The deliberate study of concrete nouns with tablet-based augmented reality

Adam Dabrowski, The University of Electro-Communications
Stuart McLean, Kindai University
Christopher Nicklin, The University of Tokyo

Abstract

Three modes of deliberate vocabulary study were investigated to determine how well they assisted learners’ recall of the meaning of target concrete nouns. Two modes of tablet-based augmented reality, one context-independent (AR1) and one context-dependent (AR2), were compared with each other and with paper-based word cards (WC) in the deliberate study of three sets of nonwords representative of concrete nouns. An orthogonal Latin square design was used to counterbalance 39 participants. We hypothesized that both AR conditions would be more beneficial than word cards in terms of participants’ ability to retain the meaning of the target words as demonstrated by performance on Yes/No and meaning-recall test items, and that AR2 conditions would be more beneficial as compared with AR1 conditions. Generalized linear mixed models revealed that both AR study modes significantly outperformed word cards. The context-dependent and context-independent augmented reality study modes did not significantly differ indicating that a visuospatial bootstrapping effect (VSB) was likely at play regardless of how dependent on or independent of their respective scenes the items studied were. These findings offer pedagogical implications of mobile-based AR use in vocabulary study and language learning in general.

Keywords: Augmented Reality, Deliberate Vocabulary Learning, Tablet-based AR, Visuospatial Bootstrapping

Language(s) Learned in This Study: Artificial language, Nonwords


Introduction

The intentional study of vocabulary items is well researched, and a variety of methods and approaches to intentional vocabulary learning have been proven to be effective. Paper-based modes of deliberate vocabulary study, particularly word cards, have been used for vocabulary study for some time and are regarded as effective for the purpose of both explicit (Beaton, Gruneberg, & Ellis, 1995; Nakata, 2019; Webb, 2007; Wilkinson, 2020) and implicit or incidental vocabulary acquisition (e.g., Elgort, 2011; Elgort & Piatecki, 2013). Intentional vocabulary study leads to greater rates of growth in shorter periods of time as compared with incidental vocabulary acquisition methods (Schmitt, 2008; Webb, 2019). Furthermore, word cards in particular have been shown to outperform other intentional study devices such as lists (Nakata, 2008; Nation, 2013; Wilkinson, 2020), which could be in part due to their ability to encourage and strengthen retrieval processes (Nakata, 2019).

Smartphones, tablets, and computers are now commonplace in society. The utility of such technology has been adopted by many teachers and students and has presented language learners with new methods for vocabulary study, which in turn have extended language learning possibilities. Augmented Reality (AR)
is an emerging technology which allows digital information to be overlaid onto a user’s current experience of reality via an electronic device and appears to have many possibilities for pedagogical application. The digital information augmenting the user’s reality is commonly visual stimuli displayed on either a mobile device or a head-mounted display, though other types of AR stimuli include audio, haptic feedback, or scent. Though the number of studies focusing on SLA with AR at present are limited, and the range of focus of these studies is wide, the role of AR in the field of SLA is now being investigated. Several researchers have reported on the efficacy of AR in intentional vocabulary acquisition (i.e. Chen & Chan, 2019; Geng & Yamada, 2020; Ibrahim et al., 2018; Larchen Costuchen et al., 2021; Liu et al., 2016; Solak & Cakır, 2015; Taskiran, 2019; Zainuddin et al., 2016).

When one thinks of AR, they might imagine a bulky, head-mounted apparatus similar in form to many of the virtual reality headsets presented in popular media. While there are relatively expensive advanced AR headsets currently available on the market, the smartphones and tablets which most teachers and students have access to are already capable of providing an AR experience. The ubiquity of these devices presents opportunities for practical pedagogical application and research of such applications. This study focuses on one such application, the deliberate study of concrete nouns, and compares two variations of tablet-based AR study modes (see Figure 1) with each other and with paper-based word cards.

**Intentional Vocabulary Study**

Vocabulary learning approaches can be subordinated into two categories: Intentional (deliberate) approaches, those which require deliberate motive, and incidental approaches, those which occur naturally and without any sense of deliberacy (Nation, 2013; Schmitt, 2000; Webb, 2020), such as when encountering an unfamiliar word in an otherwise familiar sentence and using the context of the sentence to help guess the meaning of the word. When compared with incidental vocabulary learning activities, intentional methods of vocabulary study offer learners the ability to make gains quickly, attain higher rates of retention, and reach productive mastery of the items they study faster (Schmitt, 2008; Webb, 2019). Of the many activities and approaches designed for intentional vocabulary acquisition which have been described at length by researchers (e.g., Morgan & Rinvolutri, 2004; Webb & Nation, 2017), using word cards or flashcards is one conventional approach to intentional vocabulary study whose efficiency of use and efficacy are supported by over a century of research (Nakata, 2008; Nation, 2013; Webb & Nation, 2017) and appears to prove more effective of a study method as compared with study of lists (see Komachali & Khodareza, 2012; Kuo & Ho, 2012; Mondria & Mondria-de Vries, 1994; Nakata, 2008).

**AR in Vocabulary Learning Research**

The current body of literature on AR-based vocabulary learning in SLA spans several target languages. The use of AR as a mode of study has been researched to understand various aspects of vocabulary acquisition, such as the importance of physical context, cognitive load, motivation, and engagement. Many studies focus on the acquisition of English vocabulary in EFL contexts (i.e. Chen & Chan, 2019; Liu et al., 2016; Solak & Cakır, 2015; Taskiran, 2019), but studies have also examined other target languages, for example, Japanese, Basque, Spanish, and Arabic (see respectively Geng & Yamada, 2020; Ibrahim et al., 2018; Larchen Costuchen et al., 2021; Zainuddin et al., 2016). One common theme in the few studies which examine the use of AR in vocabulary learning is the comparison of an AR device’s efficacy for vocabulary learning with traditional paper-based modes of study, such as word cards or vocabulary-focused lessons in teacher-fronted classrooms. In some of these studies researchers have concluded that AR modes compare similarly with traditional modes of study in terms of retention (e.g., Chen & Chan, 2019; Geng & Yamada, 2020), while other studies have found AR study modes to be more effective in vocabulary retention (e.g., Dabrowski, 2023; He et al., 2014; Ibrahim et al., 2018; Larchen Costuchen et al., 2021). All of these studies have made the call for more research into this new area of vocabulary study.
The Importance of Place: From the Method of Loci to Visuospatial Bootstrapping

The method of loci, also referred to as the memory palace, is an ancient mnemonic method once used by the orators of ancient Rome to remember public speeches (Yates, 1999). Users of this method mentally construct an imagined, familiar place, such as a palace or building, which houses a familiar path along which the orator places imagined images of items. The items are used as key mnemonic devices to aid in the recall of content represented by the images (Moë & De Beni, 2005). In the memory palace, the use of visual space and a pre-established path through the space is paramount in the recall of that which the user desires to remember.
Research focused on the binding of visuospatial and verbal information in working memory has suggested that in some cases a visuospatial contextualized scaffold, a spatial array represented by a keypad used for recalling strings of numbers, can significantly aid in recalling longer strings of numbers (Darling & Havelka, 2010). The researchers coined this newly discovered effect the visuospatial bootstrapping (VSB) effect, and in related studies, it was suggested that this phenomenon, which is involved in storing short-term visuospatial and verbal information, is also connected to verbal knowledge stored in long-term memory (Calia, Darling, & Havelka, 2019; Darling, Allen, & Havelka, 2017). In another related study, Darling et al. (2020) conducted two experiments which demonstrated that the availability of spatialized information contributed to effectively learning sequences of digits as well as nonwords. They also concluded that spatial information contributing to the VSB effect need not be well known; spatial information can be learned while on task and still be of benefit to longer-term memory creation. Other studies focused on the role of spatiality in memory have observed that spatialization is involved with working memory (e.g., Guida et al., 2018).

