P-ISSN 2355-2794
E-ISSN 2461-0275

Comparing the Effectiveness of Multimodal vs Monomodal Digital Flashcards for L2 Vocabulary Learning

Joshua Hicks*
Rina Marnita
Oktavianus Oktavianus

Department of Linguistics, Faculty of Humanities, Universitas Andalas, Padang 25163, INDONESIA

Abstract

This applied psycholinguistics study explores whether multimodal flashcards (containing text, audio, and a picture) are more effective than monomodal flashcards (containing text only) as a tool for learning the meanings of novel second-language (L2) concrete nouns. The research instrument was Anki, a flashcard application that utilises active recall and spaced repetition. The study used a within-subject design, where each participant (n = 25) studied a total of 30 L2-L1 (Esperanto–Indonesian) word pairs over the course of seven study sessions utilising an assortment of 15 multimodal and 15 monomodal flashcards, with each word pair being presented multimodally to approximately half of the participants and monomodally to the other half. When (re)viewing the answer side of a card, participants were instructed to tap 'Good' if they recalled the answer correctly or 'Again' if not. Recall accuracy data for the two card types were collected and then analysed using a Wilcoxon signed-rank test, which indicated that the number of user-initiated reviews ('Again' count, which is indicative of the number of memory lapses) was significantly higher for monomodal flashcards (Mdn = 61, n = 25) than for multimodal flashcards (Mdn = 50, n = 25), Z = -3.4, p < 0.001, r = -0.7. These results support the hypothesis that multimodal flashcards are more effective than monomodal flashcards as a tool for learning the meanings of L2 concrete nouns. By implication, language learners can enhance their recall accuracy of L2 concrete nouns by creating and using flashcards that utilise multiple semantically congruent modes.

Keywords: Dual-encoding, multimodal, multisensory, recall, vocabulary.

* Corresponding author, email: joshua.academia@icloud.com

Citation in APA style: Hicks, J., Marnita, R., & Oktavianus, O. (2025). Comparing the effectiveness of multimodal vs monomodal digital flashcards for L2 vocabulary learning. *Studies in English Language and Education, 12*(3), 1133-1152.

Received June 26, 2024; Revised October 30, 2024; Accepted August 6, 2025; Published Online September 30, 2025

https://doi.org/10.24815/siele.v12i3.39630

1. INTRODUCTION

Digital flashcards have been shown to be an effective tool for L2 vocabulary learning in a range of studies (Bakla & Çekiç, 2017; Mujahidah et al., 2024; Nguyen, 2021). However, these studies do not test the effectiveness of different types of flashcard design, and while some studies have demonstrated that multimodal (or multimedia or multisensory) learning can be more effective than monomodal learning (Chun & Plass, 1996; K. M. Mayer et al., 2015; R. E. Mayer, 2002; Moran et al., 2013; Okray et al., 2023; Shams & Seitz, 2008; Thelen & Murray, 2013), there are no studies that specifically test the effectiveness of using multimodal versus monomodal digital flashcards for L2 vocabulary learning (Google Scholar as on 24-10-2024). Flashcard applications such as Anki offer a myriad of design possibilities; users are not limited to text only, but they may create multimodal flashcards that include text, images, audio, and video. The inclusion of pictures and audio in addition to text is often recommended and practised by those who use Anki for L2 vocabulary learning (Refold, n.d.). This practice can be supported by cognitive theories such as Dual Coding Theory (Clark & Paivio, 1991), Cognitive Theory of Multimedia Learning (R. E. Mayer, 2002), and by applying insights gleaned from studies in multisensory learning to L2 vocabulary learning (see Moran et al., 2013; Okray et al., 2023; Shams & Seitz, 2008). However, it should be acknowledged that neither the application of theory to practice nor the application of pedagogical principles gleaned from one kind of education (monolingual instruction) to another kind of education (second language learning) will necessarily result in the intended enhancement of learning. Therefore, this study aims to empirically test one particular application of theory to practice, namely, to compare the effectiveness of multimodal and monomodal Anki flashcards as a tool for learning the meanings of L2 concrete nouns. The use of the free flashcard application Anki as a research instrument means that the research design reflects a real-world L2 vocabulary learning method that may be freely replicated by learners, teachers, and researchers alike. The research questions of this study are as follows:

1. Are multimodal flashcards (containing text, audio, and a picture) significantly more effective than monomodal flashcards (containing text only) as a tool for learning the meanings of L2 concrete nouns (i.e. resulting in significantly higher recall accuracy)?
2. Does learning L2 vocabulary multimodally result in better recall accuracy, even in response to monomodal (text-only) test cues?
3. If multimodal learning is shown to be more effective than monomodal learning in this study, why is this the case?

Based on Dual Coding Theory and supporting evidence, it was predicted that multimodal flashcards would result in significantly higher recall accuracy of L2 vocabulary compared to monomodal flashcards. To test this prediction, the present study was designed to test the null hypothesis (H0) and alternative hypothesis (H1). Empirical data were collected, and the statistical hypothesis was tested by performing a one-tailed Wilcoxon signed-rank test on the raw data.

Null hypothesis (H0)	: Multimodal flashcards are not significantly more effective than monomodal flashcards as a tool for learning the meanings of L2 concrete nouns.
Alternative hypothesis (H1)	: Multimodal flashcards are significantly more effective than monomodal flashcards as a tool for learning the meanings of L2 concrete nouns.

2. LITERATURE REVIEW

2.1 Multimodal vs. Monomodal L2 Vocabulary Learning

Lin and Yu (2017, p. 9) compared the effectiveness of monomodal (text only) and multimodal (text + audio + picture) presentation types for English vocabulary learning via

multimedia message (MMS). An analysis of recall accuracy data from an immediate post-test showed no significant effect of presentation type. However, an analysis of recall accuracy data from a delayed post-test (two weeks after vocabulary learning) indicated that recall accuracy was significantly higher for vocabulary that had been presented multimodally compared to vocabulary that had been presented monomodally. Similarly, in a study by K. M. Mayer et al. (2015), an analysis of immediate post-test data showed no significant difference in the recall accuracy of vocabulary that had been learned monomodally (audio-only) and multimodally (audio + picture; audio + gesture). However, analyses of results from delayed post-tests (two months and six months after learning) indicated that recall accuracy was significantly higher for multimodally learned words than for monomodally learned words.

The results of these studies suggest that the benefits of multimodal vocabulary learning are best observed in a delayed post-test, not in an immediate post-test. One possible explanation is that the advantage of multimodal learning over monomodal learning may only become apparent once learning has been sufficiently consolidated, e.g., through repeated spaced retrieval. The consolidation of multimodal learning would mean the formation and strengthening of an interconnected network of mental representations corresponding to the multiple modes used, enabling retrieval to operate on a richer, more informative network of representations, thus improving recall accuracy (see Moran et al., 2013). In the current study, Anki study sessions provide learners with a built-in opportunity for repeated spaced retrieval (active recall testing), which consolidates learning. In addition, rather than using an immediate post-test, recall accuracy data from the entire study phase (seven study sessions) were collected and analysed, followed by data from monomodal (text-only) delayed post-tests.

