Tatsuya Nakata*

Effects of retrieval formats on second language vocabulary learning

DOI 10.1515/iral-2015-0022

Abstract: The present study set out to examine how we can optimize paired-associate learning of second language (L2) vocabulary. In paired-associate learning, retrieval, where learners are required to access information about an L2 word from memory, is found to increase vocabulary learning. Retrieval can be categorized according to dichotomies of (a) recognition versus recall and (b) receptive versus productive. In order to identify the optimal retrieval format, the present study compared the effects of the following four conditions: recognition, recall, hybrid (combination of recall and recognition), and productive recall only. In this study, 64 English-speaking college students studied 60 Swahili-English word pairs using computer-based flashcard software. Results suggested that for paired-associate learning of L2 vocabulary, (a) recall formats are more effective than recognition for the acquisition of productive knowledge of orthography and (b) recognition formats are more desirable than recall when knowledge of spelling is not required.

Keywords: vocabulary learning, retrieval, recognition, recall, receptive / productive learning

Second language (L2) vocabulary learning is a developmental process (e. g., Barcroft 2012; Jarvis 2009; Jiang 2000; Nation 2013). In the initial stages, L2 words are typically associated with their first language (L1) translation equivalents, a process known as the lexical association stage (Jiang 2000). In later stages, learners acquire aspects of word knowledge such as the precision of meaning, spoken form, morphology, associations, grammatical functions, collocations, constraints on use, and fluency (e. g., Jiang 2000; Nation 2013; Read 2004). From a psycholinguistic point of view, vocabulary acquisition can be conceptualized as establishing lexical (or form-level) and semantic representations and incorporating them into the mental lexicon (e. g., Elgort 2011; Elgort and Piasecki 2014; Jarvis 2009). When the lexical representations of newly learned L2 words are established, they are typically mapped onto their L1

***Corresponding author: Tatsuya Nakata,** Faculty of Foreign Language Studies, Kansai University, 3-3-35 Yamate-cho, Suita-shi, Osaka 564–8680, Japan, E-mail: nakata@kansai-u.ac.jp

translation equivalents, and conceptual representations are often accessed via L1 translations (revised hierarchical model; Kroll and Stewart 1994). As L2 proficiency develops, the links between L2 lexical representations and conceptual representations become stronger, and the latter can be activated directly from the former (Kroll and Stewart 1994).

The purpose of this study was to examine how we can optimize paired-associate learning of L2 vocabulary. Paired-associate learning refers to an instructional activity where learners deliberately attempt to form a connection between an L2 word form and its meaning such as the L1 translation (e. g., Barcroft 2007; Elgort 2011; Elgort and Piasecki 2014; Steinel et al. 2007). Flashcard (word card) learning is an example of paired-associate learning. Since paired-associate learning is primarily concerned with the lexical association stage (Jiang 2000), it does not necessarily allow learners to acquire precise phonological, morphological, or semantic properties of L2 words (Nation 2013). However, because paired-associate learning is shown to be one of the most effective (e. g., Elgort 2011; Elgort and Piasecki 2014; Laufer and Shmueli 1997; Nation 2013; Webb 2007, 2009a) and common (e. g., Nakata 2011; Schmitt 1997; Wissman et al. 2012) vocabulary learning techniques, the current study set out to examine how we can optimize paired-associate learning of L2 vocabulary. Note that due to its focus on paired-associate learning, the present study is concerned only with lexical connections between L2 words and their L1 translations, and the issue of how they are linked to conceptual representations (e. g., Elgort 2011; Elgort and Piasecki 2014; Kroll and Stewart 1994) is outside the scope of this study. As a result, vocabulary learning is operationalized as the ability to associate L2 words with their L1 translation equivalents in this study.

Research shows that retrieval increases paired-associate learning of L2 vocabulary (e. g., Barcroft 2007; Karpicke and Roediger 2008). Retrieval refers to the process of accessing information about an L2 word from memory. For instance, when learners are presented with an L2 word and asked to produce its L1 translation, they need to remember the L1 word from memory. As a result, the treatment involves retrieval. In contrast, when learners are presented with an L2 word together with its L1 translation, they are not required to access any information about the L2 word from memory. The treatment, therefore, does not involve retrieval. Given that retrieval increases learning, which kinds of retrieval formats should be used to optimize paired-associate learning of L2 vocabulary? Retrieval formats can be categorized according to dichotomies of (a) recognition versus recall and (b) receptive versus productive. Recall formats involve generating a response, whereas recognition formats involve choosing a correct response from a number of

options. Likewise, receptive retrieval involves translating from L2 to L1, whereas productive retrieval involves translating from L1 to L2. Note that the receptive versus productive dichotomy can also be regarded as the direction of association between L1 and L2 words at the lexical level (i. e., L2→L1 or L1→L2 direction).

The dichotomies of (a) recognition versus recall and (b) receptive versus productive result in the following four types of retrieval formats: receptive recall, productive recall, receptive recognition, and productive recognition (e. g., Laufer and Goldstein 2004). In receptive recall, learners are asked to produce the meaning of target words while in productive recall, they produce the target word form corresponding to the meaning provided. Receptive recognition requires learners to choose, rather than to produce, the correct meaning of target words from a number of options, whereas productive recognition requires learners to choose the target word form corresponding to the meaning provided. For example, suppose that the learner was asked to learn a Swahili word *hadithi* (*story*). Receptive recall requires learners to produce the L1 translation (*story*) of the L2 word (*hadithi*) while in productive recall, they produce the L2 word form (*hadithi*) corresponding to the L1 translation (*story*). In receptive recognition, learners choose the correct meaning of the L2 word (*hadithi*) from a number of options (e. g., *story, invoice, ornament, cinnamon*), whereas in productive recognition, learners are presented with an L1 word (*story*) and asked to select the most appropriate L2 translation from a number of options (e. g., *kaputula, nira, funumu, hadithi*).

The above four types of retrieval formats are common within a range of different vocabulary learning activities such as flashcard learning, fill-in-the-blank exercises, multiple-choice questions, and word-definition matching. Nation and Webb (2011), for instance, analyzed 12 vocabulary learning activities and note that five of them involve recall while the other two involve recognition. As for the receptive-productive dichotomy, Nation and Webb observe that out of 12 vocabulary learning activities, five of them involve receptive retrieval while the other two involve productive retrieval. These four types of retrieval formats are also common among existing vocabulary learning software where vocabulary is learned in a paired-associate format. Nakata (2011), for instance, surveyed nine popular computer-based flashcard programs and found that eight of them support receptive recognition, productive recognition, and productive recall, while six of them support receptive recall.

Despite the widespread use of the above four retrieval formats, it is not yet clear which kinds of retrieval formats should be used to optimize paired-associate learning of L2 vocabulary. For instance, is productive retrieval more effective than receptive? Which is more effective, recall or recognition?

Does a combination of recall and recognition increase paired-associate learning more than either one alone? In order to identify the optimal retrieval format, the present study compared the effects of the following four conditions: recognition, recall, hybrid, and productive recall only. In the recognition condition, target items were practiced in receptive and productive recognition formats, whereas the recall condition consisted of receptive and productive recall. In the hybrid condition, target items were studied in receptive recognition, productive recognition, receptive recall, and productive recall. In the productive recall only condition, target items were learned only in a productive recall format. Findings of this study are of value because they may help us to determine how we can optimize paired-associate learning of L2 vocabulary.

1 Review of literature

As described in the previous section, retrieval refers to the process of accessing information about an L2 word from memory. Retrieval formats can be categorized according to dichotomies of (a) recognition versus recall and (b) receptive versus productive. Recall formats involve generating a response, whereas recognition formats involve choosing a correct response from a number of options. Likewise, receptive retrieval involves translating from L2 to L1 (i. e., lexical-level link from L2 to L1), whereas productive retrieval involves translating from L1 to L2 (i. e., lexical-level link from L1 to L2). Note that when the direction of learning matches that of testing (e. g., studied productively and tested productively), it is called a forward association, whereas when the direction of learning does not match that of testing (e. g., studied productively but tested receptively), it is called a backward association (e. g., Griffin and Harley 1996).

1.1 Effects of receptive and productive retrieval

Previous studies on receptive and productive retrieval indicate that receptive retrieval promotes larger gains in receptive vocabulary knowledge while productive retrieval is more beneficial for gaining productive knowledge (e. g., Griffin and Harley 1996; Mondria and Wiersma 2004; Steinel et al. 2007; Webb 2009a, 2009b). These findings may be explained by *transfer-appropriate processing* (hereafter, TAP) *theory* (Morris et al. 1977), according to which performance is enhanced if the testing condition corresponds to that of learning. Another

explanation is the *forward asymmetry effect* (Kahana and Caplan 2002). The forward asymmetry effect refers to a phenomenon where the forward association is easier than the backward association. TAP theory and the forward asymmetry effect predict that to increase both receptive and productive vocabulary knowledge efficiently, it is valuable to practice receptive as well as productive retrieval (e. g., Griffin and Harley 1996; Mondria and Wiersma 2004; Nation 2013; Webb 2009b). The study conducted by Mondria and Wiersma (2004) offers support for this prediction. They compared the following three conditions: receptive, productive, and receptive + productive. Mondria and Wiersma found that the receptive + productive condition was (a) as effective as the productive condition on a productive posttest, (b) as effective as the receptive condition on a receptive posttest, and (c) more effective than the productive condition on a receptive posttest.