In a study that employed AR in Spanish idiom learning, Larchen Costuchen et al. (2021) compared an AR-VSB method of vocabulary study, one which makes use of AR in physical space to benefit the creation of salient memory, with electronic word cards to observe vocabulary retention as measured at different intervals of time. QR codes, which triggered language and images to appear within a mobile-based AR application, were placed within the participants’ living spaces either on the floors or walls near familiar objects. The participants spent 15 minutes studying the target idioms with the method assigned to them. Translation tasks and meaning-recall questions were used to assess the participants’ idiom knowledge. The results indicated that participants who studied with the AR-VSB condition performed significantly better than an electronic word card group on both 15-minute delayed posttests ($p < .01, d = 1.33$) and on one-week delayed posttests ($p < .01, d = 1.20$).

The Present Study

As augmented reality is a new and emerging technology, even in the subcategory of deliberate vocabulary learning with AR, there are numerous gaps worthy of investigation. The present study differs from previous research in that it examines the acquisition of nonwords to help control for natural learning effects that present potentially confounding variables. The deliberate study of nonwords via an AR-VSB method has not yet been investigated. Nonwords are meaningless strings of letters which orthographically and phonologically resemble the features of real words. Nonwords were used in the current study to guard against any influence that prior language learning experience might have contributed to observed growth. The process of recalling a nonword involves similar discretion processes to the production of a natural word (Ebert, 2014) and are a good tool for measuring both production and comprehension (Chuang et al., 2021). For this reason they were selected for this study. The present study adopts elements of previous studies in that it compares variations of the AR conditions in deliberate vocabulary study and it incorporates variations of preconstructed scenes which use an AR-VSB method. The current study also examines potential differences in vocabulary retention based on variations of the contextual uniformity and relevance of a scene. The present study was guided by the following research questions:

1. How do tablet-based AR technology modes compare with paper-based word cards in terms of their ability to help a learner recall the meaning of concrete nouns?
2. When concrete nouns are studied via AR, is there a difference as to whether those lexical items are anchored to a contextually relevant scene arranged for utility or not?

Based on the findings in previous research it was hypothesized that AR modes would perform better than word cards in the retention of the target noun meanings and that the context-dependent AR scenes would help learners recall the meaning of the nouns studied more so than the context-independent AR scenes.
Methodology

Design
The deliberate study of three separate sets of nonwords using three modes of study: word cards (WC), context-independent tablet-based AR-VSB preconstructed scenes (AR1), and context-dependent tablet-based AR-VSB preconstructed scenes (AR2) were investigated in this study. The study employed a pretest, posttest, and delayed posttest design with one treatment session per set of words per participant. A three-by-three counterbalanced Latin square design was employed, whereby each of the 39 participants were assigned to one of three experimental groups: A, B, or C (see Table 1). The participants were assigned through stratified random sampling, accounting for their native language(s) and educational level. However, both variables proved to be non-significant predictors and were excluded from the final model (see the Supplementary Materials in Appendix D).

Table 1

3 x 3 Latin Square Design

<table>
<thead>
<tr>
<th>Word Set / Treatment Mode</th>
<th>WC Word Cards</th>
<th>AR1 Context-Independent AR-VSB Scene</th>
<th>AR2 Context-Dependent AR-VSB Scene</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word Set 1 (DESK)</td>
<td>Group A</td>
<td>Group B</td>
<td>Group C</td>
</tr>
<tr>
<td>Word Set 2 (KITCHEN)</td>
<td>Group C</td>
<td>Group A</td>
<td>Group B</td>
</tr>
<tr>
<td>Word Set 3 (MUSIC)</td>
<td>Group B</td>
<td>Group C</td>
<td>Group A</td>
</tr>
</tbody>
</table>

Word Sets
Due to the necessity for a cohesive and realistic context required by the AR2 scenes, three sets of ten concrete nouns loosely related by semantic set were created. Each set of ten words was related thematically so that they could be embodied by actual objects and placed in a realistic context for the AR2 scenes. The three thematic scenes included a desk scene, a kitchen scene, and a music scene. The words for each set were selected with the findings of Tinkham (1993) and Waring (1997) in mind; care was taken not to include items which too closely overlapped with each other semantically. The descriptive statistics for the target words are reported in Table 2. Confidence interval overlap can be considered as representative of non-significant difference between two variables (Cumming, 2012), thus the confidence interval overlap between the three thematic scenes in terms of letter length, difficulty, COCA frequency and range, and concreteness indicated that the three sets were well matched in terms of these variables. The kitchen item English equivalents were significantly shorter than the other two sets in terms of syllable length, but in practical terms the difference was negligible.

In order to closely control the orthographic and phonological features of the nonwords we selected, a set of seventy-five nonwords was generated using the ARC Nonword Database (Rastle et al., 2002), which had only orthographically existing onsets, only orthographically existing bodies, and were only legal bigrams. The number of letters in each word was fixed at five, a minimum neighborhood size was set at 4, and the summed frequency of orthographic and phonological neighbors was set to 0 (Zhang et al., 2020).
These words were then reviewed for any instances of -ed or -en. These items were removed as they orthographically resembled verbs. Ten words were chosen at random from the remaining nonwords for each of the three sets. These words were then randomly assigned to the ten English words selected for each set. The Japanese translation of the English word was then determined and assigned to its nonword (see Appendix A).