2.2 Learning L2 Vocabulary from Pictures vs. L1 Translations

Carpenter and Olson (2012) explored whether novel L2 concrete nouns are learned better by being paired with pictures or L1 translations. Carpenter and Olson's (2012, p. 95) first experiment replicated the pattern of results reported by Lotto and de Groot (1998) in that there was no advantage in cued recall of L2 words from pictures compared with L1 translations. However, when they asked the participants to verbally free recall in L1 the pictures presented vs. the L1 translations, the picture superiority effect was present in that participants were able to recall more pictures than the L1 translations. Therefore, Carpenter and Olson (2012, p. 99) concluded that the picture itself had been sufficiently encoded but that participants had failed to establish a sufficient association between the picture and the L2 word.

It stands to reason that once a sufficient association between the picture and the L2 word has been established, recall accuracy of L2 words learnt from pictures may be greater than recall accuracy of L2 words learnt from L1 translations. This is evident in Carpenter and Olson's (2012, p. 96) second experiment, which involved three tests with immediate feedback; these tests would have served as additional opportunities for spaced retrieval, strengthening the association between the L2 vocabulary item and the picture or L1 translation. As in Experiment one, no significant advantage emerged for picture–L2 pairs over L1 translation–L2 pairs in Test 1; however, this advantage was apparent in Tests 2 and 3 ($ts > 3.23$, $ps < .005$), and a repeated-measures ANOVA revealed that this interaction was significant by participants as well as by items.

In the current study, pictures are included on multimodal cards in addition to L1 translations since these two modes can enhance and clarify each other. Additionally, repeated spaced retrieval is integrated into Anki study sessions to help participants establish a sufficient association between the L2 word and other information on the card (e.g., picture, L2 audio, and L1 word). This consolidation of learning through repeated spaced repetition is an important part of the study since the results of other studies suggest that the advantage of multimodal L2 vocabulary learning can only be observed once learning has been sufficiently consolidated.

2.3 The More Modes the Merrier?

In a study by Li et al. (2022), participants presented with two verbal modes (text + audio) performed better than those presented with four modes (text + audio + picture + video) in the two post-tests. This suggests that presenting more modes does not necessarily improve learning outcomes. Li et al. (2022) noted that a possible explanation for these results is that the four-mode presentation slides forced participants to handle additional visual information within a limited time, which increased their cognitive load, negatively impacting learning outcomes. This explanation draws on the Limited Capacity Assumption of Mayer's Cognitive Theory of Multimedia Learning (CTML), which states that each channel in the human cognitive system has limited processing capacity; as a consequence, presenting too much visual information at once can overload the visual-pictorial channel and presenting too auditory information at once can overload the auditory-verbal channel (R. E. Mayer, 2002).

A major contribution of Mayer's Cognitive Theory of Multimedia Learning (CTML) is that Mayer takes into account John Sweller's Cognitive Load Theory when investigating the optimal conditions in which dual-coding can occur, and he developed principles to guide educators in the most effective use of multimedia for learning (R. E. Mayer, 2002). On the one hand, presenting in multiple modes has the potential to facilitate dual-encoding, which benefits recall. However, presenting in multiple modes can result in cognitive overload, hindering encoding and negatively impacting recall. Multimodal teaching and learning should therefore be done in a way that both maximises the chances of dual-encoding while managing the risk of cognitive overload by reducing unnecessary cognitive load. To this end, Mayer developed several principles of multimedia design. These principles were designed for and applied to the use of explanatory animations (words + moving pictures), but they are also relevant for multimedia L2 vocabulary teaching and learning. Mayer's principles of multimedia design have informed the design of the multimodal flashcards used in this study.

2.4 Dual Coding Theory

According to Paivio's Dual Coding Theory (DCT), the mind uses two distinct types of mental representation or "code"; verbal representations in the Verbal System (V) correspond to linguistic stimuli, and non-verbal representations (imagens) in the Image System (I) correspond to non-linguistic stimuli (Clark & Paivio, 1991, p. 152; Paivio & Csapo, 1973, p. 177). While monolinguals have one verbal system (V) and one image system (I), bilinguals or language learners have two verbal systems (V1 and V2) corresponding to two languages (L1 and L2) plus one shared image system (I) (Paivio & Desrochers, 1980, pp. 390-391).

The independence and partial interconnectedness of each system means that one code can be transformed into another, meaning – for example – that pictures can be named, and words can evoke nonverbal images (Paivio & Csapo, 1973, p. 178). In the context of second-language learning, this means that L2 words can be pictured (using V2→I) or translated into L1 (using V2→V1). In addition, if one representation or connection within a representational network decays (i.e., becomes unviable), the independence and partial interconnectedness of each symbolic system means that the rest of the network remains functional and may even be able to retrieve the required information by means of other representations and connections in the network. For example, if a V2–V1 connection is unviable, the image system can provide a means of indirect access from one language to another, enabling a person to translate from L2 to L1 by means of the image system (V2 → I → V1) (Paivio & Desrochers, 1980, p. 391). According to DCT, using both verbal stimuli (e.g., auditory words) and non-verbal stimuli (e.g., pictures) in teaching and learning facilitates the building of connections between the verbal and nonverbal systems (i.e., dual-encoding), resulting in a larger number of possible retrieval routes, which can have an additive effect on recall (Paivio & Csapo, 1973, p. 178). A 'Verbal Only' monomodal learning method (L2 + L1) facilitates the formation of V2–V1 connections only (, left), whereas a 'Three System' multimodal learning method (L2 + Picture + L1) facilitates the formation, activation, and consolidation of connections between all three systems, resulting in a larger

number of possible retrieval routes, which can have an additive effect on recall (, right). The key implication for L2 vocabulary teaching and learning is summed up in what Nation (2013, p. 467) calls the dual-encoding principle; having both linguistic and non-linguistic (e.g., pictorial) associations for a word aids word retention.

