information Technologies*, *26*(2), 1995–2016. doi:10.1007/s10639-020-10348-4

Ebadi, S., & Rahimi, M. (2019). Mediating EFL learners' academic writing skills in online dynamic assessment using Google Docs. *Computer Assisted Language Learning*, *32*(5-6), 527–555. doi:10.1080/09588221.2018.1527362

Ebadi, S., Weisi, H., Monkaresi, H., & Bahramlou, K. (2018). Exploring lexical inferencing as a vocabulary acquisition strategy through computerized dynamic assessment and static assessment. *Computer Assisted Language Learning*, *31*(7), 790–817. doi:10.1080/09588221.2018.1451344

Fryer, L. K., Coniam, D., Carpenter, R., & Lăpușneanu, D. (2020). Bots for language learning now: Current and future directions. *Language Learning & Technology*, *24*(2), 8–22. https://www.lltjournal.org/item/3143.

Fuente, M. J. (2003). Is SLA interactionist theory relevant to CALL? A study on the effects of computer-mediated interaction in L2 vocabulary acquisition. *Computer Assisted Language Learning*, *16*(1), 47–81. doi:10.1076/call.16.1.47.15526

Google. (2020). *Dialogflow ES documentation*. Retrieved from https://cloud.google.com/dialogflow/es/docs.

Heift, T. (2017). History and key developments in intelligent computer-assisted language learning (ICALL). In S. L. Thorne & S. May (Eds.), *Language, education and technology: Encyclopedia of language and education* (3rd ed.). Springer. doi:10.1007/978-3-319-02237-6_23

Herazo, J. D., Davin, K. J., & Sagre, A. M. (2019). L2 dynamic assessment: An activity theory perspective. *The Modern Language Journal*, *103*(2), 443–458. doi:10.1111/modl.12559

Hulstijn, J. (1992). Retention of inferred and given word meanings: Experiments in incidental vocabulary learning. In P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 113–125). Macmillan.

Hulstijn, J., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal*, *80*(3), 327–339. doi:10.1111/j.1540-4781.1996.tb01614.x

Ko, H. (2005). Glosses, comprehension, and strategy use. *Reading in a Foreign Language*, *17*(2), 125–143.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*(2), 202–220. doi:10.1177/1094428105284919

Lantolf, J. (2000). Second language learning as a mediated process. *Language Teaching*, *33*(2), 79–96. doi:10.1017/S0261444800015329

Lantolf, J. P., & Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics*, *1*(1), 49–72. doi:10.1558/japl.v1i1.647

Lantolf, J. P., & Poehner, M. E. (2011). Dynamic assessment in the classroom: Vygotskian praxis for L2 development. *Language Teaching Research*, *15*(1), 11–33. doi:10.1177/1362168810383328

Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development.* Oxford University Press.

Lee, J. H., Yang, H., Shin, D., & Kim, H. (2020). Chatbots. *ELT Journal*, *74*(3), 338–344. doi:10.1093/elt/ccaa035

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.

Mehri Kamrood, A., Davoudi, M., Ghaniabadi, S., & Amirian, S. M. R. (2019). Diagnosing L2 learners' development through online computerized dynamic assessment. *Computer Assisted Language Learning*, 1–30. Advance online publication. doi:10.1080/09588221.2019.1645181

Nation, I. S. P. (1990). *Teaching and learning vocabulary.* Newbury House.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13. https://jalt-publications.org/tlt/issues/2007-07_31.7.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1–4. doi:10.1207/S15326985EP3801_1

Poehner, M. E. (2008). *Dynamic assessment: A Vygotskian approach to understanding and promoting L2 development.* Springer.

Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment (C-DA). *Language Teaching Research*, *17*(3), 323–342. doi:10.1177/1362168813482935

Poehner, M. E., & Leontjev, D. (2020). To correct or to cooperate: Mediational processes and L2 development. *Language Teaching Research*, *24*(3), 295–316. doi:10.1177/1362168818783212

Poehner, M. E., & van Compernolle, R. (2013). L2 development around tests: Learner response processes and dynamic assessment. *International Review of Applied Linguistics in Language Teaching*, *51*(4), 353–377. doi:10.1515/iral-2013-0015

Poehner, M. E., Zhang, J., & Lu, X. (2015). Computerized dynamic assessment (CDA): Diagnosing L2 development according to learner responsiveness to mediation. *Language Testing*, *32*(3), 337–357. doi:10.1177/0265532214560390

Qin, T., & Van Compernolle, R. A. (2021). Computerized dynamic assessment of implicature comprehension in L2 Chinese. *Language Learning & Technology*, *25*(2), 55–74. http://hdl.handle.net/10125/73433.