Previous studies also demonstrate that productive retrieval leads to relatively large gains in receptive knowledge, whereas receptive retrieval results in only small gains in productive knowledge (e. g., Griffin and Harley 1996; Mondria and Wiersma 2004; Schneider et al. 2002; Steinel et al. 2007; Webb 2009a). These findings may be partially explained by the desirable difficulty framework (e. g., Bjork 1994), according to which more demanding tasks lead to better long-term retention and *transfer* than less demanding ones (see Ellis 1995; Schneider et al. 2002; Steinel et al. 2007, for a similar discussion). Transfer in this context refers to a situation where the testing condition does not correspond to that of learning. For instance, if target words are practiced productively but tested receptively, the posttest measures transfer (Schneider et al. 2002). Because productive retrieval is more demanding than receptive retrieval (Mondria and Wiersma 2004; Nation and Webb 2011; Schneider et al. 2002), the desirable difficulty framework predicts that productive retrieval facilitates transfer more than the latter. This may be partly the reason why productive retrieval tends to be effective even on receptive posttests, where the testing condition does not match that of learning (Schneider et al. 2002).

Another explanation is related to the relative ease of receptive vocabulary use compared with productive (Ellis and Beaton 1993; Nation 2013). In the learners' mental lexicon, L1 words have more active links (e. g., conceptual links and paradigmatic and syntagmatic associations) than newly learned L2 words, which are typically only weakly associated with their L1 translations. Productive use (i. e., L1→L2 direction) is harder than receptive because the former involves activating a relatively weak link (L2 word) from a number of competing and active connections, whereas receptive use (i. e., L2→L1 direction) involves no or less competing paths because newly learned L2 words often have connections only to their L1 translations (Ellis and Beaton 1993; Nation 2013).

Although TAP theory (Morris et al. 1977) and the forward asymmetry effect (Kahana and Caplan 2002) predict that productive retrieval is not effective for gaining receptive knowledge, productive retrieval may nonetheless lead to relatively large gains in receptive knowledge because reception is easier than production (e. g., Ellis and Beaton 1993; Laufer and Goldstein 2004; Webb 2008).

1.2 Effects of recall and recognition

There exist conflicting views regarding the effects of recall and recognition on paired-associate learning. The desirable difficulty framework (e. g., Bjork 1994; see above) and the retrieval effort hypothesis (Pyc and Rawson 2009) suggest that recall enhances paired-associate learning more than recognition. This is because recall introduces more retrieval difficulty for the learner (e. g., Laufer and Goldstein 2004; Nation and Webb 2011), which facilitates learning according to the desirable difficulty framework and the retrieval effort hypothesis. The retrieval practice effect (Baddeley 1997; Ellis 1995), in contrast, predicts that recognition increases paired-associate learning more than recall. According to the retrieval practice effect, successful retrieval attempts lead to superior retention to unsuccessful retrieval attempts. Because recognition, which is easier than recall (e. g., Laufer and Goldstein 2004; Nation and Webb 2011), produces more successful retrievals during learning, recognition may be more effective than recall based on the retrieval practice effect. In other words, both recall and recognition have their advantage and disadvantage. Although recall is beneficial in that it introduces more retrieval difficulty, it does not necessarily produce high retrieval success. This is not desirable based on the retrieval practice effect. In contrast, while recognition tends to facilitate successful retrieval during learning, it does not necessarily help introduce retrieval difficulty. This is not effective according to the desirable difficulty framework and the retrieval effort hypothesis.

L2 vocabulary studies have produced mixed results regarding the effects of recall and recognition (Chen and Huang 2014; Van Bussel 1994; Yun et al. 2008). In Van Bussel (1994, Experiment 2), for instance, 32 speakers of Dutch studied 40 English words under recall and recognition conditions. At the beginning of the treatment, participants were presented with the target words. After the initial presentation, in the recall condition, participants were presented with a cloze sentence and asked to supply an appropriate target word to complete the sentence. In the recognition condition, participants were presented with a sentence containing a target word and judged whether the target word was used appropriately in the sentence. Learning was measured by recall and

recognition posttests. Van Bussel did not find any significant difference between the recall and recognition conditions in their posttest scores.

In Yun et al. (2008), 120 Korean students (aged 16 on average) studied 20 English words (e. g., *successful, vocabulary, mistake, require*) under recall and recognition conditions. In the recall condition, participants were asked to type the definition of target words, whereas in the recognition condition, learners were presented with a target word and asked to choose the correct definition from five options. Learning was measured by recall and transfer posttests. The former required participants to produce the meaning of target words. The latter required participants to write original sentences using target words. Yun et al. found that the recall group significantly outperformed the recognition group on both recall and transfer posttests. In Chen and Huang (2014), 157 Taiwanese sixth graders (aged 12 on average) were divided into recall and recognition groups and studied English words related to animals. Learning was measured by recall and recognition posttests. The former required participants to type the target word form corresponding to the picture provided, whereas the latter involved choosing the target word form corresponding to the picture from four options. Chen and Huang found that the recall group significantly outperformed the recognition group on the posttest.

The inconsistent findings of previous research (Chen and Huang 2014; Van Bussel 1994; Yun et al. 2008) may be in part due to at least two methodological differences. First, existing studies differ in how recall and recognition were operationalized during the treatment or on the posttest. For instance, the recognition format in Van Bussel (1994, Experiment 2) required participants to judge whether the target word was used appropriately in the sentence, whereas the recognition format in Yun et al. (2008) involved choosing the correct definition of target words from five options (receptive recognition). The recall format in Van Bussel asked participants to supply an appropriate L2 target word to complete the sentence (productive recall), whereas the recall format in Yun et al. involved producing the definition of target words (receptive recall). Furthermore, while the recall posttest in Yun et al. involved producing the definition of target words (receptive recall), the recall posttest in Chen and Huang (2014) involved producing the target word form corresponding to the picture provided (productive recall). Second, previous studies also differ in the type of participants and materials. In Chen and Huang, Taiwanese sixth graders (aged 12 on average) studied English words related to animals. In Yun et al., Korean tenth graders (aged 16 on average) studied English words such as *successful, vocabulary, mistake*, and *require*. In Van Bussel, Dutch-speaking university students studied English words. These methodological differences may in part explain the mixed findings of existing research.

Non-L2 vocabulary research has also produced mixed results regarding the effects of recall and recognition on learning (for a review, see Smith and Karpicke 2014). Smith and Karpicke (2014) argue that the inconsistent results were caused possibly because the effects of recall may interact with learning phase performance. As discussed above, although recall is beneficial in that it introduces more retrieval difficulty, its limitation is that it does not necessarily ensure retrieval success. This suggests that when retrieval success during learning is low, recall formats may not be very effective because the positive effects of retrieval difficulty are likely outweighed by the negative effects of retrieval failures. When retrieval success during learning is high, in contrast, recall may be particularly effective because it enables learners to benefit from the positive effects of not only retrieval difficulty but also retrieval success (Smith and Karpicke 2014).

Although the findings of earlier L2 vocabulary studies on recall and recognition are valuable, they also suffer from at least five limitations. One limitation is that some studies do not provide detailed information about their methodology or results. For instance, the method section of existing studies does not give information such as the following: (a) whether the posttest was receptive or productive (Van Bussel 1994), (b) what the interval was between the treatment and posttest (Chen and Huang 2014; Van Bussel 1994), (c) what was given as the cue (e. g., L2 target word, L1 translation, or cloze sentence) in the posttest (Van Bussel 1994) or during the treatment (Chen and Huang 2014), (d) what kind of response (e. g., L2 target word or L1 translation) was required from participants during the treatment (Chen and Huang 2014), or (e) how many items were used as target words (Chen and Huang 2014). Furthermore, in the results section, Van Bussel (1994) does not provide the mean, *SD*, *F* value, or *p* value for the comparison of the recall and recognition conditions. Effect sizes are reported in neither Van Bussel nor Yun et al. (2008). The lack of sufficient information regarding the methodology and results makes interpretation of previous studies difficult.

Second, as discussed earlier, Smith and Karpicke (2014) observe that recall may be particularly effective when retrieval success during learning is high. A possible interaction between the recall-recognition dichotomy and retrieval success during learning suggests that it is valuable to examine the proportion of correct responses during learning when comparing the effects of recall and recognition. None of the earlier L2 vocabulary research, however, reports learning phase performance (Chen and Huang 2014; Van Bussel 1994; Yun et al. 2008). Third, the desirable difficulty framework, according to which more demanding tasks lead to better long-term retention than less demanding ones (e. g., Bjork 1994), predicts that recall is particularly effective if the posttest is

given after a long delay. A possible interaction between the recall-recognition dichotomy and posttest timing suggests that to obtain a comprehensive picture regarding the effects of recall and recognition, it is useful to give posttests at multiple intervals, preferably one after a short delay and the other after a long delay. All existing L2 vocabulary studies (Chen and Huang 2014; Van Bussel 1994; Yun et al. 2008), however, gave posttests at only one interval.