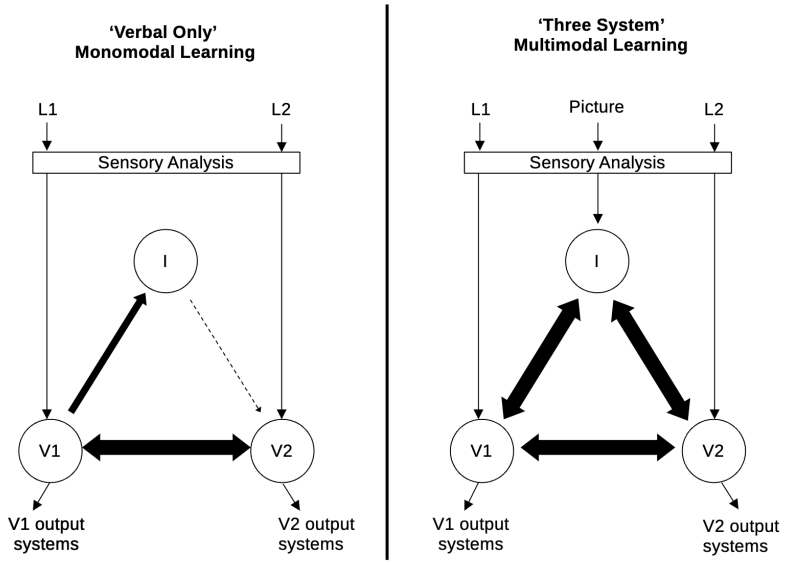

Figure 1. Between-system connections resulting from 'Verbal only' monomodal language learning (left) and 'Three System' multimodal language learning (right) (based on Paivio & Desrochers, 1980, p. 391).

In addition to between-system connections, connections also exist between subsystems that correspond to different sensory modalities (e.g., visual, auditory, motor) (Clark & Paivio, 1991, p. 153). These subsystems are capable of functioning more-or-less independently of one another, as evidenced by the selective effects of focal brain injuries which might impair one subsystem while leaving others functionally intact (Paivio, 1986, p. 57). For example, visual memory of the shapes of words (stored in the verbal-visual subsystem) may be impaired in an adult with brain injury, while motor memory of the shapes of words (the verbal-motor subsystem) remains unimpaired, allowing the patient to decode the meaning of words by tracing letters with his finger (see Carreker & Birsh, 2018, p. 101). Since each subsystem is more-or-less independent, Paivio and Csapo's (1973, p. 178) claim that "the two codes can have additive effects on recall" may be expanded to suggest that, in addition to this, interconnections between subsystems can also have an additive effect on recall. The implication of this for the current study is that the inclusion of both visual and auditory words (facilitating the formation, activation and consolidation of connections between verbal-visual and verbal-auditory subsystems) may have an additive effect on recall.

3. METHODS

The present study employed a quantitative, within-subjects design to investigate the effectiveness of multimodal flashcards (containing text, audio, and a picture) compared to monomodal flashcards (containing text only) for learning second-language (L2) vocabulary. The Esperanto language was selected as the target language to ensure all participants began with no prior knowledge of the target vocabulary. Recall accuracy data was collected during two distinct phases: the Study Phase (across seven sessions) and the Test Phase (which includes three delayed post-tests). Data collection was performed remotely via the Anki application, which tracked participants' recall accuracy. The subsequent subsections detail the research design, including the

participant recruitment process, the technique of data collection and analysis, and the research instrument.

3.1 Participants

A total of 38 participants were initially recruited to participate in the study; of these, 25 participants completed the study by completing seven study sessions with built-in active recall testing (the study phase) followed by three delayed post-tests (the test phase). Meanwhile, 13 recruited participants failed to complete the study and therefore their data were excluded from analysis. The eligibility criteria for participation in the study were as follows:

1. Nationality: Indonesian
2. Language: Can speak Indonesian
3. Experience: Has never studied the Esperanto language
4. Age: ≥17 years old
5. Device: Owns an Android phone that can install AnkiDroid

These criteria also describe the target population of the study. The accessible population was, however, much more specific. The researcher had access to two participant pools, namely:

1. First-semester English Literature students at Andalas University, Padang, Indonesia.
2. Members of Sunset English Club and their friends (a free weekly club at *1 Nusantara Cafe* where people can informally learn English and practice speaking English with others).

The sampling method used can be described as convenience sampling, one type of non-probability sampling in which participants are recruited based on their availability and willingness to participate (Suen et al., 2014). Of the final group of 25 participants, 10 were from the first participant pool and 15 from the second participant pool. The following variation was present in the sample:

1. First Language (or mother tongue): *Bahasa Minangkabau* (14 participants), *Bahasa Indonesia* (11 participants)
2. Occupation: Students (17 participants), working (8 participants)
3. Age range: 17–41 years old

Since many participants are bilingual, in this paper the abbreviation L1 (first language) is used in a non-technical sense to refer to a known language to which participants were exposed from childhood and in which participants are already fully communicatively competent (i.e., Indonesian), regardless of whether the language was acquired 'first' or acquired simultaneously with another language. The variation present in the sample – including variation not measured, such as language learning ability and working memory capacity – was not expected to affect the outcome of the study since this study uses a within-subject design, i.e., each participant's performance was compared to his / her own on L2 words learned using monomodal flashcards and L2 words learned using multimodal flashcards.

3.2 Technique of Data Collection and Analysis

To answer research question one, participants (described in section 3.1) were asked to learn a total of 30 Esperanto–Indonesian word pairs (Appendix B) over the course of seven study sessions, presented as an assortment of 15 multimodal and 15 monomodal Anki cards (card outlines are shown in Figures 2 and 3, example cards shown in Figures 2 and 5, the Anki application is described in section 3.4, and the Anki settings used are listed in Appendix A). During the Study Phase, participants' receptive retrieval (recalling L1 in response to an L2 test cue) for multimodal and monomodal cards was repeatedly tested using multimodal test cues (L2 text and L2 audio) and monomodal test cues (L2 text only), respectively. Participants were instructed to attempt to recall the answer side of the card (L1) when presented with the question side of a card (L2 test cue), and then to tap 'Reveal Answer' (*Tampilkan Jawaban*) to check their answer, build an

association between the different elements of the card, and provide feedback regarding their recall accuracy. Participants were instructed to tap 'Good' (*Baik*) if they had recalled the answer correctly or 'Again' (*Ulang*) if not, which provided the recall accuracy data for analysis. Participants' 'Again' count for multimodal and monomodal cards was collected and analysed using a one-tailed Wilcoxon signed-rank test to determine whether multimodal cards (containing text, audio, and a picture) are significantly more effective than monomodal flashcards (containing text only) as a tool for learning the meanings of L2 concrete nouns, (i.e. resulting in significantly higher recall accuracy). The Wilcoxon signed-rank test – a non-parametric paired test – was selected because a Shapiro-Wilk test indicated that the assumption of normality for the differences between the two dependent samples was violated in three of the four data sets analysed (Scheff, 2016). With a sample size of n=25, this violation makes a non-parametric paired test the appropriate choice.