**Table 2**

**Descriptive Statistics for Target Words**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Set a</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>95% CI</th>
<th>Skew</th>
<th>SES</th>
<th>Kurt</th>
<th>SEK</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>D</td>
<td>10</td>
<td>5.6</td>
<td>2.54</td>
<td>[4.13, 7.07]</td>
<td>0.35</td>
<td>0.68</td>
<td>-0.85</td>
<td>1.33</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>10</td>
<td>4.7</td>
<td>2.11</td>
<td>[3.48, 5.92]</td>
<td>1.99</td>
<td>0.68</td>
<td>4.65</td>
<td>1.33</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>10</td>
<td>5.6</td>
<td>1.95</td>
<td>[4.47, 6.73]</td>
<td>1.15</td>
<td>0.68</td>
<td>0.05</td>
<td>1.33</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Syllables</td>
<td>D</td>
<td>10</td>
<td>1.7</td>
<td>0.67</td>
<td>[1.31, 2.09]</td>
<td>0.43</td>
<td>0.68</td>
<td>-0.28</td>
<td>1.33</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>10</td>
<td>1.1</td>
<td>0.31</td>
<td>[0.92, 1.28]</td>
<td>3.16</td>
<td>0.68</td>
<td>10</td>
<td>1.33</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>10</td>
<td>2</td>
<td>0.81</td>
<td>[1.53, 2.47]</td>
<td>0.68</td>
<td>-1.39</td>
<td>1.33</td>
<td>1</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Difficulty Y/N c</td>
<td>D</td>
<td>10</td>
<td>-1.32</td>
<td>0.39</td>
<td>[-2.20, -0.43]</td>
<td>-0.11</td>
<td>0.68</td>
<td>-1.06</td>
<td>1.33</td>
<td>-1.81</td>
<td>-0.7</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>10</td>
<td>-1.07</td>
<td>0.71</td>
<td>[-2.67, 0.54]</td>
<td>0.45</td>
<td>0.68</td>
<td>-0.06</td>
<td>1.33</td>
<td>-2</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>10</td>
<td>-1.08</td>
<td>0.81</td>
<td>[-2.90, 0.74]</td>
<td>-0.15</td>
<td>0.68</td>
<td>-0.47</td>
<td>1.33</td>
<td>-2.5</td>
<td>0.01</td>
</tr>
<tr>
<td>Difficulty MR c</td>
<td>D</td>
<td>10</td>
<td>1.79</td>
<td>0.51</td>
<td>[0.63, 2.96]</td>
<td>-0.56</td>
<td>0.68</td>
<td>0.76</td>
<td>1.33</td>
<td>0.79</td>
<td>2.63</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>10</td>
<td>1.74</td>
<td>0.59</td>
<td>[0.40, 3.09]</td>
<td>-0.14</td>
<td>0.68</td>
<td>-0.43</td>
<td>1.33</td>
<td>0.73</td>
<td>2.57</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>10</td>
<td>1.79</td>
<td>0.82</td>
<td>[-0.06, 3.64]</td>
<td>-1.67</td>
<td>0.68</td>
<td>3.71</td>
<td>1.33</td>
<td>-0.19</td>
<td>2.62</td>
</tr>
<tr>
<td>COCA Freq (per mil)</td>
<td>D</td>
<td>10</td>
<td>86.42</td>
<td>134.9</td>
<td>[8.35, 164.49]</td>
<td>2.27</td>
<td>0.68</td>
<td>5.44</td>
<td>1.33</td>
<td>3.26</td>
<td>434.13</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>10</td>
<td>30.89</td>
<td>32.67</td>
<td>[11.98, 49.80]</td>
<td>1.73</td>
<td>0.68</td>
<td>3.4</td>
<td>1.33</td>
<td>1</td>
<td>109.59</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>10</td>
<td>82.01</td>
<td>147.94</td>
<td>[0.00, 167.62]</td>
<td>2.26</td>
<td>0.68</td>
<td>4.94</td>
<td>1.33</td>
<td>0.44</td>
<td>457.06</td>
</tr>
<tr>
<td>Range (%)</td>
<td>D</td>
<td>10</td>
<td>6.77</td>
<td>7.53</td>
<td>[2.41, 11.13]</td>
<td>1.34</td>
<td>0.68</td>
<td>1.35</td>
<td>1.33</td>
<td>0.4</td>
<td>23.3</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>10</td>
<td>2.57</td>
<td>2.06</td>
<td>[1.38, 3.76]</td>
<td>0.58</td>
<td>0.68</td>
<td>-1.09</td>
<td>1.33</td>
<td>0.1</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>10</td>
<td>5.52</td>
<td>9.21</td>
<td>[0.19, 10.85]</td>
<td>2.43</td>
<td>0.68</td>
<td>6.13</td>
<td>1.33</td>
<td>0.1</td>
<td>29.7</td>
</tr>
<tr>
<td>MRC Concreteness d</td>
<td>D</td>
<td>7</td>
<td>562</td>
<td>40.534</td>
<td>[524.46, 599.54]</td>
<td>-0.73</td>
<td>0.79</td>
<td>-1.27</td>
<td>1.59</td>
<td>502</td>
<td>599</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>7</td>
<td>592.714</td>
<td>24.757</td>
<td>[569.79, 615.64]</td>
<td>-1.17</td>
<td>0.79</td>
<td>0.45</td>
<td>1.59</td>
<td>548</td>
<td>615</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>8</td>
<td>584.125</td>
<td>29.255</td>
<td>[559.61, 608.64]</td>
<td>0.33</td>
<td>0.75</td>
<td>0.47</td>
<td>1.48</td>
<td>539</td>
<td>635</td>
</tr>
</tbody>
</table>

a D = Desk, K = Kitchen, and M = Music.

b 95% CIs calculated with t(9) = 2.26 to account for n = 10, t(8) = 2.31 to account for n = 9, t(7) = 2.37 to account for n = 8, and t(6) = 2.45 to account for n = 7.

c Difficulty in Rasch logits: see Dabrowski (2022) for details of the Rasch analysis.

d MRC concreteness values were obtained from Tool for Automatic Analysis of Lexical Sophistication (TAALES; Kyle et al., 2018) but were not available for all words.
Treatment Modes

The words were presented to participants under one of three conditions, comprising word cards and two AR conditions. Under the word card condition, participants were presented with a set of paper-based word cards containing the target vocabulary. Word cards were created beforehand by the researchers for each of the three sets of nonwords (see Appendix B). In this study, we sought to understand if AR was capable of helping to better facilitate a form-meaning link. As all participants were fluent speakers of either English and/or Japanese, both languages were provided to demonstrate the creation of a form-meaning link between a target nonword and a concrete noun known to the participant. Access to Japanese and English was provided to mitigate any burden presented in learning an unfamiliar L2 word in addition to the nonword. Though these differences in proficiency present a possible confound, differential item functioning (DIF) analyses were carried out in a separate study by Dabrowski (2022) which collected and presented validity evidence for the tests used in this study and found no unfair advantages of four possible membership classes, one of which was L1 language representation on the word cards or AR labels, and no unfair advantages on the basis of native lexis representation were detected (this is discussed in more depth in the discussion section below).

Each of the AR treatment modes described next made use of a 2018 iPad Pro, which measures 10.5 inches and has a resolution of 1668 x 2224 pixels, a 4:3 ratio, and 264 pixels per square inch. The creation of the AR scenes and assignment of virtual labels within the scenes were carried out by the researchers before treatment with a paid subscription-model application called ARIO (Ario Technologies Incorporated, 2022), which was used to assign virtual labels to the items within each of the AR-VSB scenes. Because ARIO can recognize a previously scanned environment, the AR labels can be stored within the scene and restored with the software by scanning the scene again at a later time with the camera of the tablet.

The context-independent tablet-based AR-VSB preconstructed scene conditions (AR1) involved viewing a context-independent real-world scene populated with actual objects or realia with the tablet-based augmented reality application. The AR labels were formatted the same as the word cards. The items within these scenes were arranged in a random manner atop of a table. The placement of the items were indicative of an inventory being laid out. The arrangement of the items was the same for all participants who experienced them in that scene. AR1 scenes were created for each of the three sets of nonwords (see Appendix B).

The context-dependent tablet-based AR-VSB preconstructed scene conditions (AR2) involved viewing a theorized context-dependent real-world scene populated with actual objects or realia with ARIO, which was again used to assign virtual labels to the items within the scene. The AR labels were formatted the same as the word cards and AR1 mode of study. The items within these scenes were arranged in a manner which invited utility. The placement of the items in the AR2 scenes was indicative of a usable space. For example, the items in the Desk scene were arranged such that they resembled the workspace of a student who had left the scene briefly for a short break: the computer open, notebook and writing utensils next to the computer, and other stationery available to a possible user of the scene; the Music scene was arranged so that the piano and its accouterments were facing a user who might sit down in front of the table, the guitar was perched in a guitar stand, and other items were placed in a way so that they could be used with ease; and the Kitchen scene made use of an actual kitchenette area—the items to be studied as vocabulary were placed on shelves and counters within the real-world scene. In each of these three AR2 scenes, the arrangement of the items was the same for all participants who experienced them in that scene. AR2 scenes were created for each of the three sets of nonwords (for pictures please see Appendix B).