Another limitation is that with the exception of Van Bussel (1994), existing studies do not examine whether the effects of recall and recognition interact with the type of posttest (i. e., recall or recognition posttest). According to TAP theory (Morris et al. 1977), learning in a recall condition should be superior to learning in a recognition condition on a recall posttest, and learning through recognition should be superior to learning through recall on a recognition posttest. This suggests that to obtain a comprehensive picture regarding the effects of recall and recognition, it is useful to measure learning by both recall and recognition posttests. In Yun et al. (2008), however, learning was measured only by a recall posttest. Chen and Huang (2014) administered both recall and recognition posttests. Yet, because they only present and analyze the combined scores of the two posttests, it is not clear whether recall was effective only on the recall posttest or on both recall and recognition posttests.

Lastly, previous studies on receptive and productive learning (e. g., Griffin and Harley 1996; Mondria and Wiersma 2004; Nation 2013; Webb 2009b) suggest that to gain both receptive and productive vocabulary knowledge efficiently, learners need to practice receptive as well as productive retrieval (see Effects of receptive and productive retrieval). However, existing studies on recall and recognition used only receptive or productive retrieval (Van Bussel 1994; Yun et al. 2008) or do not specify whether the treatment was receptive, productive, or a combination of both (Chen and Huang 2014). In order to maximize paired-associate learning, it would be useful to use both receptive and productive retrieval. As the above review of literature shows, existing studies on recall and recognition suffer from a number of limitations. The present study investigated the effects of recall and recognition formats on paired-associate learning of L2 vocabulary while addressing the limitations of earlier research.

1.3 Effects of a combination of recall and recognition

As discussed in the previous section, both recall and recognition have their advantage and disadvantage. Some researchers (e. g., Park 2005; Smith and Karpicke 2014) argue that using both recall and recognition formats

(hereafter referred to as the *hybrid format*; Smith and Karpicke 2014) may offer a solution to this problem. Specifically, by practicing retrieval in a recognition format initially and in a recall format later, learners may be able to benefit from the positive effects of both retrieval success afforded by recognition and retrieval difficulty afforded by recall. Although none of the existing studies examined the effects of the hybrid format on L2 vocabulary learning (Chen and Huang 2014; Van Bussel 1994; Yun et al. 2008), Clariana and Lee (2001) compared the effects of recognition (i. e., choose the most appropriate word corresponding to the definition), recall (i. e., type the appropriate word corresponding to the definition), and hybrid formats (i. e., a recognition question followed by a recall question) on the learning of L1 technical vocabulary. In their study, 133 American graduate students studied 35 technical words in the field of instructional design. Clariana and Lee found that the hybrid format increased the learning of L1 technical vocabulary more than recognition alone.

Although Clariana and Lee's (2001) findings are significant, one limitation of their research is that the recognition and hybrid formats were not controlled for the frequency of retrievals. More specifically, while there were two retrieval attempts per target word in the hybrid condition (one recognition + one recall), target items were practiced only once in the recognition condition. The superiority of the hybrid format, therefore, may be partly due to the difference in the retrieval frequency rather than the retrieval format. Furthermore, as Clariana and Lee looked into the learning of L1 technical vocabulary, their findings may not necessarily be applicable to L2 vocabulary learning. In order to examine the value of the hybrid format for paired-associate learning of L2 vocabulary, the effects of the hybrid format were also investigated in this study.

1.4 Effects of using only productive recall

It would also be useful to examine the effects of using only a productive recall format (hereafter referred to as the *productive recall only treatment*) for paired-associate learning of L2 vocabulary. There exist conflicting views about the effectiveness of the productive recall only treatment. First, as productive recall tends to produce a low rate of retrieval success (e. g., Laufer and Goldstein 2004), the productive recall only treatment may not be effective based on the retrieval practice effect (Baddeley 1997; Ellis 1995). Second, TAP theory (Morris et al. 1977) and the forward asymmetry effect (Kahana and Caplan 2002) predict that the productive recall only treatment enhances the acquisition of productive, but not receptive, vocabulary knowledge. Existing studies on receptive and

productive learning (e. g., Griffin and Harley 1996; Mondria and Wiersma 2004; Steinel et al. 2007; Webb 2009b) support this prediction.

In contrast, the desirable difficulty framework (e. g., Bjork 1994) and the retrieval effort hypothesis (Pyc and Rawson 2009) suggest that the productive recall only treatment is effective. Because productive recall is more demanding than receptive recognition, productive recognition, and receptive recall (e. g., Laufer and Goldstein 2004), the productive recall only treatment may maximize paired-associate learning according to the desirable difficulty framework and the retrieval effort hypothesis. The desirable difficulty framework also partially contradicts the retrieval practice effect, TAP theory, and forward asymmetry effect. First, unlike the retrieval practice effect, the desirable difficulty framework states that retrieval success during learning is not necessarily a reliable index of long-term retention. The productive recall only treatment, therefore, may turn out to be effective despite a low level of retrieval success during learning (e. g., Bjork 1994; Schneider et al. 2002).

Second, unlike TAP theory and the forward asymmetry effect, the desirable difficulty framework suggests that the productive recall only treatment enhances the acquisition of not only productive but also receptive knowledge. Because productive recall is the most demanding retrieval format (e. g., Laufer and Goldstein 2004), the desirable difficulty framework predicts that the productive recall only treatment results in a kind of knowledge that is transferable to novel environments (e. g., Bjork 1994; Schneider et al. 2002). As a result, the productive recall only treatment may turn out to be effective even on receptive posttests, where the testing condition does not match that of learning.

2 The present study

The first purpose of the present study was to examine the effects of recall and recognition on paired-associate learning of L2 vocabulary. This study differs from existing studies in four important respects. First, although the effects of recall may interact with the level of retrieval success during learning (Smith and Karpicke 2014), none of the earlier L2 vocabulary studies on recall and recognition reports the proportion of correct retrievals during the learning phase (Chen and Huang 2014; Van Bussel 1994; Yun et al. 2008). With this limitation in mind, the present study examined not only posttest but also learning phase performance. Second, all existing L2 vocabulary studies on recall and recognition (Chen and Huang 2014; Van Bussel 1994; Yun et al.

2008) administered a posttest at only one interval. This is problematic because the desirable difficulty framework (e. g., Bjork 1994) predicts that the effects of recall and recognition are conditional upon the timing of posttest. To test a possible interaction between the recall-recognition dichotomy and posttest timing, the present study administered posttests at two intervals: immediately and 1 week after the treatment. The interval of 1 week was chosen for the delayed posttest because research has indicated that most forgetting occurs immediately after learning (e. g., Cepeda et al. 2008). Scores on a 1-week delayed posttest may thus provide an accurate assessment of retention over time. Third, with the exception of Van Bussel (1994), earlier studies failed to examine whether the effects of recall and recognition are conditioned by the type of posttest (recall or recognition). With this limitation in mind, the present study administered both recall and recognition posttests. Lastly, because previous studies on receptive and productive learning suggest that it is desirable to practice both receptive and productive retrieval (e. g., Griffin and Harley 1996; Mondria and Wiersma 2004; Nation 2013; Webb 2009b), the recall and recognition formats in this study involved productive as well as receptive retrieval.

Another goal of this study was to test whether a combination of recall and recognition (hybrid format) increases paired-associate learning of L2 vocabulary more than either one alone. Although the hybrid format is considered to maximize learning (e. g., Park 2005; Smith and Karpicke 2014), none of the earlier studies examined the effects of the hybrid format on L2 vocabulary learning. Lastly, because there exist conflicting views about the effectiveness of the productive recall only treatment (i. e., retrieval practice effect, TAP theory, and forward asymmetry effect vs. desirable difficulty framework and retrieval effort hypothesis), the effectiveness of the productive recall only treatment was also investigated. By comparing the effects of recognition, recall, hybrid, and productive recall only formats, the present study may help us to determine which kinds of retrieval formats should be used to optimize paired-associate learning of L2 vocabulary.

The following three research questions were addressed in this study:

1. Which is more effective for paired-associate learning of L2 vocabulary, recall or recognition?
2. Is a hybrid format more effective than recall or recognition alone for paired-associate learning of L2 vocabulary?
3. Is the productive recall only treatment effective for paired-associate learning of L2 vocabulary?

3 Method

3.1 Participants

The participants were 64 English-speaking students at a university in New Zealand. Their average age was 20.77 ($SD = 3.91$). None of the participants had prior knowledge of Swahili, the target language in this study. The participants were randomly divided into four groups of 16 participants. Sixty target word pairs were also divided into four sets of 15 pairs (hereafter, Sets A, B, C, and D; see Target items for details). In order to counterbalance the effects of target items, the four groups of participants studied different sets of items under different conditions (Figure 1). Participants received a $20 NZD shopping voucher in exchange for their participation.

	Learning conditions (within-participant variable)			
	Item set A	Item set B	Item set C	Item set D
Group 1	Recognition	Recall	Hybrid	Productive recall only
Group 2	Productive recall only	Recognition	Recall	Hybrid
Group 3	Hybrid	Productive recall only	Recognition	Recall
Group 4	Recall	Hybrid	Productive recall only	Recognition

Figure 1: Design of the present study. Sixty-four participants were randomly divided into four groups of 16 participants (Groups 1, 2, 3, and 4). Sixty target word pairs were also divided into four sets of 15 word pairs (Sets A, B, C, and D). The four groups of participants studied different sets of word pairs under different conditions, thus counterbalancing the effects of target items. For instance, Group 1 studied Set A under the recognition, Set B under the recall, Set C under the hybrid, and Set D under the productive recall only conditions. Group 2 studied Set B under the recognition, Set C under the recall, Set D under the hybrid, and Set A under the productive recall only conditions, and so forth.