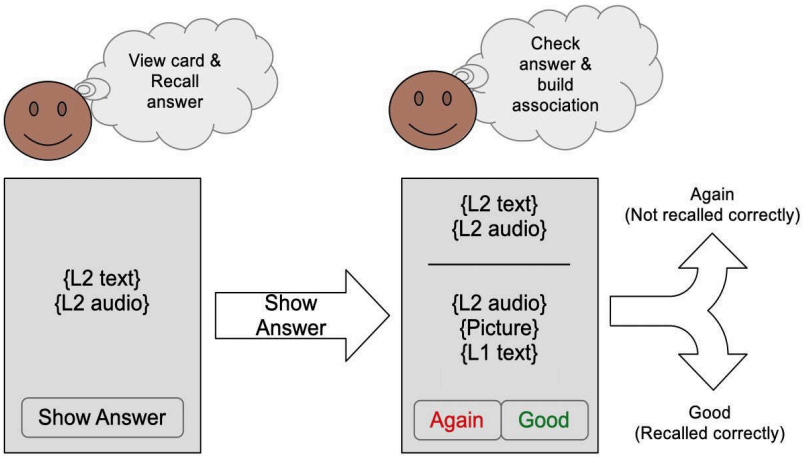

Figure 2. Multimodal flashcard outline; recall testing (left), building associations (right).

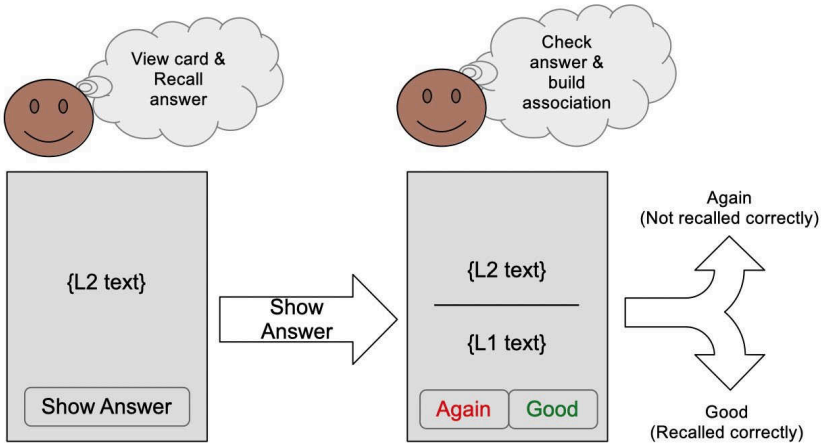

Figure 3. Monomodal flashcard outline; recall testing (left), building association (right).

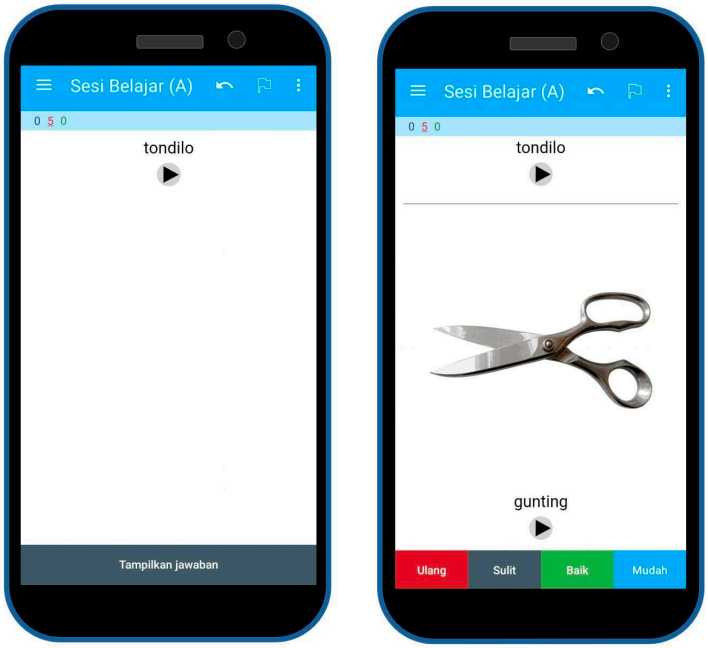

Figure 4. A multimodal Anki card for *tondilo – gunting* 'scissors', question side for recall testing (left), answer side for building associations (right).

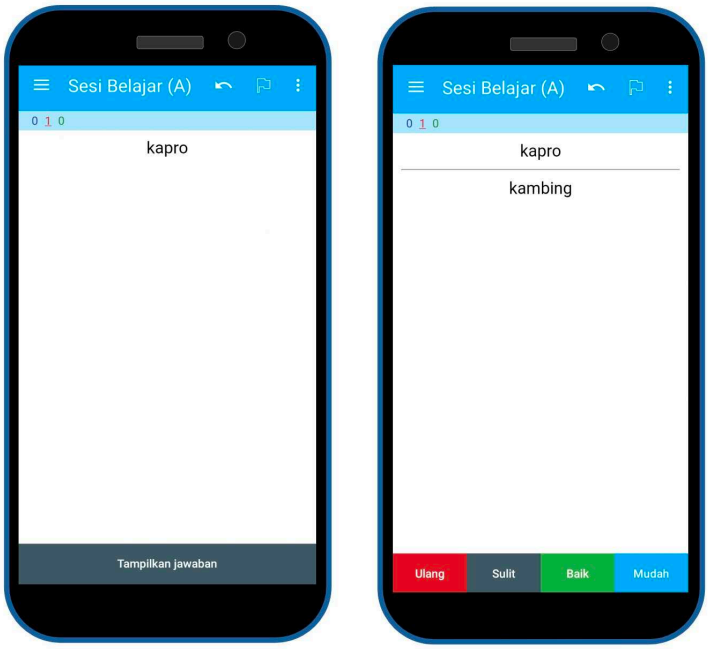

Figure 5. A monomodal Anki card for *kapro – kambing* 'goat'; question side for recall testing (left), answer side for building associations (right).

During the Study Phase, cards marked 'Again' (i.e. did not recall correctly) were shown again after a short interval within the same study session (waiting in the 'learning queue' for that study session), whereas cards marked 'Good' (i.e. recalled correctly) were scheduled for the next day unless the card was only on the first learning step (see Figures 6 and 7). This repeated recall testing is key to Anki's effectiveness as a learning tool. Research has shown that it is primarily the number of test episodes (spaced retrievals), not the number of study episodes, that determines retention, as demonstrated by Karpicke and Roediger (2008), who found that increasing the number study episodes for learning foreign vocabulary words had little effect on retention (33 →

36%), whereas increasing the number of test episodes increased retention significantly (33 → 81%).