Participants

The participants were 39 adult learners, 20 females and 19 males, who reside and work or study in Tokyo, Japan. Seventeen of the participants were doctoral students attending an American university in Japan who were also working as professors or instructors, six were undergraduate students pursuing degrees in psychology, and the remaining 15 participants were university professors or language instructors. All
participants had a native or near-native command of either English or Japanese. There were 21 native English speakers, 13 native Japanese speakers, two native Cantonese speakers, one native speaker of Mandarin, one native speaker of Dutch, and one native speaker of Portuguese. Though the sample was somewhat heterogeneous, it should be reinforced that the sample frame was drawn at an American university. Of membership classes examined in four separate DIF analyses, neither L1 nor L2 representation in word cards or AR labels unfairly rewarded any L1 membership classes indicating that the L1 variety did not pose any issues (Dabrowski, 2022). Based on Brysbaert and Steven’s (2018) recommendation to collect a minimum of 1,600 observations to ensure a well-powered design, we initially settled on a target $n$ size of 36 participants, which when multiplied by the 75 questions in the delayed posttests (as only delayed posttest data were to be analyzed), yields a total of 2,700 observations. This $n$ size also lent to equally distributing the participants across the Latin square design, as it is divisible by three. In the process of recruiting, we recruited three additional participants beyond our target, increasing the final total number of observations to 2,925, nearly double that of Brysbaert and Steven’s recommendation (2018).

**Assessments**

A pretest, an immediate posttest, and a delayed posttest of participant knowledge of the target nonwords were created for each of the three sets of words. All of these assessments were administered with Google Forms in an internet browser. The pretests included the ten items from their corresponding sets. Although Yes/No questions have been found to overestimate knowledge (Stubbe, 2012), in this study the Yes/No questions were used to establish a baseline. All participants scored zero across all pretests indicating that the participants were completing the tasks with honesty. The immediate posttests contained two sections: a Yes/No section and a meaning-recall section. Meaning-recall items were chosen for their ability to discriminate whether or not a form-meaning link had been created. Such recall test items are more difficult than multiple-choice receptive test items (Laufer & Goldstein, 2004). In the Yes/No section, 15 words were presented to the participant and they judged if they remembered having learned the word. In addition to the ten words within the set, an additional five distractors (selected from the ARC-generated nonword list from which the target words were selected) were included in the Yes/No section. The meaning-recall section included only the ten items the participant studied within that word set. The form of the target word was presented to the participant, and they were asked to write the corresponding English or Japanese meaning of the target word (Read, 2019; Stoeckel et al., 2021). The primary purpose of the immediate posttests was to train, orient, and familiarize the participants with the test format and the meaning-recall question types that would also appear on the delayed posttest (whose data were used for our analyses). The delayed posttests were duplicates of the immediate posttests yet with the order of the questions randomized to guard against testing, listing, or learning effects (to access copies of the Google Forms used in this study, see Appendix C). The meaning-recall sections of the delayed posttest results were scored independently by two raters, $K = .92$ [.89, .94], indicating very good inter-rater reliability.

**Procedure**

Each participant met with the lead researcher individually at a university in Tokyo. The participants read and signed an informed consent agreement and were oriented to the experiment assessment formats. First, participants completed the pretest for the relevant set, before being oriented to the specific mode of study (either WC, AR1, or AR2). They were then given ten minutes to study the words and were informed that there would be an immediate posttest following the study session and a delayed posttest one week later. Following the task, the researcher distracted the participants with exactly one minute of unrelated conversation. The participants then took the immediate posttest which was made up of 15 Yes/No items and 10 meaning-recall items via a Google Forms survey. Each participant repeated this process according to their assigned treatment order until all three sets had been studied with the treatment type assigned to the participant as per their grouping within the Latin square design. They then completed the delayed posttests one week after the treatment sessions via a Google Forms survey in a quiet location. The delayed posttests were administered in the same order that the participants had studied the sets. A procedure
overview can be referenced in Figure 2.

**Figure 2**

**Procedure Overview**

1. Consent form and orientation to the experiment
2. Session 1: Word set and treatment mode determined by the participant’s assignment within the Latin square design
   2.1. Pretest
   2.2. 10 minutes of study with assigned treatment mode
   2.3. 1 minute of verbal distraction
   2.4. Posttest
3. Session 2: Word set and treatment mode determined by the participant’s assignment within the Latin square design
   3.1. Pretest
   3.2. 10 minutes of study with assigned treatment mode
   3.3. 1 minute of verbal distraction
   3.4. Posttest
4. Session 3: Word set and treatment mode determined by the participant’s assignment within the Latin square design
   4.1. Pretest
   4.2. 10 minutes of study with assigned treatment mode
   4.3. 1 minute of verbal distraction
   4.4. Posttest
5. Delayed Posttests
   5.1. Conducted 1 week after treatment
   5.2. Conducted in the same order as the participant’s assigned set order

**Analyses**

To address the two research questions, generalized linear mixed-effects models (GLMM) were fit to the Yes/No and meaning-recall delayed posttest data with the *lme4* package (Bates et al., 2015) for R (R Core Team, 2022). We chose to include only the delayed posttest data in our analyses for three reasons: (1) As was mentioned in the Assessments section above, the main purpose of the immediate posttests was to familiarize the participants with the MR question types which were not included in the pretests, (2) as participants had all answered ‘No’ to all items in the pretests, a clear baseline in word knowledge was established, and (3) we were less interested in observing the short- to medium-term retention of vocabulary and instead interested in the creation of a lasting form-meaning link. A GLMM was the most appropriate model because of the binary nature of the data, which was modeled with a binomial distribution. The model building process and assumption checks are reported in Supplementary Materials B (see Appendix D). The final model was a random-intercepts-only model, with intercepts modeled for each participant and item. Although such models are prone to Type I error (e.g., Barr et al., 2013), this parsimonious model resulted in the best fit. The GLMM fixed effects comprised four categorical variables. *Group* was a three-category variable representing which group a participant was assigned to. The levels consisted of Group A, B, and C, with A being the reference level. *Treatment* was a three-category variable relating to the treatment and consisted of WC, AR1, and AR2, with WC being the reference level. Finally, *Set* was a three-category variable representing the word group being tested. The levels comprised desk, kitchen, and music, with desk constituting the reference level. Odds ratios (ORs), which are the exponential of the beta coefficient and represent the size of the effect, were reported, along
with confidence intervals (CIs) calculated with the *broom.mixed* (Bolker & Robinson, 2021) package for R. An OR of 1.20 implies that for every 1 unit increase in the predictor (i.e., the difference between the category levels in the present model) there is a 20% increase in the likelihood of a positive response to the dependent variable. Pseudo-$R^2$ effect sizes for the GLMM were estimated with the *MuMIn* (Barton, 2020) package for R.

**Results**

The descriptive statistics for the delayed posttest results displayed in Table 3 offered initial evidence that the augmented reality treatments did not result in increased vocabulary knowledge in comparison to word cards. Because the Yes/No and meaning-recall scores consisted of 15 and 10 items respectively, the scores are reported as percentages for comparability. For the Yes/No data, confidence interval (CI) overlap between the three treatments was observed within each set (i.e., Desk, Kitchen, and Music) and also between sets. The same pattern was also observed for the meaning-recall data. The lack of overlap between the CIs for the Yes/No and meaning-recall scores indicated a significant difference between the scores on both tests. With this in mind, separate GLMMs were constructed for each set of test results, and each will now be considered in turn (see Table 4 for the results from both models).