3.2 Experimental design

The independent variable was the type of learning condition: recognition, recall, hybrid, and productive recall only. The learning condition was a within-participant variable. The dependent variables were effectiveness and efficiency of the four learning conditions. The present study operationally defined

effectiveness as the number of correct responses on the posttest. Efficiency was operationalized as the number of word pairs learned per minute (posttest score divided by the study time; e. g., Nakata in press; Kornell 2009; Mondria 2003).

3.3 Dependent measures

Four types of posttests were given in this order: productive recall, productive recognition, receptive recall, and receptive recognition. A productive recall test was followed by a productive recognition test because correct responses in the former test were given as multiple-choice options in the latter, and administering the recognition test prior to the former may affect performance on the recall test. A receptive recall test was followed by a receptive recognition test for the same reason. The two productive tests were given prior to the two receptive tests because administering the receptive tests earlier may have a larger effect on test scores than vice versa. More specifically, in the receptive test, Swahili (L2) words such as *chakura, malkia*, and *hadithi* were provided as cues. Since the participants have never met these words prior to the treatment, they might acquire knowledge of orthography of these words by seeing these words used as a cue in the receptive test. In contrast, in the productive test, English (L1) words such as *food, queen*, and *story* were given as cues. Since these words were already familiar to participants prior to the experiment, they are not likely to acquire new vocabulary knowledge through the productive test. As there seems to be a larger learning effect from the receptive test than the productive test, the productive test was administered prior to the receptive test.

All 60 target word pairs were tested, and each posttest consisted of 60 items. Unlike the treatment, feedback was not provided in the posttest. Other than that, the posttests were exactly the same as the corresponding retrieval formats in the treatment (Figure 3). In the recognition posttest, the distractors were chosen randomly from the correct responses for other target items and fixed for all participants. An item that was used as a distractor for a given item during the treatment was not chosen as a distractor for the same item in the recognition posttest. To reduce effects of the productive posttests on the receptive tests, the productive recognition posttest was followed by a 3-minute distractor task (two-digit additions). The immediate posttest was given on the same day as the treatment. The delayed posttest was conducted 1 week after the treatment. A pretest was not given because none of the participants had prior knowledge of the target language (Swahili).

3.4 Target items

Sixty Swahili-English word pairs (e. g., *chakura-food*) selected from Nelson and Dunlosky (1994) were used as target items. The 60 word pairs were divided into Sets A to D so that the learning difficulty would be distributed as evenly as possible. More specifically, the four sets of items were matched for the following four variables: (a) Nelson and Dunlosky's difficulty norms, (b) L2 word length (number of letters and syllables), (c) pronounceability of L2 words (wordlikeness ratings reported by Nelson and Dunlosky 1994), and (d) orthographic similarity of L2 words to the L1 lexicon (average positional bigram frequency and orthographic neighborhood size). Although the four sets may not be completely comparable in their difficulty, a potential difference, if any, would not have a major effect on the results of this experiment because effects of target word pairs would be counterbalanced across four groups of participants (Figure 1).

3.5 Procedure

The study was conducted with computer software developed by the author. After receiving explanations about the study, participants studied 60 Swahili-English word pairs in a paired-associate format using a computer-based flashcard program. During the treatment, there were five cycles of 60 items, and each item was encountered only once in each cycle. The items from four sets (Sets A to D) occurred once every four items (e. g., ABCD ABCD ABCD...). The order of items was randomized anew for each cycle. In the first cycle, the target Swahili-English word pairs were presented for 7 seconds per word pair. In the rest of the cycles, participants practiced retrieval. Target items were studied in a different retrieval format depending on the condition to which they were assigned (Figure 2). In the recognition condition, target items were practiced in a receptive recognition format twice and then a productive recognition format twice. The recall condition consisted of two receptive recall questions followed by two productive recall questions. In the hybrid condition, target items were studied once in each of the receptive recognition, productive recognition, receptive recall, and productive recall formats in that order. In the productive recall only condition, target items were practiced four times in productive recall. In the first three conditions, questions were arranged in order of increasing difficulty. This is based on the view that gradually increasing retrieval difficulty maximizes learning (e. g., Logan and Balota 2008). The order of the formats is based on Laufer and Goldstein (2004).

Conditions	Retrieval formats during the treatment
Recognition	2 receptive recognition + 2 productive recognition questions
Recall	2 receptive recall + 2 productive recall questions
Hybrid	1 receptive recognition + 1 productive recognition + 1 receptive recall + 1 productive recall questions
Productive recall only	4 productive recall questions

Figure 2: Retrieval formats in the four conditions. Target items were studied in a different retrieval format during the treatment depending on the condition to which they were assigned. In the recognition condition, target items were practiced in a receptive recognition format twice and then a productive recognition format twice. In the recall condition, target items were practiced in a receptive recall format twice and then a productive recall format twice. In the hybrid condition, target items were studied once in each of the receptive recognition, productive recognition, receptive recall, and productive recall formats in that order. In the productive recall only condition, target items were practiced four times in productive recall.

As shown in Figure 3, four kinds of retrieval formats were used: receptive recognition, productive recognition, receptive recall, and productive recall. In receptive recognition, participants were presented with a Swahili word and asked to select the most appropriate English translation from four options (Figure 3(a)). The distractors were chosen randomly from the correct responses for other target items and fixed for all participants. If participants were not sure about the correct answer, they were instructed to choose the *I DON'T KNOW* option. In productive recognition, participants were presented with an English word and asked to pick the most appropriate Swahili translation from among four alternatives (Figure 3(b)). In receptive recall, learners were asked to type the meaning of Swahili words (Figure 3(c)). In productive recall, participants produced the Swahili word form corresponding to the English translation provided (Figure 3(d)). Participants were given as much time as necessary to respond. After each response, the target Swahili-English word pair was provided as feedback.[1] After completing the treatment, participants answered two-digit additions (e. g., 39 + 34 = ?) as a distractor task for 5 minutes. The immediate posttest was conducted immediately after the distractor task (see below). One week after the treatment, the delayed posttest was conducted without prior notice.

1 For half of the participants, the feedback duration was fixed to 5 seconds per response. The other half of the participants were allowed to close the feedback window before 5 seconds elapsed. The pacing of feedback was manipulated because this study was part of a larger study that investigated feedback pacing as a variable. Since no significant interaction existed between the condition and feedback pacing (Nakata 2013), the use of the two types of feedback is assumed to have had little effect on the results of the current study.

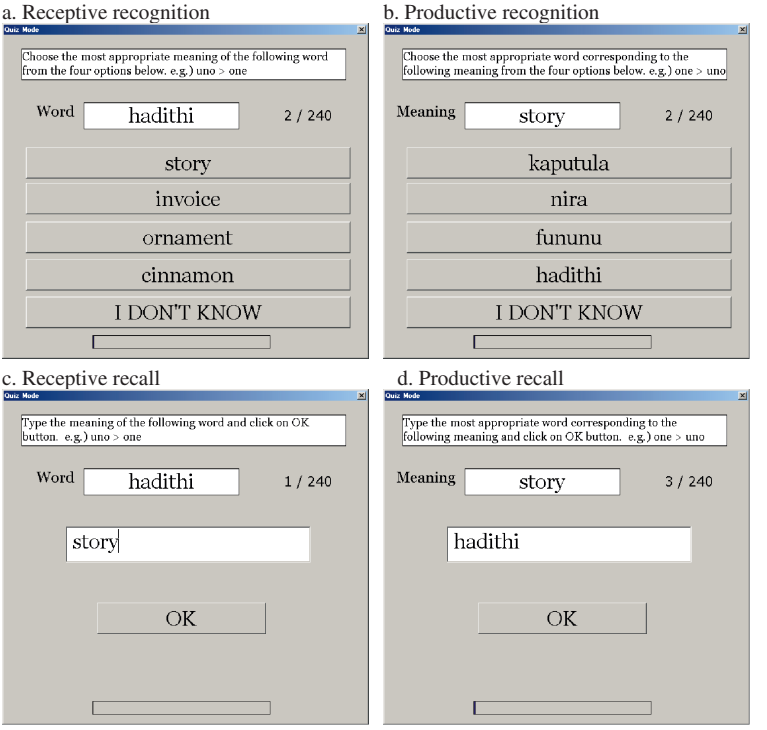

Figure 3: Examples of the four retrieval formats. In receptive recognition (a), participants were presented with a Swahili word and asked to select the most appropriate English translation from four options. In productive recognition (b), participants were presented with an English word and asked to select the most appropriate Swahili translation from four options. In receptive recall (c), participants were asked to type the meaning of Swahili words. In productive recall (d), participants were asked to type the Swahili word form corresponding to the English translation provided.