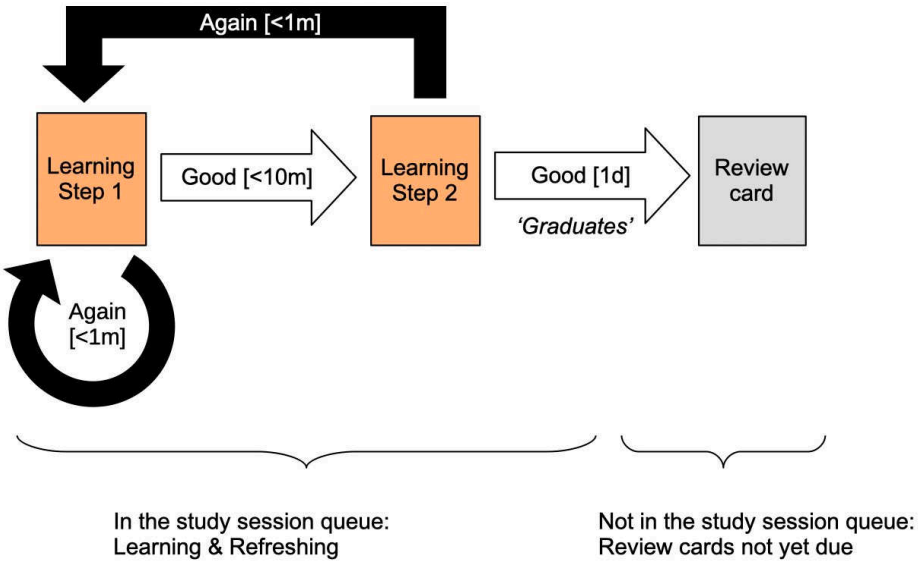

Figure 6. A New of Learning Step 1 Card flow chart.

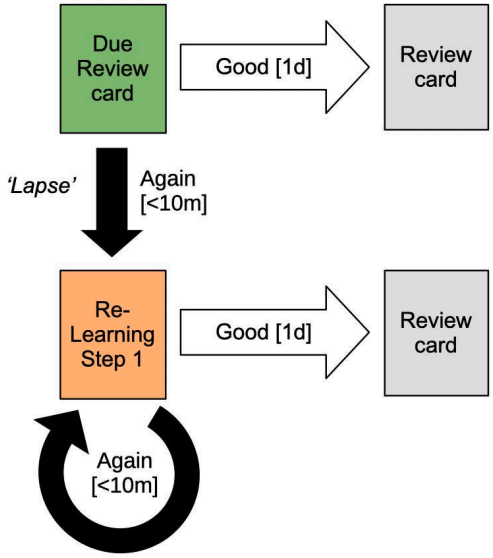

Figure 7. A Due Review Card flow chart.

The Study Phase consisted of seven study sessions. Sessions one to five introduced six new items per session. Each session presented items for review that had been learned in previous sessions. As a result of this design, sessions six and seven allowed the participants to consolidate their learning of all 30 items without any new words being introduced. The participants were instructed to complete one study session per day, meaning that the study phase would be completed over the course of seven days. However, some participants failed to complete one study session each day, so the participants completed the study sessions over the course of 7-14 days.

To answer research question two, during a subsequent Test Phase, participants' recall accuracy in response to only monomodal test cues (L2 text) was tested for all 30-word pairs in each of three delayed post-tests. In other words, monomodal Anki cards (Figure 5) were used to test participants' recall of all word pairs regardless of whether the word pair was initially learnt

using a monomodal or multimodal card in the Study Phase. Post-tests were carried out one, three, and seven days after the Study Phase. The use of increasingly larger intervals between post-tests in the Test Phase was intended to progressively increase the chances that participants would forget what they had learnt (c.f. Boros, n.d.). This was intended to enable us to compare participants' retention of vocabulary learnt using monomodal and multimodal flashcards over time. As in the Study Phase, participants were asked to grade their recall with 'Good' or 'Again' depending on whether they recalled the meaning correctly or not. Unlike during the Study Phase, in each post-test each card was seen only once, even when the participant selected 'Again'. Recall accuracy data (the 'Good' count) for cards that had been learnt in a multimodal and monomodal mode during the Study Phase were collected, and the data were analysed using a one-tailed Wilcoxon Signed-Rank test to determine whether learning L2 vocabulary multimodally can result in better recall accuracy even in response to monomodal (text-only) test cues. Lastly, and to answer research question three, the question of why multimodal learning can be more effective than monomodal learning was addressed by discussing the results in light of Paivio's Dual Coding Theory and with reference to insights gleaned from studies in multisensory research (section 4.3).

3.3 Research Design

This study uses a quantitative, within-subjects design (c.f. Carpenter & Olson, 2012). For each participant, half of the word pairs were presented multimodally and half monomodally. This within-subjects design controls for a possible difference in ability between participants (Jhangiani et al., 2019). The insertion order of new cards was random (see Appendix A) to avoid order effects. In addition, the combinations of word pairs and card types were counterbalanced across participants according to a Latin square design to control for possible variation in word pair difficulty (Jhangiani et al., 2019). In a Latin Square design, each treatment condition occurs in every column and row (Rayner & Livingston, 2023; Richardson, 2018); as shown in Table 1, each word pair – divided into sublist 1 and sublist 2 (see Appendix B) – was presented multimodally to (approximately) half the participants and monomodally to the other half. Latin square counterbalancing means that any overall difference in recall accuracy between the two conditions (multimodal/monomodal cards) cannot have been caused by a difference in the difficulty of vocabulary between sublists (Jhangiani et al., 2019).

Table 1. Latin Square.

	Sublist 1	**Sublist 2**
Deck A (used by Group A participants, *n* = 13)	Multimodal	Monomodal
Deck B (used by Group B participants, *n* = 12)	Monomodal	Multimodal

An important part of the research design was to choose a target language that would be completely new to all research participants. The chosen target language to be learned by participants in this study was Esperanto, an artificial language constructed by L. L. Zamenhof in 1887 (The Editors of Encyclopaedia Britannica, n.d.). Esperanto was chosen because – unlike the English language – it is very rare to find someone who has studied or been exposed to Esperanto in Padang, Indonesia; therefore, it would be easy to find participants with no lexical knowledge of Esperanto, which would eliminate the bias of certain participants having pre-existing knowledge of the target language, eliminate the need for a pre-test, and make it easy for the researcher to find novel (i.e., previously not encountered) concrete nouns for participants to learn.

3.4 Research Instrument

The Anki application – a free, open-source application for creating and studying digital flashcards within a spaced repetition system – was used as a research instrument. The first author

(hereafter, 'the researcher') used Anki for macOS, participants used AnkiDroid for Android, and data were synced between the participants' and the researcher's devices via AnkiWeb (Anki's syncing service), enabling the him to upload Anki decks to participants' accounts and collect recall accuracy data remotely. The researcher created an AnkiWeb account for each participant – to be used exclusively for the experiment – and provided each participant with login details for their participant account.

A list of 30 Esperanto–Indonesian word pairs was compiled (see Appendix B). All Esperanto words were concrete nouns of 2-3 syllables in length, denoting objects with which the participants were likely to be familiar (e.g., a chair). The researcher endeavoured to balance the difficulty of L2 words in each sublist based on an analysis of the phonological complexity of each word. The Anki deck options used in the current study can be found in Appendix A. Audio for each Esperanto word was recorded by the researcher, added to multimodal flashcards, and set to play automatically. Pictures for each vocabulary item were sourced online, primarily from https://publicdomainvectors.org/en/. The researcher selected pictures that were easily recognisable, with minimal visual noise, and enough visual context to help participants recognise the pictured object.