**Table 3**

*Descriptive Statistics for the Delayed Posttest Results*

<table>
<thead>
<tr>
<th>Test</th>
<th>Set</th>
<th>Treatment</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>95% CI</th>
<th>Skew</th>
<th>Kurt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y/N</td>
<td>Desk</td>
<td>WC</td>
<td>13</td>
<td>81.54</td>
<td>15.67</td>
<td>53.33</td>
<td>100</td>
<td>[72.06, 91.00]</td>
<td>-0.24</td>
<td>-0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR1</td>
<td>13</td>
<td>94.36</td>
<td>5.99</td>
<td>86.67</td>
<td>100</td>
<td>[90.73, 97.97]</td>
<td>-0.34</td>
<td>-1.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR2</td>
<td>13</td>
<td>89.23</td>
<td>10.38</td>
<td>73.33</td>
<td>100</td>
<td>[82.95, 95.50]</td>
<td>-0.29</td>
<td>-1.43</td>
</tr>
<tr>
<td></td>
<td>Kitchen</td>
<td>WC</td>
<td>13</td>
<td>85.64</td>
<td>14.36</td>
<td>53.33</td>
<td>100</td>
<td>[76.96, 94.31]</td>
<td>-1.01</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR1</td>
<td>13</td>
<td>84.10</td>
<td>12.92</td>
<td>60.00</td>
<td>100</td>
<td>[76.29, 91.91]</td>
<td>-0.41</td>
<td>-0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR2</td>
<td>13</td>
<td>93.85</td>
<td>8.37</td>
<td>73.33</td>
<td>100</td>
<td>[88.78, 98.90]</td>
<td>-1.36</td>
<td>1.55</td>
</tr>
<tr>
<td></td>
<td>Music</td>
<td>WC</td>
<td>13</td>
<td>89.23</td>
<td>11.07</td>
<td>73.33</td>
<td>100</td>
<td>[82.54, 95.91]</td>
<td>-0.47</td>
<td>-1.48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR1</td>
<td>13</td>
<td>91.79</td>
<td>9.09</td>
<td>73.33</td>
<td>100</td>
<td>[86.30, 97.28]</td>
<td>-0.68</td>
<td>-0.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR2</td>
<td>13</td>
<td>89.23</td>
<td>10.38</td>
<td>66.67</td>
<td>100</td>
<td>[82.95, 95.50]</td>
<td>-0.61</td>
<td>0.13</td>
</tr>
<tr>
<td>MR</td>
<td>Desk</td>
<td>WC</td>
<td>13</td>
<td>33.85</td>
<td>28.44</td>
<td>0.00</td>
<td>90</td>
<td>[16.65, 51.03]</td>
<td>0.75</td>
<td>-0.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR1</td>
<td>13</td>
<td>43.08</td>
<td>28.10</td>
<td>10.00</td>
<td>90</td>
<td>[26.09, 60.05]</td>
<td>0.69</td>
<td>-0.70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR2</td>
<td>13</td>
<td>47.69</td>
<td>29.76</td>
<td>10.00</td>
<td>80</td>
<td>[29.70, 65.67]</td>
<td>-0.25</td>
<td>-1.69</td>
</tr>
<tr>
<td></td>
<td>Kitchen</td>
<td>WC</td>
<td>13</td>
<td>55.38</td>
<td>30.17</td>
<td>10.00</td>
<td>100</td>
<td>[37.15, 73.61]</td>
<td>0.35</td>
<td>-1.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR1</td>
<td>13</td>
<td>38.46</td>
<td>32.62</td>
<td>0.00</td>
<td>100</td>
<td>[18.74, 58.17]</td>
<td>0.83</td>
<td>-0.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR2</td>
<td>13</td>
<td>58.46</td>
<td>33.13</td>
<td>0.00</td>
<td>100</td>
<td>[38.44, 78.48]</td>
<td>-0.33</td>
<td>-1.08</td>
</tr>
<tr>
<td></td>
<td>Music</td>
<td>WC</td>
<td>13</td>
<td>39.23</td>
<td>27.53</td>
<td>0.00</td>
<td>100</td>
<td>[22.59, 55.86]</td>
<td>0.86</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR1</td>
<td>13</td>
<td>60.77</td>
<td>29.57</td>
<td>0.00</td>
<td>100</td>
<td>[42.90, 78.63]</td>
<td>-0.71</td>
<td>-0.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AR2</td>
<td>13</td>
<td>49.23</td>
<td>29.85</td>
<td>0.00</td>
<td>100</td>
<td>[31.19, 67.26]</td>
<td>0.33</td>
<td>-0.51</td>
</tr>
</tbody>
</table>

*Note.* Y/N = Yes/No items ($n = 15$ items per set); MR = meaning-recall items ($n = 10$ items per set).

**Yes/No GLMM Delayed Posttest Results**

The Yes/No GLMM random effects indicated that participants, $SD = 0.84$, were responsible for more variance than the items, $SD = 0.25$. The effect size for the Y/N model, conditional $R^2 = 0.23$, indicated that it explained approximately 23% of the variance in the data, with the fixed effects accounting for 5%,
$R^2 = 0.05$. The fixed effects for the Yes/No model indicated that there was no significant difference between Groups A and B, $\beta = 0.86, z = 2.04, OR = 2.36 [1.03, 5.40]$, and Groups A and C, $\beta = 0.49, z = 1.20, OR = 1.64 [0.73, 3.66]$. The lack of OR-CI overlap suggested that there was no significant difference between the two comparisons. There was also no significant difference displayed between the word sets, as attested to by the desk vs. kitchen, $\beta = -0.11, z = 0.40, OR = 0.90 [0.54, 1.51]$, and desk vs. music comparisons, $\beta = 0.02, z = 0.08, OR = 1.02 [0.61, 1.73]$. The fixed effects for the Yes/No model indicated also that there was no significant difference between Groups A and B, $\beta = 0.48, z = 0.91, OR = 1.61 [0.58, 4.49]$, and Groups A and C, $\beta = 0.80, z = 1.53, OR = 2.23 [0.80, 6.23]$.  

**Meaning-Recall GLMM Delayed Posttest Results**

The random effects for the meaning-recall delayed posttest results GLMM also indicated that participants accounted for more variance than the items, where spread for the persons, $SD = 1.26$, was greater than that for the items, $SD = 0.41$. The meaning-recall model explained 37% of the variance, $R^2 = 0.37$, with the fixed effects accounting for 4%, $R^2 = 0.04$. In contrast to the Yes/No data, the meaning-recall results indicated that scores on the kitchen-related words were slightly more difficult than the desk words, $\beta = 0.51, z = 2.03, OR = 1.66 [1.02, 2.70]$, although there was no significant difference observed for desk vs. music comparison, $\beta = 0.46, z = 1.84, OR = 1.58 [0.97, 2.58]$. These results indicate that both the groups and word sets were well matched to each other. Similar to Y/N GLMM, the majority of the variance in the vocabulary test scores was accounted for by participant and item differences, again with the treatment and word sets accounting for a smaller portion.