3.6 Scoring

Responses on the recognition posttests were scored as either correct or incorrect. On the productive recall posttest, responses were scored using strict and sensitive protocols. Using the strict scoring protocol, only correctly spelled responses were marked as correct. Using the sensitive scoring protocol, responses that would be awarded 0.75 using a lexical production scoring protocol-written (e. g., Barcroft 2007) were also treated as correct (e. g., *chumbo*, *chembo*, and *chumba* for *chimbo*). On the receptive recall posttest, responses with spelling mistakes were marked as correct (e. g., *cinammon* for *cinnamon*). Plural forms of the target

word were also marked as correct (e. g., *ornaments* for *ornament*). Responses were scored as incorrect if they were of a different part of speech as the translation given during the treatment (e. g., *scientific* for *science*).

4 Results

4.1 Learning phase data

Since the effects of recall and recognition may be affected by the level of retrieval success during learning (Smith and Karpicke 2014; see Review of literature), learning phase performance was analyzed. When collapsed across the four retrieval attempts, the average number of correct responses during the treatment (*SD*s in parentheses) was 12.16 (2.19), 4.70 (2.90), 8.61 (2.75), and 4.06 (2.87) out of 15 in the recognition, recall, hybrid, and productive recall only conditions, respectively. A one-way ANOVA found a statistically significant difference among the four conditions, F (2.20, 138.57) = 555.72, $p < .001$, $\eta^2 = .90$. (The *df*s contain decimal values due to the Greenhouse-Geiser correction.) The Bonferroni method of multiple comparisons showed that all four conditions were significantly different from each other ($p \leq .005$), producing medium to large effect sizes ($.39 \leq r \leq .97$).

4.2 Posttest performance

Table 1 provides the immediate and delayed posttest results for the four conditions. K-R21 was .93 or higher (.93-.96) for all dependent measures, demonstrating good reliability. The productive and receptive recall posttest scores were analyzed by a two-way 4 (learning conditions: recognition/recall/hybrid/productive recall only) x 2 (posttest timing: immediate/1-week delayed) ANOVA. The main effect of condition was significant on the productive recall posttest, strict: F (3, 189) = 18.19, $p < .001$, $\eta_p^2 = .22$; sensitive: F (3, 189) = 2.75, $p = .044$, $\eta_p^2 = .04$, but was not significant on the receptive recall posttest, F (2.52, 159.05) = 0.56, $p = .615$, $\eta_p^2 = .01$. The interaction between the condition and posttest timing was not significant on either the productive or receptive recall posttest, productive strict: F (2.64, 166.30) = 1.86, $p = .145$, $\eta_p^2 = .03$; productive sensitive: F (2.64, 166.14) = 2.15, $p = .105$, $\eta_p^2 = .03$, receptive: F (3, 189) = 0.64, $p = .593$, $\eta_p^2 = .01$.

As the main effect of condition proved significant on the productive recall posttest, contrasts were performed to examine where the significant differences existed when collapsed across the immediate and delayed posttests. With strict

Table 1: Average number of correct responses on the posttests (Standard deviations in italics).

Posttests	Posttest timing							
	Immediate posttest				Delayed posttest			
	Recognition	Recall	Hybrid	Productive recall only	Recognition	Recall	Hybrid	Productive recall only
Productive recall (strict)	6.73	8.33	7.72	8.45	2.94	3.88	3.25	4.25
	3.42	*4.16*	*3.85*	*4.16*	*2.54*	*3.33*	*3.11*	*3.30*
Productive recall (sensitive)	10.73	11.00	10.69	10.75	5.53	6.39	5.80	6.36
	3.85	*3.92*	*3.94*	*4.01*	*3.56*	*4.02*	*3.82*	*4.09*
Productive recognition	13.75	13.92	14.00	13.59	12.88	12.89	13.02	13.00
	2.22	*1.85*	*1.89*	*2.68*	*2.82*	*2.81*	*2.62*	*2.91*
Receptive recall	11.05	11.25	11.38	11.41	10.70	10.95	10.81	10.88
	3.23	*3.67*	*3.32*	*3.71*	*3.14*	*3.75*	*3.40*	*3.69*
Receptive recognition	13.84	13.77	13.77	13.75	13.42	13.50	13.58	13.47
	2.18	*2.50*	*2.22*	*2.61*	*2.58*	*2.66*	*2.33*	*2.76*

Note: $n = 64$. The maximum score is 15 for each cell.

scoring, the recall condition was significantly more effective than the recognition ($p < .001$, $r = .52$) and hybrid conditions ($p = .008$, $r = .32$), and medium to large effect sizes were observed. The productive recall only condition fared significantly better than the recognition ($p < .001$, $r = .64$) and hybrid conditions ($p < .001$, $r = .48$), producing medium to large effect sizes. The difference between the recall and productive recall only conditions was rather small as indicated by the lack of statistical significance ($p = .251$) as well as the small effect size ($r = .14$). The contrasts also showed that the hybrid condition was significantly more effective than the recognition condition ($p = .003$), producing a medium sized effect ($r = .37$). Taken together, the findings suggest the following order with strict scoring on the productive recall posttest: recall = productive recall only > hybrid > recognition.

When the sensitive scoring procedure was used, the difference among the four conditions was smaller. With sensitive scoring, the recall condition fared significantly better than the recognition condition ($p = .025$) when collapsed across the immediate and delayed posttests. However, only a small effect size was observed ($r = .28$). The productive recall only condition was no more effective than the recognition ($p = .091$, $r = .21$) or hybrid condition ($p = .151$, $r = .18$). The difference between the recall and hybrid conditions fell short of statistical significance ($p = .053$, $r = .24$). No significant difference existed between the recall and productive recall only conditions ($p = .489$, $r = .09$) or between the recognition and hybrid conditions either ($p = .594$, $r = .07$), producing very small effect sizes. These results indicate that although the recall and productive recall

only conditions resulted in higher scores than the recognition and hybrid conditions with sensitive scoring, the advantage was smaller compared with when the strict method was used.

Next, the recognition posttest scores were analyzed using a non-parametric Friedman test. ANOVA was not used because the coefficients of skewness or kurtosis for the recognition posttest scores were greater than 2 or smaller than −2, indicating that the distributions of these scores were significantly different from the normal distribution. The Friedman tests detected no statistically significant difference among the four conditions regardless of the timing or type of posttest: immediate productive: $\chi^2 = 4.49$, $p = .213$; immediate receptive: $\chi^2 = 1.10$, $p = .777$; delayed productive: $\chi^2 = 0.82$, $p = .844$; delayed receptive: $\chi^2 = 1.31$, $p = .727$. The results suggest that there was little difference among the four conditions in their recognition posttest scores.[2]

Next, in order to investigate whether the four learning conditions affected durability of learning, the attrition scores in the four conditions were compared. The attrition scores were calculated by subtracting the delayed posttest score from the immediate posttest score. No significant difference was found among the four conditions in the attrition scores regardless of the type of posttest: $F(3, 189) \leq 2.15$, $p \geq .105$, $\eta_p^2 \leq .03$. The results suggest that the learning condition did not have a major effect on the durability of vocabulary knowledge.[3]

---

2 One anonymous reviewer pointed out that the recognition posttest scores might have been affected by guessing. In order to correct for guessing, the following formula was used: corrected score = number of correct responses − 1/3 * number of incorrect responses (e. g., Harris 1969; Prihoda et al. 2006). The corrected and uncorrected scores correlated significantly ($r = .99$, $p < .001$) on all posttests. Friedman tests detected no statistically significant difference among the four conditions regardless of the timing or type of posttest, $\chi^2 \leq 4.38$, $p \geq .226$, mirroring the results calculated without correction for guessing. The analysis suggests that guessing perhaps did not have a major effect on the results of this study.

3 One anonymous reviewer suggested examining possible effects of the learning condition, learning phase performance (retrieval success during the treatment), and time on task on posttest performance. Based on the reviewer's suggestion, a binomial logit mixed-effects model (Jaeger 2008) was used to explore a possible relationship between these variables. The analysis showed that the main effect of learning phase performance was significant on all dependent variables ($p < .001$ for all). This suggests that when collapsed across learning conditions, better learning phase performance was associated with better posttest performance. The main effect of time on task was not significant on the delayed productive recall posttest (strict scoring: $p = .632$, sensitive scoring: $p = .147$), but was significant on all other posttests ($p \leq .019$). The results suggest that when collapsed across learning conditions, longer time on task was associated with better posttest performance on all posttests except on the delayed productive

4.3 Efficiency

On average, the participants spent 9.14 (2.20), 11.56 (3.20), 10.49 (2.38), and 11.18 (2.81) minutes (*SD*s in parentheses) studying the target word pairs under the recognition, recall, hybrid, and productive recall only conditions, respectively. As there was a statistically significant difference in study time among the four conditions, F (1.68, 105.87) = 58.80, $p < .001$, $\eta^2 = .48$, the efficiency of the four

recall posttest. The main effect of time on task was not significant on the delayed productive recall posttest possibly because it was the most difficult type of posttest (e. g., Laufer and Goldstein 2004; also see Table 1), and just spending longer time studying did not necessarily guarantee successful performance on this test.