Each Anki card was tagged according to card type (multimodal/monomodal). These tags did not appear to participants but were used by the researcher to filter each participant's Anki statistics according to card type, so that recall accuracy data for all multimodal cards and all monomodal cards could be viewed separately and manually input into a spreadsheet. In the Anki deck used for the post-tests (Test Phase), card tags (multimodal/monomodal) indicated whether the word pair was learnt initially in a monomodal or a multimodal way during the Study Phase.

4. RESULTS

4.1 The Study Phase (Research Question One)

The independent variable in this study is the flashcard type used (multimodal or monomodal). The dependent variable measured in the Study Phase is the 'Again' count (i.e., the number of user-initiated reviews) for monomodal and multimodal cards. Since participants were instructed to tap the 'Again' button if they failed to correctly recall the answer for a card, the 'Again' count is indicative of the number of memory lapses. Thus, a higher 'Again' count (number of user-initiated reviews) is indicative of lower recall accuracy, and a lower 'Again' count is indicative of higher recall accuracy. Median recall accuracy was higher for multimodally learnt items than for monomodally learnt items (see Figure 8).

Results of the Wilcoxon signed-rank test (one-tailed) indicated that there were significantly more user-initiated reviews (i.e., 'Again' count) for monomodal flashcards ($Mdn = 61$, $n = 25$) than for multimodal flashcards ($Mdn = 50$, $n = 25$), $Z = -3.4$, $p < 0.001$, $r = -0.7$. Since the number of user-initiated reviews is indicative of the number of memory lapses, the results indicate that significantly more memory lapses occurred for L2 words that were learnt monomodally than for L2 words that were learnt multimodally. Therefore, the null hypothesis (that multimodal flashcards are no more effective than monomodal flashcards as a tool for learning L2 concrete nouns) can be rejected. This finding answers research question one. For this sample, multimodal flashcards are significantly more effective than monomodal flashcards as a tool for learning the meanings (L1 translations) of L2 concrete nouns. The chance of a type I error (rejecting a correct H0) is very small (0.035%), and the results strongly support H1, as the smaller the p-value, the more it supports H1.

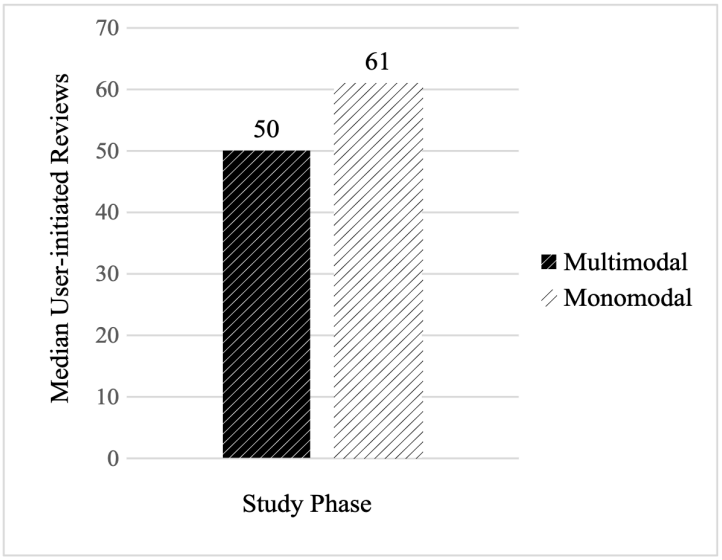

Figure 8. Median number of user-initiated reviews for multimodal and monomodal cards during the Study Phase.

4.2 The Test Phase (Research Question Two)

The dependent variable measured in the Test Phase is the 'Good' count (i.e., the number of correct recalls). Unlike during the Study Phase, each card was presented only once during each post-test, and all test cues were monomodal (text only). Participants were instructed to tap 'Good' if they successfully recalled the answer, and therefore, a higher 'Good' count is indicative of higher recall accuracy. Median recall accuracy was higher for multimodally learnt items than for monomodally learnt items (see Figure 5).

A one-tailed Wilcoxon signed-rank test analysis of the results from post-test 1 showed that there were significantly more correct recalls of multimodally learned items (Mdn = 14, n = 25) than of monomodally learned items (Mdn = 13, n = 25), Z = -2.6, p = 0.005, r = -0.6. Significantly more items that had been learned in a multimodal way were recalled correctly than items that had been learned in a monomodal way, even in response to monomodal test cues, and even though monomodal cards had been reviewed significantly more times on average than multimodal cards during the study phase. Therefore, in answer to research question two, the null hypothesis can be rejected. For this sample, multimodal flashcards are significantly more effective than monomodal flashcards as a tool for learning the meanings of L2 concrete nouns.

As mentioned in section 3.2, the use of increasingly larger intervals between post-tests in the Test Phase was intended to progressively increase the chances that participants would forget what they had learnt, enabling us to compare participants' retention of vocabulary learnt using monomodal and multimodal flashcards over time. However, recall accuracy did not significantly reduce between post-tests as expected. On the contrary, recall accuracy for multimodally learnt cards remained relatively stable between post-tests, with median recall remaining at 14 for each post-test, while recall accuracy for monomodally learned cards actually improved between post tests (13 → 13 → 14). As a result, the effect size of the difference between multimodally and monomodally learned cards reduced between post-tests, and a Wilcoxon signed-rank test analysis of the results from post-test 3 found no significant difference between the number of correct recalls ('Good' count) of multimodally learned items (Mdn = 14, n = 25) and monomodally learned items (Mdn = 14, n = 25), Z = -1.2, p = 0.119, r = -0.3.

The relative stability and progressive improvement of recall accuracy for multimodally and monomodally learnt items between post-tests can be attributed to a weakness in the design of the Test Phase. Firstly, the length of time between post-tests was not sufficient for participants to forget what they had learnt during the Study Phase. Secondly and more significantly, participants were able to learn from the post-tests because they were self-marked, enabling many participants

to improve their scores for monomodally learnt items in post-tests 2 and 3. Progressive improvement in median recall accuracy can be observed for monomodally learnt cards, but not for multimodally learnt cards, because there was much more room (potential) for improvement in the recall accuracy scores for monomodally learnt items. In contrast, participant score for multimodally learnt cards was already very high in post-test 1 (with 10 out of 25 participants scoring full marks), leaving little room (potential) for improvement. This practice effect (see Jhangiani et al., 2019) could have been avoided by using a test in which participants received no feedback.