**Figure 3**

*Treatments Compared*
**Figure 4**

*Treatments Compared Per Set*

![Boxplot showing treatments compared per set](image)

**Table 4**

*GLMM Coefficients*

<table>
<thead>
<tr>
<th>Test</th>
<th>Fixed Effects</th>
<th>Random Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta$</td>
<td>SE</td>
</tr>
<tr>
<td>Y/N $^a$ (Intercept)</td>
<td>1.57</td>
<td>0.34</td>
</tr>
<tr>
<td>Group A: B</td>
<td>0.86</td>
<td>0.42</td>
</tr>
<tr>
<td>Group A: C</td>
<td>0.49</td>
<td>0.41</td>
</tr>
<tr>
<td>Treatment WC: AR1</td>
<td>0.66</td>
<td>0.24</td>
</tr>
<tr>
<td>Treatment WC: AR2</td>
<td>0.57</td>
<td>0.23</td>
</tr>
<tr>
<td>Set Desk: Kitchen</td>
<td>-0.11</td>
<td>0.26</td>
</tr>
<tr>
<td>Set Desk: Music</td>
<td>0.02</td>
<td>0.27</td>
</tr>
<tr>
<td>MR $^b$ (Intercept)</td>
<td>-1.13</td>
<td>0.42</td>
</tr>
<tr>
<td>Group A: B</td>
<td>0.48</td>
<td>0.52</td>
</tr>
<tr>
<td>Group A: C</td>
<td>0.80</td>
<td>0.52</td>
</tr>
<tr>
<td>Treatment WC: AR1</td>
<td>0.27</td>
<td>0.17</td>
</tr>
<tr>
<td>Treatment WC: AR2</td>
<td>0.49</td>
<td>0.17</td>
</tr>
<tr>
<td>Set Desk: Kitchen</td>
<td>0.51</td>
<td>0.25</td>
</tr>
<tr>
<td>Set Desk: Music</td>
<td>0.46</td>
<td>0.25</td>
</tr>
</tbody>
</table>

$a$ Model formula: Yes/No score $\sim$ (1|participant) + (1|item) + Group + Treatment + Set

$b$ Model formula: Meaning-recall score $\sim$ (1|participant) + (1|item) + Group + Treatment + Set
RQ 1: AR vs. WC

With respect to RQ1, which asked how results observed for the augmented reality conditions compared word card results, the GLMM results indicated that in most cases the AR treatments led to higher gains. The AR1 condition, which represented a context-independent real-world scene, resulted in significantly higher Yes/No scores than the word cards, $\beta = 0.66, z = 2.80, OR = 1.93 [1.22, 3.06]$, but not for the meaning-recall scores, $\beta = 0.27, z = 1.58, OR = 1.30 [0.94, 1.81]$. The effect of the AR2 condition, which represented a context-independent real-world scene, was slightly weaker than the AR1 effect when compared to the word cards for the Yes/No data, $\beta = 0.57, z = 2.45, OR = 1.76 [1.12, 2.77]$, but stronger for the meaning-recall data, $\beta = 0.49, z = 2.95, OR = 1.64 [1.18, 2.27]$. 

RQ 2: Post Hoc Comparison

In relation to RQ2, which asked how results observed for the context-independent real-world scene (AR1) compared with context-dependent real-world scene (AR2), post hoc comparison of the two treatments were conducted with the `emmeans` (Lenth, 2021) package for R. The results in Table 5 show that there was no significant difference between the two AR conditions for both the Yes/No, $\beta = 0.09, z = 0.36, OR = 1.09$, and meaning-recall results, $\beta = -0.23, z = -1.38, OR = 0.79$.

### Table 5

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Test</th>
<th>$\beta$</th>
<th>SE</th>
<th>z</th>
<th>p</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR1 vs AR2</td>
<td>Yes/No</td>
<td>0.09</td>
<td>0.25</td>
<td>0.36</td>
<td>.930</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>Meaning-recall</td>
<td>-0.23</td>
<td>0.17</td>
<td>-1.38</td>
<td>.353</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Discussion

This study compared and examined Yes/No and meaning-recall delayed posttest item data gathered from adult learners who deliberately studied concrete nouns in three conditions: study with paper-based word cards, study with a tablet-based AR application in a context-independent real-world VSB scene (AR1), and study with a tablet-based AR application in a context-dependent real-world VSB scene (AR2). In respect to our two research questions, this section discusses the implications of the findings and one confirmed hypothesis and offers insight into the limitations and future directions of this research.

Research Question 1

*How do tablet-based AR technology modes compare with paper-based word cards in terms of their ability to help a learner recall the meaning of concrete nouns?*

Both AR conditions outperformed word cards as measured by Yes/No test items on delayed posttests. Specifically, in Table 4 we can see significant differences between the word card and AR1 conditions, and between the word card and AR2 conditions. In respect to the delayed posttest meaning-recall items, statistically significant higher scores were achieved by participants who studied with AR2 as compared with word cards, though there were no statistically significant differences in the delayed posttest meaning-recall scores achieved by those who studied with AR1 as compared with word cards (see Figure 3 and Figure 4).

The Yes/No and meaning-recall test scores observed in the present study are reminiscent of Laufer and Goldstein’s hierarchy (2004). The authors reported a hierarchy of form-meaning link word knowledge from the easiest, meaning-recognition (learners are prompted with the L2 form and select associated L1
meaning), form-recognition (learners are prompted with the L1 meaning and select associated L2 form),
to the most difficult, meaning recall (L2 to L1 translation), and form recall (L1 to L2 translation). The
Yes/No format is outside of this format as it does not require that learners demonstrate knowledge of the
meaning of target L2 forms. However, McLean et al. (2020), found that among 103 Japanese EFL
learners, scores from a Yes/No test (62.1%) were lower than meaning-recognition scores (78.5%), and
also lower than meaning recall (42.3%) and form-recall (28.2%) for the same 1,000 words. Thus, it might
be concluded that the threshold of knowledge necessary to correctly answer a meaning-recall item is
greater than that necessary to select ‘Yes’ on a Yes/No test. This might explain why Pellicer-Sanchez
(2016) found 85% learners selected ‘Yes’ to knowing nonwords embedded six times in a reading text, but
only 55% were able to recall the correct meaning of the same nonwords. In accordance with this, the
results of the current study suggest that AR2 appears stronger than word cards for acquiring a deeper
knowledge of a word as demonstrated by participants’ ability to display a form-meaning link on the
meaning-recall format test questions. While the current study found this significant difference in AR2’s
ability to help form a deeper form-meaning link as compared with the word cards, research of AR in its
application in vocabulary acquisition is still within its infancy and this finding needs to be confirmed
within future replications.

Research Question 2

When concrete nouns are studied via AR, is there a difference as to whether those lexical items are
anchored to a contextually relevant scene arranged for utility or not?

The post hoc comparisons (see Table 5) of the results of the GLMM indicated that there were no
statistically significant differences between AR1 and AR2 study modes as measured by either delayed
posttest Yes/No items ($p = 0.930$) or by delayed posttest meaning-recall items ($p = 0.353$). These results
indicate that regardless of the physical arrangements of all AR1 and AR2 scenes (i.e., whether or not a
scene was arranged for a user’s utility), similar learning took place despite the differences presented by
the different theorized conditions. One possible explanation for this could be that a VSB effect was
present regardless of the theorized contextual-dependence or contextual-independence variations within
the AR1 and AR2 scenes. This appears to be consistent with the findings of Darling et al. (2020) which
showed that when recalling series of digits or nonwords, both familiar known spatial information and
novel spatial information contributes to the VSB effect. This might explain the lack of any statistically
significant difference between the AR1 and AR2 conditions whose theorized variations of contextual
dependence depended on spatial arrangement. It is also possible that the strength of any VSB effect might
also be leveraged by how personally meaningful to the participants the objects and the scenes were. In
Larchen Costuchen et al.‘s (2021) study, participants made use of their own personalized living
environments to place AR element triggers which participants scanned to study L2 idioms and phrases in
order to assess the effect of a theorized AR-VSB treatment (Larchen Costuchen et al., 2021). This raises
an important question worthy of further research: Is the VSB effect amplified when participants use
personalized environments for study versus environments such as those represented by the preconstructed
AR1 and AR2 scenes of the current study? Important personally regulated differences, such as how
personally meaningful, relevant, or familiar a participant finds certain objects and their environments,
might be unaccounted for factors worthy of investigation.