Next, in order to examine whether learning phase performance or time on task had a larger effect on posttest performance, two further analyses were carried out using a binomial logit mixed-effects model. First, using time on task as a covariate, an interaction between the learning condition and learning phase performance was tested. Second, using learning phase performance as a covariate, an interaction between the learning condition and time on task was examined. The results of the analyses can be summarized as follows: First, the interaction between the learning condition and learning phase performance was not statistically significant on the productive recall posttest in the recognition condition (immediate: $p = .128$; delayed: $p = .567$). In all other cases, the interaction between the learning condition and learning phase performance was statistically significant ($p \leq .009$). The results suggest that successful learning phase performance was associated with successful posttest performance on all posttests except on the productive recall posttest in the recognition condition. One possible explanation is that unlike the other three conditions, the recognition condition did not involve any productive recall format. TAP theory (Morris et al. 1977) predicts that the best way to learn the spelling of L2 words would be to practice retrieval of the word forms. It is probably as a result of this that on the productive recall posttest, the recognition condition, which did not involve productive recall, performed poorly regardless of the level of retrieval success during learning.

Second, the analyses also suggested that learning phase performance was a better predictor of the posttest performance than time on task. This is because although the interaction between the learning condition and learning phase performance was statistically significant in 38 out of 40 cases (4 conditions x 5 dependent measures x 2 posttest timings), the interaction between the learning condition and time on task was statistically significant only in 24 out of 40 cases. The results suggest that overall, learning phase performance was a better predictor of the posttest performance than time on task. In other words, just spending longer time studying did not necessarily lead to better posttest performance. One caveat, though, is that the results might have been partly caused by item difficulty effects. In other words, items that were correctly answered during the learning phase were probably easier than others, and hence, more likely to be answered correctly on the posttest (Karpicke and Roediger 2007; Pyc and Rawson 2009). As a result, it should be noted that the significant interaction between learning phase and posttest performance does not necessarily mean that successful learning phase performance caused successful posttest performance.

conditions was compared. The present study operationally defined efficiency as the number of word pairs learned per minute, and efficiency scores were calculated by dividing the posttest score by the study time (e. g., Nakata in press; Kornell 2009; Mondria 2003). For instance, suppose that a given participant spent 10 minutes studying the target word pairs under a given condition and scored 10 and 5 on the immediate and delayed posttests, respectively. The efficiency score for this participant will be 1.00 (10/10 = 1.00) for the immediate posttest and 0.50 (5/10 = 0.50) for the delayed posttest. Table 2 provides the efficiency scores in the four conditions. To test whether any significant difference existed among the four conditions, the efficiency scores were entered into a two-way 4 (learning conditions: recognition/recall/hybrid/productive recall only) x 2 (posttest timing: immediate/1-week delayed) ANOVA. Table 3 shows the results of the ANOVAs.

Table 2: Average efficiency scores in the four conditions (Standard deviations in italics).

Posttests								**Posttest timing**
	Immediate posttest				**Delayed posttest**			
	Recognition	**Recall**	**Hybrid**	**Productive recall only**	**Recognition**	**Recall**	**Hybrid**	**Productive recall only**
Productive recall (strict)	0.79	0.74	0.77	0.76	0.34	0.34	0.32	0.38
	0.47	*0.38*	*0.42*	*0.35*	*0.32*	*0.31*	*0.32*	*0.30*
Productive recall (sensitive)	1.24	0.98	1.05	0.96	0.64	0.55	0.57	0.57
	0.56	*0.37*	*0.46*	*0.34*	*0.46*	*0.36*	*0.40*	*0.37*
Productive recognition	1.59	1.29	1.40	1.27	1.49	1.17	1.29	1.21
	0.48	*0.37*	*0.36*	*0.36*	*0.50*	*0.34*	*0.37*	*0.37*
Receptive recall	1.29	1.00	1.13	1.03	1.25	0.97	1.07	0.99
	0.54	*0.35*	*0.42*	*0.35*	*0.52*	*0.35*	*0.42*	*0.37*
Receptive recognition	1.61	1.26	1.37	1.29	1.55	1.23	1.35	1.25
	0.49	*0.39*	*0.37*	*0.38*	*0.49*	*0.36*	*0.35*	*0.37*

Note: $n = 64$.

As Table 3 shows, the main effect of condition was significant on all posttests except with strict scoring on the productive recall posttest. The interaction between the condition and posttest timing was significant only with sensitive scoring on the productive recall posttest. Due to the significant main effect of condition, contrasts were performed to examine where the significant differences existed when collapsed across the immediate and delayed posttests. Results of the contrasts are reported in Table 4. The table presents the F values, p values, and effect sizes r for the pair-wise contrasts. Contrasts were not

Table 3: Results of two-way ANOVAs for the efficiency scores.

Posttests	Condition				Condition X Posttest timing			
	df	*F*	*p*	η_p^2	*df*	*F*	*p*	η_p^2
Productive recall (strict)	2.37, 149.57	.79	.476	.01	2.52, 158.62	2.21	.100	.03
Productive recall (sensitive)	2.12, 133.74	18.92	.000	.23	2.53, 159.19	13.15	.000	.17
Productive recognition	2.14, 134.86	37.76	.000	.37	2.60, 163.51	2.25	.093	.03
Receptive recall	2.32, 146.16	28.52	.000	.31	3, 189	.35	.788	.01
Receptive recognition	1.84, 116.14	46.79	.000	.43	2.35, 148.07	.70	.520	.01

Note: Because Mauchly's test showed that sphericity assumptions were not met for all effects except the interaction between the condition and posttest timing on the receptive recall posttest, the Greenhouse-Geiser correction was used. As a result, the *df*s for all effects except the interaction between the condition and posttest timing on the receptive recall posttest contain decimal values.

performed for the productive recall test with strict scoring because the main effect of condition was not significant (Table 3).

First, the contrasts (Table 4) showed that with sensitive scoring on the productive recall test, the recognition condition was significantly more efficient than the other three conditions, producing large effect sizes ($.54 \leq r \leq .58$). No statistically significant difference was detected among the recall, hybrid, and productive recall only conditions. The results are supported by the effect sizes ($.02 \leq r \leq .24$), which are regarded as having no more than small effects (Cohen 1988). These results suggest the following order with sensitive scoring on the productive recall posttest: recognition > recall = hybrid = productive recall only. Next, on the productive recognition, receptive recall, and receptive recognition posttests, the contrasts (Table 4) revealed that (a) the recognition condition was significantly more efficient than the other three conditions, (b) the hybrid condition was significantly more efficient than the recall and productive recall only conditions, (c) no statistically significant difference existed between the recall and productive recall only conditions, and (d) medium to large effect sizes ($.32 \leq r \leq .75$) were found for all comparisons except between the recall and productive recall only conditions ($.05 \leq r \leq .16$). These results suggest the following order on the productive recognition, receptive recall, and receptive recognition posttests: recognition > hybrid > recall = productive recall only. Overall, the results show that the recognition condition was the most efficient, followed by the hybrid condition.

Table 4: Results of pair-wise contrasts for efficiency scores.

Posttests	Conditions	Recognition			Recall			Hybrid		
		F	*p*	*r*	*F*	*p*	*r*	*F*	*p*	*r*
Productive recall (sensitive)	Recall	25.99	.000	.54						
	Hybrid	26.47	.000	.55	2.99	.089	.21			
	Productive recall only	31.09	.000	.58	.01	.905	.02	3.83	.055	.24
Productive recognition	Recall	53.12	.000	.68						
	Hybrid	42.01	.000	.64	23.04	.000	.52			
	Productive recall only	59.91	.000	.70	.13	.721	.05	12.28	.001	.41
Receptive recall	Recall	50.89	.000	.67						
	Hybrid	25.76	.000	.54	16.69	.000	.46			
	Productive recall only	48.85	.000	.66	1.55	.217	.16	7.19	.009	.32
Receptive recognition	Recall	73.27	.000	.74						
	Hybrid	77.98	.000	.75	22.54	.000	.52			
	Productive recall only	58.36	.000	.70	1.28	.262	.14	8.57	.005	.35

Note: $df = (1, 62)$. The above table should be read as follows: With sensitive scoring on the productive recall posttest, the difference between the recognition and recall conditions was statistically significant, $F(1, 62) = 25.99$, $p < .001$, and a large effect size was observed ($r = .54$). Effect sizes (r) of .10, .30, and .50 indicate small, medium, and large effects, respectively (Cohen 1988).

5 Discussion

The first research question in this study asked whether recall increases paired-associate learning of L2 vocabulary more than recognition. The current study demonstrated only a limited advantage of recall. First, let us consider the scores on the receptive recall posttest. On this test, no statistically significant difference existed between the recall and recognition conditions. The result is at odds with Yun et al. (2008), who found the benefits of recall over recognition on a receptive recall posttest. There may be two explanations for the incongruent results. First, the contradictory results may be ascribed in part to a difference in learning phase performance. As discussed in Review of literature, recall formats may be particularly effective when retrieval success during learning is high (Smith and Karpicke 2014). In this study, the recall condition produced a significantly lower retrieval success rate during learning

($M = 4.70$) than the recognition condition ($M = 12.16$). As a result, the positive effects of retrieval difficulty were possibly outweighed by the negative effects of retrieval failures. This may partially account for the lack of significant difference between the recall and recognition conditions on the receptive recall posttest.

Alternatively, the results could be partially due to the test order. While the productive recall posttest was given before the receptive recall posttest in this study, no posttest preceded the receptive recall posttest in Yun et al. (2008). Since correct responses in the receptive recall test were used as cues in the productive recall test, the productive recall test might have affected performance on the receptive recall test. This may have possibly diminished a potential difference between the recall and recognition conditions on the receptive recall posttest in this study.