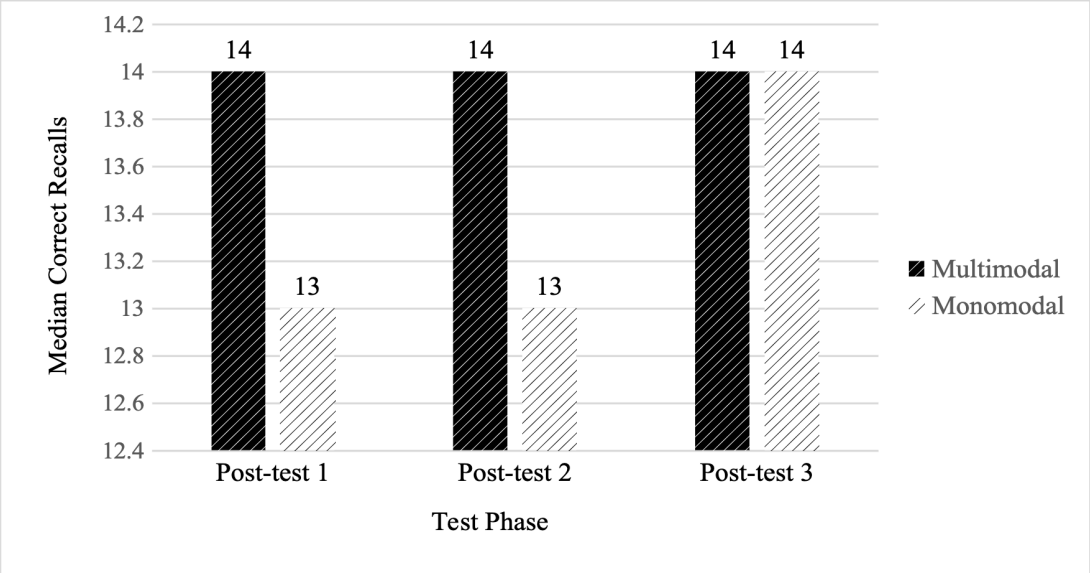

Figure 9. Median number of correct recalls for multimodal and monomodal cards during the Test Phase.

4.3 Interpretation (Research Question Three)

Having shown that multimodal flashcards are significantly more effective as a tool for learning the meanings of L2 concrete nouns compared to monomodal flashcards (research questions one), even in response to monomodal test cues (research question two), this subsection explores why this is the case in light of Dual Coding Theory, and with reference to research in multisensory learning (research question three). The multimodal cards used in this study facilitated the building of connections between the Verbal systems and the Image system (dual-encoding), which can have an additive effect on recall due to the interconnected yet independent nature of each symbolic system (Paivio & Csapo, 1973). This additive effect on recall was due to a larger number of viable retrieval routes for word pairs learned multimodally. Using multimodal cards will have resulted in the availability of twice as many viable retrieval routes compared to using monomodal cards (V2 → I → V1 and V2 → V1, compared to V2 → V1 only, see Figure 10, middle row); meaning that if one retrieval route became decayed (unviable) then an alternative retrieval route could be used to successfully recall the meaning of an L2 word in L1 (see Figure 10, bottom row).

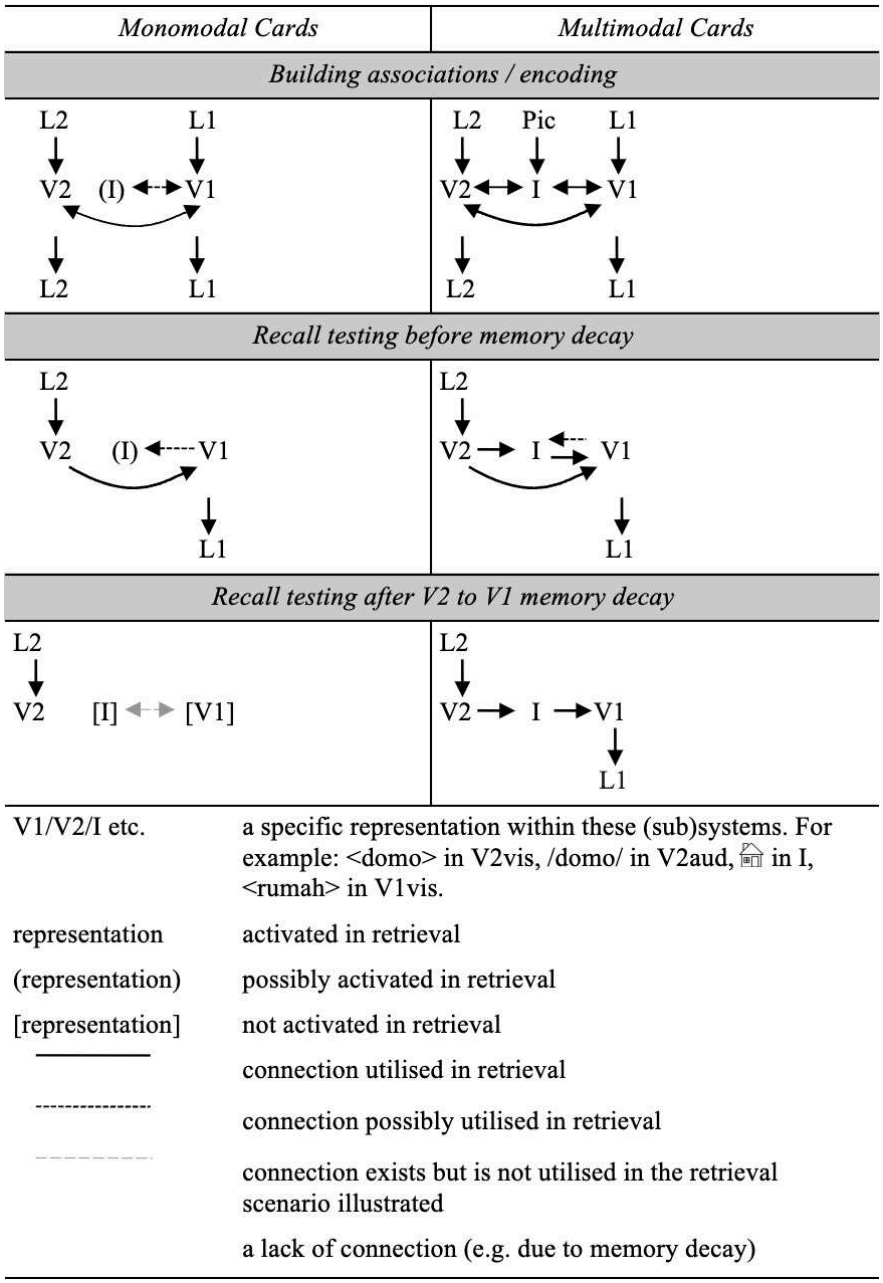

Figure 10. Symbolic system activation mapping.

In addition to facilitating connections between the verbal systems (V1 and V2) and the image system (I), the multimodal cards also facilitated interconnections between the Verbal-visual (V2vis) and Verbal-auditory (V2aud) subsystems by providing both L2 text and audio, which may also have had an additive effect on recall. Although Paivio (1986, p. 57) did not directly conclude this, interconnections between subsystems can be expected to have an additive effect on recall. This expectation stems from evidence, such as the selective effects of focal brain injuries, that shows subsystems corresponding to different sensory modalities can function more-or-less independently of one another.