Regarding the lack of significant difference between the AR1 and AR2 conditions, there exists the
possibility that there might have been too many similarities between the theorized variations between the
AR1 and AR2 scenes; of the six different scene variations in the AR1 and AR2 conditions, apart from the
AR2 kitchen scene, the other five scenes were constructed within a classroom and made use of a table.
Furthermore, within the AR2 condition scenes, the AR2 desk and music scenes could be considered to be
more artificial than the AR2 kitchen scene which was embodied within an actual kitchen. In a future
replication of this study, the employment of AR2 scenes embodied by more realistic environments is
warranted. In this vein, opportunities exist for both qualitative and quantitative studies. Just how
important is the arrangement of items within a scene? Also, how much does personal familiarity with the
objects or a specific space play a role in the acquisition of vocabulary within those environments? Both are questions worthy of further investigation.

The novelty of the AR technology is yet another important factor to consider. Despite the AR application having been thoroughly introduced and explained to the participants via training exercises and orientation before the experiment was conducted, AR is a somewhat new and emerging technology and its use has not yet inundated commonly used pedagogical practices. Despite AR having become a somewhat commonly implemented technology in smartphones, tablets, smartphone games, and on some social media platforms, its novelty and attractiveness might present a possible bias when it is applied in novel pedagogical scenarios. Novelty of technology is one aspect which should be further investigated in future AR vocabulary studies. Regarding the current study, all the participants were observed to be focused on their assigned tasks in each round of treatment minimizing any possible disparity between the amount of focus given to items studied in any of the conditions.

Finally, both models indicated the majority of the variance was caused by differences between the participants (random effects). In the current study, though mostly populated with participants who spoke English or Japanese as an L1, participants with other L1s were also included in this sample. Because all these participants were highly proficient in English and this experiment focused on the learning of nonwords, some of the participants’ L1s were not represented on the word cards or AR labels and word cards. The L1 lexis of the participants might be responsible for some of this variance between the participants. In a future study, regardless of if nonwords are used, the L1 of the participants should be more tightly controlled to help eliminate this possible source of variation. For instance, a homogenous group of learners with the same L1 might counteract this potential confound. Despite this, it should be noted that when these data were subjected to Rasch analyses in a separate study with the goal of collecting validation evidence for the posttests, four separate DIF were conducted to determine whether certain items were unfairly rewarding participants based on four different membership classifications: gender differences, differences based on the treatment modes, differences based on having an L1 different from one of the languages of the target-word meaning available on the word cards or AR labels, and differences based on total L1 variations of the participants (i.e., English / Japanese / Other). The DIF analyses results indicated that no items or questions unfairly rewarded participants on the basis of any of these membership classes (see Dabrowski, 2022).

Limitations and Future Directions

As mentioned previously, regarding the AR2 scenes, only the desk and kitchen scenes used actual embodied locations (classroom desk spaces and a kitchen) for the placement of the items which embodied the target words whereas the music scene was constructed to display utility by arranging the items in a somewhat artificial manner as a real-world music studio was not available to the researchers. In light of this limitation, the AR2 scene for Set 3 was created in a way that used the objects which embodied the concrete nouns of the scene in an arrangement that was theorized to cultivate practical utility of a real-world music studio (his scene can be referenced in Appendix B). Despite this limitation and the somewhat artificial nature of this scene, it is of interest that learners who studied the sets of words within the AR2 scenes were rewarded in terms of their ability to recall meaning as compared with the word-card study condition, whereas the AR1 scenes which are arguably more artificially arranged than the AR2 music scene only rivaled the word-card study condition significantly on the Yes/No test questions. This might indicate that if this experiment were replicated, and instead of this artificial AR2 scene, a real-world music studio were used in place of these scenes, perhaps a stronger effect could be observed. It seems to also indicate that the arrangement of the items into a scene which offered practical utility benefited the participants significantly in their recall of meaning as demonstrated by the Yes/No and meaning-recall test items of the delayed posttest regardless of the more artificial nature of the AR2 music scene.

The word cards and the AR application labels used presentation-style labels to display the words and their meanings to the learners. In this format the target word and its meaning are both visually accessible at once. Presentation style was argued to be less superior to methods of vocabulary study which use retrieval
modes, such as double-sided word cards, because the nature of retrieval modes in study forces the learner to retrieve knowledge, thus making the connection between form and meaning stronger (Barcroft, 2007; McNamara & Healy, 1995; Nakata 2011; Royer, 1973). The decision to use presentation style labels was made to control and match the word cards to the limitation of the ARIO application, which was not reasonably capable of creating and displaying double-sided word cards. Despite this limitation, the participants who studied with AR made statistically greater gains as compared to those who used the word cards. A future study which compares a paper-based retrieval mode and an AR-based retrieval mode is warranted. Future experiments could also expand the length of the treatment and the number of treatments. Finally, while this study focused exclusively on the use of an AR-VSB treatment method to deliberately study concrete nouns, this treatment method could be examined in future studies to focus on comparisons of concrete nouns with abstract nouns, verbs, other parts of speech, and other lexical attributes to better understand the role of the VSB effect in AR-based vocabulary acquisition.

Though a randomized stratified sampling strategy was used in the construction of the groups within the Latin square design, the sample was drawn partially from an intact PhD cohort and a snowball sampling strategy was used to recruit the other participants. While the sample in this study represents adult learners, the experiment should be conducted again with a random sample drawn from a different or broader population. More specifically replications of this experiment should more closely control for the L1 of the participants. Though such variation was found not to present any confounds in the present study, it introduced an unnecessary complexity into our design. Controlling for this in future studies might contribute to a reduction in the noise associated with the random factors, potentially making it easier to observe the fixed effects.

One aspect to be further explored is spatiality, or the intentional and systematic incorporation of spatial variability in vocabulary study and other areas of SLA. The current study shed some light on how AR technology compares with word cards in its ability to help learners recall the meaning of concrete nouns and whether variations of contextual arrangements of the AR scenes caused any differences in retention. Though we were able to provide answers for both of our questions, a variety of different yet related study conditions not attended to in the current study could be used to probe further into spatial information, how it supports learning, and the mechanisms underlying the positive performance advantage observed in the AR treatment modes. Word cards could be placed into familiar grids or arrays (similar to those arrays present in the work of Darling et al., 2020) to further probe the VSB effect in spatialized vocabulary learning. Furthermore, vocabulary learning in an AR-VSB condition could be compared with a similar real-world scene populated with paper-based labels, such as Post-it notes, could help to isolate and more closely approach observing spatial and other factors that might underpin the effectiveness of AR. However, a major advantage of AR technology over paper-based labels is the ability to virtually annotate and gloss the public world around us, hinting at possible future pedagogical uses of AR that paper-based study methods may be unable to rival.

**Conclusion**

This study can conclude with tentative confidence that both AR modes were effective in helping learners deliberately study and learn the meaning of the target words. According to the data collected from this experiment, AR appears to be as effective, if not more so, than presentation-style word cards in leveraging recognition of word form and meaning as well as the recall of word meaning; more word meanings were accurately recalled and more word forms and meanings were accurately recognized on one-week delayed posttests when AR was the mode of study. The arrangement and contextual significance of the AR scenes in this study appeared not to offer any additional leverage in the recall or recognition of the words and meanings studied.