Next, the recall condition resulted in significantly higher scores than the recognition condition on the productive recall posttest. Why was the advantage of recall found not on the receptive but on the productive recall posttest? One possible explanation is that the productive recall posttest required precise knowledge of orthography unlike the receptive recall posttest. TAP theory (Morris et al. 1977) predicts that the best way to learn the spelling of L2 words would be to practice retrieval of the word forms. It is probably as a result of this that on the productive recall posttest, the recall condition, which involved productive recall, fared better than recognition. As noted above, due to the rather low retrieval success rate during learning, the advantage of recall over recognition might have been relatively small in this study. However, because of TAP, the superiority of recall on the productive recall posttest perhaps did not diminish to the extent that it would completely disappear. This could be in part the reason why recall was significantly more effective than recognition on the productive recall posttest, but not on the receptive. On the productive or receptive recognition posttests, no significant difference was detected between the recognition and recall conditions. The findings are consistent with Van Bussel (1994) and suggest that recall does not necessarily increase paired-associate learning of L2 vocabulary more than recognition when learning is measured by a recognition posttest.

The second research question in this study asked whether a combination of recall and recognition (hybrid format) increases paired-associate learning of L2 vocabulary more than either one alone. Some researchers argue that the hybrid format maximizes learning because it combines the benefits of both recall and recognition (e. g., Park 2005; Smith and Karpicke 2014). The present study, yet, demonstrated only limited benefits of the hybrid format. Although the hybrid condition was significantly more effective than the recognition condition with

strict scoring on the productive recall posttest, there was no difference between it and the recognition or recall conditions on the other posttests. In addition, the hybrid condition was less effective than the recall condition on the productive recall posttest. These results are inconsistent with Clariana and Lee (2001), who found that a hybrid format fared better than recognition on the retention of L1 vocabulary.

There are two explanations for the contradictory results between the present study and Clariana and Lee (2001). First, in Clariana and Lee, the recognition and hybrid formats were not controlled for the retrieval frequency: While target items were practiced twice in the hybrid condition, there was only one retrieval attempt in the recognition condition. In this study, in contrast, all four learning conditions had the same number of retrieval attempts (four). This may partially be the reason why the hybrid format in Clariana and Lee fared better than in this study. Second, the contradictory results could be ascribed in part to a difference in learning phase performance. The hybrid format in this study produced a lower retrieval success rate during learning ($M = 57.4\,\%$; see Results) than in Clariana and Lee ($M = 95\,\%$). As a result, the participants might not have been able to benefit from the positive effects of retrieval success under the hybrid condition in this study. This may also be responsible for the incongruent results between Clariana and Lee and this study.

In the current study, the hybrid condition was less effective than the recall condition on the productive recall posttest. This is probably because the recall condition involved more productive recall questions per target word (two) than the hybrid condition (one). As productive recall is rather demanding (e. g., Laufer and Goldstein 2004), practicing L2 words in a productive recall format once perhaps did not guarantee successful performance on the productive recall posttest. Consequently, the hybrid format might have proved less effective than recall on the productive recall posttest.

The third research question in this study asked whether the productive recall only treatment is effective. The present study found the advantage of the productive recall only condition over the recognition and hybrid conditions with strict scoring on the productive recall posttest. The results can be explained by TAP theory (Morris et al. 1977). The productive recall only condition, however, was no more effective than any of the other conditions on the other dependent variables. The limited advantage of the productive recall only treatment was caused possibly because the positive effects of retrieval difficulty (e. g., Bjork 1994; Pyc and Rawson 2009) were outweighed by the negative effects of retrieval failures (Baddeley 1997; Ellis 1995). The productive recall only condition produced a smaller number of correct retrievals during learning ($M = 4.06$) than the other three conditions (recognition: 12.16, recall:

4.70, hybrid: 8.61; see Results). Because learners were not able to benefit from the positive effects of retrieval success, the advantage of the productive recall only condition might have been limited.

At the same time, it is noteworthy that the productive recall only condition turned out to be as effective as the other three conditions even on the receptive posttests. Because the productive recall only treatment consists only of productive retrieval, TAP theory (Morris et al. 1977) predicts that on the receptive tests, it may be less effective than the other three treatments, which involve both receptive and productive retrieval. The forward asymmetry effect also predicts only limited positive effects of the productive recall only treatment on the receptive posttests because the receptive tests involve a backward association, which is less effective than a forward association (Kahana and Caplan 2002). No significant difference, however, was found between the productive recall only condition and the other three conditions on either the receptive recall or receptive recognition posttest. The results may be in part accounted for by the desirable difficulty framework, according to which a difficult learning condition facilitates transfer (e. g., Bjork 1994). Because the productive recall only condition was the most demanding among the four conditions, it might have resulted in a kind of knowledge that was transferable to novel environments. The productive recall only treatment, thus, enhanced the acquisition of not only productive but also receptive knowledge. Pedagogically, the results contradict the view that L2 words should be practiced both receptively and productively (e. g., Griffin and Harley 1996; Mondria and Wiersma 2004; Nation 2013; Webb 2009b) and imply that productive retrieval may be sufficient as long as the treatment introduces difficulty for the learner.

The lack of significant difference between the recall and productive recall only conditions on the receptive posttests in this study is inconsistent with the findings of Mondria and Wiersma (2004), who found the superiority of the receptive + productive condition over the productive only condition on a receptive posttest. The conflicting results might be due in part to methodological differences between the present study and Mondria and Wiersma. First, while the treatment was controlled by a computer program in this study, it was controlled by participants in Mondria and Wiersma. As a result, there might have been more variations in how the treatments were conducted in Mondria and Wiersma than in this study, which potentially affected the results. Furthermore, although practicing both receptive and productive retrieval is assumed to require more time than practicing only productive retrieval, the participants in the receptive + productive and productive only conditions were given the same amount of study time in Mondria and Wiersma. The study time, however, was not controlled in this study. Thus the treatments in this study

perhaps had more ecological validity in terms of study time than those in Mondria and Wiersma. These differences in the treatments could be partially responsible for the incongruent results between the present and previous research. In addition, while participants in Mondria and Wiersma took either a receptive or productive posttest, all participants took both receptive and productive posttests in this study. Because the productive recall and recognition posttests were given prior to the receptive recall and recognition posttests (see Dependent measures), the productive tests might have influenced performance on the receptive tests, possibly reducing a potential difference in the receptive test scores in this study. To diminish possible learning effects from posttests, it may be useful to administer fewer posttests per participant in future research.

5.1 Pedagogical implications

One pedagogical implication of this study is that in paired-associate learning of L2 vocabulary, recognition formats are more desirable than recall as long as the acquisition of productive knowledge of orthography is not the goal. This is based on three findings. First, this study found that the recognition condition was as effective as the other three conditions on all posttests except the productive recall posttest. The results suggest that recognition formats alone are sufficient in paired-associate learning when the precise knowledge of orthography is not required. Second, the recognition condition required the least study time and was the most efficient. Recognition, hence, may be more effective than recall in terms of efficiency. Third, the recognition condition produced significantly more correct responses during the treatment than the other three conditions. Because incorrect responses during the learning phase could potentially demotivate learners (e. g., Logan and Balota 2008), recognition formats may be more motivating. These three findings suggest that if knowledge of spelling is not required, recognition is more desirable than recall.

Although the present study demonstrated the value of the recognition formats, most existing computer-based flashcard programs offer limited capabilities regarding recognition (Nakata 2011). The limited support for the recognition formats in existing vocabulary learning software may be partly ascribed to apparent misconceptions about the value of recognition for learning. For instance, the developers of *LearnThatWord*, a web-based vocabulary learning program, claim that "Multiple choice questions are not a teaching tool at all" and "To use multiple choice questions in instructional materials is problematic!" (eSpindle Learning 2015). Based on the results of the present

study, however, it would be useful to reconsider the value of recognition for vocabulary learning when designing vocabulary learning software or exercises. Next, let us consider the implications of this study for paper-based learning. The present study showed that in paired-associate learning, (a) recall is more effective than recognition for the acquisition of productive knowledge of orthography, and (b) recall is as effective as recognition as long as the precise knowledge of orthography is not required. The findings may translate well to paper-based learning, where recall formats may be easier to implement than recognition.

On the productive recall posttest, the recall and productive recall only conditions fared better than the recognition and hybrid conditions. The former were effective on the productive recall posttest probably because they involved more productive recall questions (two in the recall and four in the productive recall only conditions per target word) than the latter (zero in the recognition and one in the hybrid conditions). A pedagogical implication of these results is that for the acquisition of productive knowledge of orthography, L2 words need to be practiced in a productive recall format at least twice. The retrieval frequency needed for learning to occur, needless to say, is likely affected by factors such as the learners' memory capacity or difficulty of target words. Another implication of this study is that retrieval success during learning is not necessarily a reliable index of learning. Although the recall and productive recall only conditions decreased learning phase performance, they turned out to be effective on the posttests. Based on the findings, it would be valuable to raise awareness that producing errors during the learning phase is not necessarily an indication of ineffective learning (e. g., Bjork 1994; Ellis 1995).