Ramezanali, N., & Faez, F. (2019). Vocabulary learning and retention through multimedia glossing. *Language Learning & Technology*, *23*(2), 105–124. doi:10125/44685

Rassaei, E. (2020). Effects of mobile-mediated dynamic and nondynamic glosses on L2 vocabulary learning: A sociocultural perspective. *The Modern Language Journal*, *104*(1), 284–302. doi:10.1111/modl.12629

Rassaei, E. (2021). Implementing mobile-mediated dynamic assessment for teaching request forms to EFL learners. *Computer Assisted Language Learning*, 1–31. Advance online publication. doi:10.1080/09588221.2021.1912105

Rott, S., Williams, J., & Cameron, R. (2002). The effect of multiple-choice glosses and input-output cycles on lexical acquisition and retention. *Language Teaching Research*, *6*(3), 183–222. doi:10.1191/1362168802lr108oa

Shute, V., & Ventura, M. (2013). *Stealth assessment: Measuring and supporting learning in video games*. MIT Press. doi:10.7551/mitpress/9589.001.0001

Teng, F. (2020). Vocabulary learning through videos: Captions, advance-organizer strategy, and their combination. *Computer Assisted Language Learning*, 1–33. Advance online publication. doi:10.1080/09588221.2020.1720253

Türk, E., & Erçetin, G. (2014). Effects of interactive versus simultaneous display of multimedia glosses on L2 reading comprehension and incidental vocabulary learning. *Computer Assisted Language Learning*, *27*(1), 1–25. doi:10.1080/09588221.2012.692384

van der Veen, C., Dobber, M., & van Oers, B. (2016). Implementing dynamic assessment of vocabulary development as a trialogical learning process: A practice of teacher support in primary education schools. *Language Assessment Quarterly*, *13*(4), 329–340. doi:10.1080/15434303.2016.1235577

van Merriënboer, J. J. G., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational Psychologist*, *38*(1), 5–13. doi:10.1207/S15326985EP3801_2

Vygotsky, L. S. (1978). *Mind in society. The development of higher psychological processes*. Harvard University Press.

Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021). Same benefits, different communication patterns: Comparing Children's reading with a conversational agent vs. a human partner. *Computers & Education*, *161*, 104059. doi:10.1016/j.compedu.2020.104059

Yang, Y., & Qian, D. D. (2020). Promoting L2 English learners' reading proficiency through computerized dynamic assessment. *Computer Assisted Language Learning*, *33*(5-6), 628–652. doi:10.1080/09588221.2019.1585882

Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning & Technology*, *13*, 48–67. doi:10125/44180.

Zhang, J., & Lu, X. (2019). Measuring and supporting second language development using computerized dynamic assessment. *Language and Sociocultural Theory*, *6*(1), 92–115. doi:10.1558/lst.31710

Appendix 1. *Productive test*

그림을 보고, 영어로 알맞은 뜻을 쓰세요.

[Look at the pictures and write the words in English in the blanks]

[An image was provided for each item, for a total of 10 images]

1.냉장고 r_____ (12) 2.습기 m_____(8)

3.모래 s_____ (4) 4.악마 d_____(5)

5.눈보라 b_____(8) 6.통, 담는 그릇 j_____(3)

7.씨앗, 알맹이 k_____(6) 8.보석 j_____(7)

9.횃불 t_____(5) 10.항아리 p_____(3)

Appendix 2. *Receptive test*

아래 영어 단어에 대해 동의어나, 의미 혹은 단어의 정의를 영어나 한국어로 쓰세요.

[Provide an equivalent, meaning, or definition in Korean or English for the following English words]

1.jewelry _____

2.demon _____

3.pot _____

4.refrigerator _____

5.jar _____

6.blizzard _____

7.moisture _____

8.sand _____

9.torch _____

10.kernel _____