Finally, the indication that a tablet-based AR application was effective in helping learners make a lasting form-meaning link with the words they studied is a finding that reinforces the call for more research of AR technology as applied in deliberate vocabulary study contexts and in wider language learning contexts. While expensive head-mounted AR devices are no doubt attractive for research possibilities,
this study shows that tablets and smartphones readily available at a fraction of the cost can be used for similar research and pedagogical applications.

Acknowledgements

We would like to acknowledge and thank the participants for their time. We would also like to thank the anonymous reviewers who helped to improve this manuscript. Finally, we would like to thank David Beglar, Nicholas Carr, Tomoko Nemoto, and James Sick for their valuable input on the design of this project.

References


Dabrowski, A. (2023). A case study in tablet-based augmented reality vocabulary acquisition. *International Journal on Open and Distance E-Learning, 8*(2). https://doi.org/10.58887/ijodel.v8i2.96


Appendix A. Word Sets

Table A1

Word Set 1: Desk

<table>
<thead>
<tr>
<th>Nonword</th>
<th>Japanese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>bload</td>
<td>ペン</td>
<td>pen</td>
</tr>
<tr>
<td>dudge</td>
<td>本</td>
<td>book</td>
</tr>
<tr>
<td>foose</td>
<td>パソコン</td>
<td>computer</td>
</tr>
<tr>
<td>grink</td>
<td>紙</td>
<td>paper</td>
</tr>
<tr>
<td>mence</td>
<td>テープ</td>
<td>tape</td>
</tr>
<tr>
<td>shoat</td>
<td>眼鏡</td>
<td>glasses</td>
</tr>
<tr>
<td>creet</td>
<td>ハサミ</td>
<td>scissors</td>
</tr>
<tr>
<td>slank</td>
<td>ポーチ</td>
<td>pouch</td>
</tr>
<tr>
<td>prome</td>
<td>シーディー</td>
<td>CD</td>
</tr>
<tr>
<td>zight</td>
<td>スマホ</td>
<td>smartphone</td>
</tr>
</tbody>
</table>

Table A2

Word Set 2: Kitchen

<table>
<thead>
<tr>
<th>Nonword</th>
<th>Japanese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>clush</td>
<td>洗剤</td>
<td>soap</td>
</tr>
<tr>
<td>brear</td>
<td>スポンジ</td>
<td>sponge</td>
</tr>
<tr>
<td>foint</td>
<td>お箸</td>
<td>chopsticks</td>
</tr>
<tr>
<td>plail</td>
<td>フライパン</td>
<td>pan</td>
</tr>
<tr>
<td>lorch</td>
<td>皿</td>
<td>plate</td>
</tr>
<tr>
<td>chame</td>
<td>カップ</td>
<td>cup</td>
</tr>
<tr>
<td>smole</td>
<td>フォーク</td>
<td>fork</td>
</tr>
<tr>
<td>stape</td>
<td>お茶</td>
<td>tea</td>
</tr>
<tr>
<td>stook</td>
<td>チップス</td>
<td>chips</td>
</tr>
<tr>
<td>zound</td>
<td>コカコーラ</td>
<td>coke</td>
</tr>
</tbody>
</table>
### Table A3

*Word Set 3: Music*

<table>
<thead>
<tr>
<th>Nonword</th>
<th>Japanese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>boach</td>
<td>ピアノ</td>
<td>piano</td>
</tr>
<tr>
<td>crown</td>
<td>ギター</td>
<td>guitar</td>
</tr>
<tr>
<td>fronk</td>
<td>音楽</td>
<td>music</td>
</tr>
<tr>
<td>plint</td>
<td>メトロノーム</td>
<td>metronome</td>
</tr>
<tr>
<td>prane</td>
<td>コード</td>
<td>cord</td>
</tr>
<tr>
<td>satch</td>
<td>アイポッド</td>
<td>iPod</td>
</tr>
<tr>
<td>flont</td>
<td>ピック</td>
<td>pick</td>
</tr>
<tr>
<td>slunt</td>
<td>アンプ</td>
<td>amplifier</td>
</tr>
<tr>
<td>drack</td>
<td>コーヒー</td>
<td>coffee</td>
</tr>
<tr>
<td>tudge</td>
<td>ケース</td>
<td>case</td>
</tr>
</tbody>
</table>
Appendix B. Set Documentation

Figure B1

Set 1 Word Cards

Figure B2

Set 2 Word Cards
Figure B3

Set 3 Word Cards

Figure B4

Set 1 AR1 Scene
Figure B5

Set 1 AR1 Scene with AR Labels

Figure B6

Set 2 AR1 Scene
Figure B7

Set 2 AR1 Scene with AR Labels

Figure B8

Set 3 AR1 Scene
Figure B9

Set 3 AR1 Scene with AR Labels

Figure B10

Set 1 AR2 Scene
Figure B11

Set 1 AR2 Scene with AR Labels

Figure B12

Set 2 AR2 Scene
Figure B13

Set 2 AR2 Scene with AR Labels

Figure B14

Set 3 AR2 Scene
Figure B15

Set 2 AR2 Scene with AR Labels
Appendix C. Assessments: Google Forms Links

Set 1 Pretest: https://forms.gle/nSuNP8hsi8VXGAAa7
Set 2 Pretest: https://forms.gle/BCQYCnu4dxJNSe187
Set 3 Pretest: https://forms.gle/TD67Jas8L56H9LmLA
Set 1 Posttest: https://forms.gle/pG6XqquAkD6Hmz5XA
Set 2 Posttest: https://forms.gle/R5b8wSKtHXjSZHLD9
Set 3 Posttest: https://forms.gle/CDaQvHDxAYZZACUr9
Set 1 Delayed Posttest: https://forms.gle/SoNZbJbtgftgQRXW9
Set 2 Delayed Posttest: https://forms.gle/KHbmwBmd7JuqSKiw7
Set 3 Delayed Posttest: https://forms.gle/8W7pvS4XKGbkYEoW7
Appendix D. Supplementary Materials

Please see the link below to view:
1) Supplementary Materials A: Variables Considered
2) Supplementary Materials B: GLMM Model Building and Assumption Checks
   https://osf.io/g8n5b/

About the Authors

Adam Dabrowski is a PhD candidate in applied linguistics at Temple University, Japan. He teaches as a lecturer at The University of Electro-Communications located in Tokyo, Japan. His research interests include augmented reality in language education, vocabulary acquisition, and embodied cognition. Adam Dabrowski is the corresponding author.
   E-mail: Dabrowskiowski@gmail.com
   ORCiD: https://orcid.org/0000-0003-1958-9823

Stuart McLean (TESOL and medicine PhDs) teaches at Kindai University and Temple University, Japan. He is making online self-marking form-recall and meaning-recall (orthographic and phonological) vocabulary levels tests. Automatically marked responses, actually typed responses, and the time taken to complete responses are downloadable (vlt.carleton.ca).
   See Stuart McLean’s Google Scholar page for details of his research.
   E-mail: stumc93@gmail.com
   ORCiD: https://orcid.org/0000-0002-7035-378X

Christopher Nicklin is a Project Assistant Professor in the Center for Global Education at the University of Tokyo, and recently completed his PhD in applied linguistics at Temple University, Japan. His research interests focus on vocabulary, particularly in the subdomains of psycholinguistics and research methodology.
   E-mail: christopher.nicklin79@gmail.com
   ORCiD: https://orcid.org/0000-0002-8945-0678