5.2 Directions for further research

Although the findings of the present study are valuable, this study also suffers from several limitations. One limitation is the relatively short duration of the treatment. In the present study, participants studied 60 word pairs in less than 45 minutes. It would be valuable to examine the effects of recall and recognition formats over a longer time period. Another limitation is the lack of prior knowledge of the participants. Because the participants in this study had no prior knowledge of Swahili, the target language, the findings of this study may not necessarily be applicable to more advanced learners. It would be useful to replicate this study with higher proficiency learners. Lastly, in the present study, the posttest timing was manipulated within participants, and each

learner took both the immediate and 1-week delayed posttests. It is possible that the administration of the immediate posttest affected performance on the delayed posttest. Future research may manipulate the posttest timing between participants.

6 Concluding remarks

In order to identify the optimal retrieval format for paired-associate learning of L2 vocabulary, the present study compared the effects of the following four conditions: recognition, recall, hybrid, and productive recall only. Results suggested that in paired-associate learning, (a) recall formats are more effective than recognition for the acquisition of productive knowledge of orthography and (b) recognition formats are more desirable than recall when knowledge of spelling is not required. The findings of this study have direct and immediate application to learning because the four retrieval formats used in this study are common within a range of different vocabulary learning activities such as flashcard learning, fill-in-the-blank exercises, and word-definition matching (e. g., Nakata and Webb 2016; Nation and Webb 2011). Future studies investigating the effects of retrieval formats are of value because they may further help us to determine how we can optimize L2 vocabulary from retrieval.

Acknowledgments: This research was supported by the Faculty Research Grant (#98778) and the Victoria PhD Scholarship from Victoria University of Wellington as well as Grant-in-Aid for Research Activity Start-up (#15H06746) and Grant-in-Aid for Young Scientists (A) (#16H05943) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan. This article is based on part of the author's doctoral dissertation, which was submitted to Victoria University of Wellington in 2013. I am very grateful to Stuart Webb, Paul Nation, Rod Ellis, Jan Hulstijn, Stuart McLean, and anonymous reviewers for their invaluable advice. I would also like to extend my special thanks to Yu Tamura for his assistance with data analysis.

References

Baddeley, A. D. 1997. *Human memory: Theory and practice* (Revised ed.). East Sussex, UK: Psychology Press.

Barcroft, J. 2007. Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning* 57. 35–56.

Barcroft, J. 2012. *Input-based incremental vocabulary instruction*. Alexandria, VA: TESOL.
Bjork, R. A. 1994. Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (eds.), *Metacognition: Knowing about knowing*, 185–205. Cambridge, MA: MIT Press.
Cepeda, N. J., E. Vul, D. Rohrer, J. T. Wixted & H. Pashler. 2008. Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science* 19. 1095–1102.
Chen, C.-H. & K. Huang. 2014. The effects of response modes and cues on language learning, cognitive load and self-efficacy beliefs in web-based learning. *Journal of Educational Multimedia and Hypermedia* 23. 117–134.
Clariana, R. B. & D. Lee. 2001. The effects of recognition and recall study tasks with feedback in a computer-based vocabulary lesson. *Educational Technology Research and Development* 49. 23–36.
Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*, 2nd edn. Hillsdale, NJ: Erlbaum.
Elgort, I. 2011. Deliberate learning and vocabulary acquisition in a second language. *Language Learning* 61. 367–413.
Elgort, I. & A. E. Piasecki. 2014. The effect of a bilingual learning mode on the establishment of lexical semantic representations in the L2. *Bilingualism: Language and Cognition* 17. 572–588.
Ellis, N. C. 1995. The psychology of foreign language vocabulary acquisition: Implications for CALL. *Computer Assisted Language Learning* 8. 103–128.
Ellis, N. C. & A. Beaton. 1993. Factors affecting the learning of foreign language vocabulary: Imagery keyword mediators and phonological short-term memory. *The Quarterly Journal of Experimental Psychology Section A* 46. 533–558.
eSpindle Learning. 2015. Answers to common questions about our vocabulary and spelling program. *LearnThatWord*. https://www.learnthat.org/pages/view/faq.html (accessed 3 March 2015)
Griffin, G. F. & T. A. Harley. 1996. List learning of second language vocabulary. *Applied Psycholinguistics* 17. 443–460.
Harris, D. P. 1969. *Testing English as a second language*. New York, NY: McGraw-Hill.
Jaeger, T. F. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59. 434–446.
Jarvis, S. 2009. Lexical transfer. In A. Pavlenko (ed.), *The bilingual mental lexicon: Interdisciplinary approaches*, 99–124. Clevedon, UK: Multilingual Matters.
Jiang, N. 2000. Lexical representation and development in a second language. *Applied Linguistics* 21. 47–77.
Kahana, M. J. & J. B. Caplan. 2002. Associative asymmetry in probed recall of serial lists. *Memory & Cognition* 30. 841–849.
Karpicke, J. D. & H. L. Roediger. 2007. Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language* 57. 151–162.
Karpicke, J. D. & H. L. Roediger. 2008. The critical importance of retrieval for learning. *Science* 319. 966–968.
Kornell, N. 2009. Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology* 23. 1297–1317.
Kroll, J. F. & E. Stewart. 1994. Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language* 33. 149–174.

Laufer, B. & Z. Goldstein. 2004. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning* 54. 399–436.

Laufer, B. & K. Shmueli. 1997. Memorizing new words: Does teaching have anything to do with it? *RELC Journal* 28. 89–108.

Logan, J. M. & D. A. Balota. 2008. Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition* 15. 257–280.

Mondria, J. A. 2003. The effects of inferring, verifying, and memorizing on the retention of L2 word meanings: An experimental comparison of the "Meaning-Inferred Method" and the "Meaning-Given Method." *Studies in Second Language Acquisition* 25. 473–499.

Mondria, J. A. & B. Wiersma. 2004. Receptive, productive and receptive + productive L2 vocabulary learning: What difference does it make? In P. Bogaards & B. Laufer, (eds.), *Vocabulary in a second language: Selection, acquisition, and testing*, 79–100. Amsterdam, The Netherlands: Benjamins.

Morris, C. D., Bransford, J. D. & J. J. Franks. 1977. Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior* 16. 519–533.

Nakata, T. 2011. Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning* 24. 17–38.

Nakata, T. 2013. *Optimising second language vocabulary learning from flashcards (Unpublished doctoral dissertation)*. New Zealand: Victoria University of Wellington.

Nakata, T. in press. Does repeated practice make perfect? The effects of within-session repeated retrieval on second language vocabulary learning. *Studies in Second Language Acquisition*.

Nakata, T. & S. A. Webb. 2016. Vocabulary learning exercises: Evaluating a selection of exercises commonly featured in language learning materials. In B. Tomlinson, (ed.), *SLA research and materials development for language learning*, 123–138. Oxon, UK: Routledge.

Nation, I. S. P. 2013. *Learning vocabulary in another language*, 2nd edn. Cambridge, UK: Cambridge University Press.

Nation, I. S. P. & S. A. Webb. 2011. *Researching and analyzing vocabulary*. Boston, MA: Heinle Cengage Learning.

Nelson, T. O. & J. Dunlosky. 1994. Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory* 2. 325–335.

Park, J. 2005. Learning in a new computerized testing system. *Journal of Educational Psychology* 97. 436–443.

Prihoda, T. J., R. N. Pinckard, C. A. McMahan & A. C. Jones. 2006. Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. *Journal of Dental Education* 70. 378–386.

Pyc, M. A. & K. A. Rawson. 2009. Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language* 60. 437–447.

Read, J. 2004. Plumbing the depths: How should the construct of vocabulary knowledge be defined? In P. Bogaards & B. Laufer (eds.), *Vocabulary in a second language: Selection, acquisition, and testing*, 209–227. Amsterdam, The Netherlands: Benjamins.

Schmitt, N. 1997. Vocabulary learning strategies. In N. Schmitt & M. McCarthy (eds.), *Vocabulary: Description, acquisition and pedagogy*, 199–227. Cambridge, UK: Cambridge University Press.

Schneider, V. I., A. F. Healy & L. E. Bourne. 2002. What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language* 46. 419–440.

Smith, M. A. & J. D. Karpicke. 2014. Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory* 22. 784–802.

Steinel, M. P., J. H. Hulstijn & W. Steinel. 2007. Second language idiom learning in a paired-associate paradigm: Effects of direction of learning, direction of testing, idiom image-ability, and idiom transparency. *Studies in Second Language Acquisition* 29. 449–484.

Van Bussel, F. J. J. 1994. Design rules for computer-aided learning of vocabulary items in a second language. *Computers in Human Behavior* 10. 63–76.

Webb, S. A. 2007. Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language Teaching Research* 11. 63–81.

Webb, S. A. 2008. Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition* 30. 79–95.

Webb, S. A. 2009a. The effects of pre-learning vocabulary on reading comprehension and writing. *Canadian Modern Language Review* 65. 441–470.

Webb, S. A. 2009b. The effects of receptive and productive learning of word pairs on vocabulary knowledge. *RELC Journal* 40. 360–376.

Wissman, K. T., K. A. Rawson & M. A. Pyc. 2012. How and when do students use flashcards? *Memory* 20. 568–579.

Yun, S., P. C.Miller, Y. Baek, JJung. & M. Ko. 2008. Improving recall and transfer skills through vocabulary building in web-based second language learning: An examination by item and feedback type. *Educational Technology & Society* 11. 158–172.