Providing visual and auditory verbal input simultaneously will have resulted in the formation of visual and auditory representations in V2vis and V2aud, along with strong interconnections between them, plus – when viewing the answer side of the card – strong connections between these V2 subsystems and the other symbolic systems (V1 and I). In contrast, monomodal cards presented only the visual (orthographical) form of the L2 word on the question side of the card, leaving the participant to guess at its proper pronunciation (its phonemic form).

The participant's mind may have attempted to construct an auditory representation of the word by mapping graphemes to phonemes (see Magrassi et al., 2015), but this representation (?V2aud) may have deviated from correct or standard pronunciation. Upon viewing the answer side of the card, connections between V2vis and V1 will have formed, and existing semantic knowledge of L1 may have been retrieved from the Image system, since the availability of imagens is an essential part of semantic memory (Paivio, 1986, p. 121). However, cross-system connections from (?V2aud) to other systems would presumably be much weaker than interconnections formed due to being exposed to multimodal (and multisensory) stimuli (see Figure 11, middle row). If – for example – the connection between V2vis and V1 were to decay, the possible and weak connections formed through monomodal learning may not have been sufficient for the participant to recall L2 in response to an L1 test cue, resulting in lower recall accuracy for monomodally learned items (see Figure 11, bottom row, left column). A participant who has learnt the same word multimodally would be able to rely on other retrieval routes that had been established and strengthened through multimodal learning, enabling successful recall (see Figure 11, bottom row, right column). Furthermore, it is reasonable to assume that imagined representations constructed by the mind without the aid of external input in the representation's corresponding sensory modality only form a kind of tentative knowledge that is not as resilient or likely to be activated compared to representations that have been formed as the result of real-world experience or input.

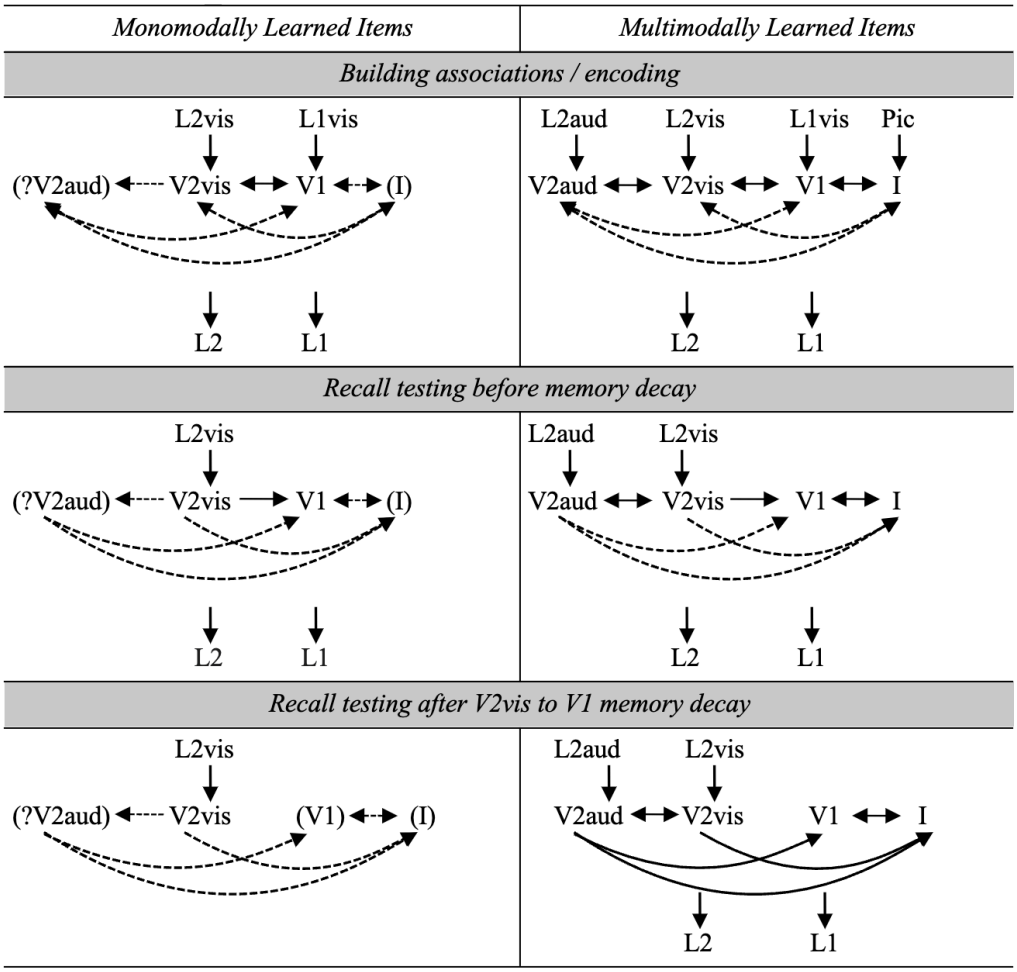

Figure 11. Symbolic system and subsystem activation mapping.

The concept of redintegration as a mechanism of memory retrieval can also be used to explain why learning L2 vocabulary multimodally results in better recall accuracy even in response to monomodal (text-only) test cues. Redintegration is "the capacity of a portion of a consolidated memory to re-activate the entire extended original network" (Thelen & Murray,

2013, p. 497). In the context of the present study, this means that if a word is learned multimodally, resulting in the formation of a cognitive network of representations corresponding to multiple modes (e.g. text, audio, and picture), then a subsequent monomodal test cue that stimulates just one part of this network (e.g. text that stimulates a Verbal-visual representation) can (re-)activate the whole network (i.e. including Verbal-auditory and Image representations). In this way, retrieval of a multimodally learned word operates on a richer and more informative network of interconnected representations, resulting in higher recall accuracy (see Moran et al., 2013, p. 589). From a Dual Coding Theory perspective, Clark and Paivio (1987, p. 9, 1991, p. 154) used the term 'spreading activation' to describe an equivalent (or at least similar) cognitive process by which a stimulus which directly stimulates a representation in one symbolic (sub)system can also indirectly stimulate representations in other symbolic (sub)systems by means of established connections between (sub)systems.

In the current study, the mnemonic advantage of multimodal (multisensory) cards was observed not only in response to multisensory test cues (i.e. during the Study Phase, in which the front side of the card included both L2 text and audio), but also in response to unisensory test cues (i.e. during the Test Phase, in which the front side of the card included only L2 text). This finding is consistent with Thelen and Murray's (2013, p. 483) conclusion that semantically congruent multisensory experience at one point in time improves subsequent unisensory visual (and auditory) object recognition, when compared to objects encountered exclusively in a unisensory context: The visual objects in this study were L2 words, and recognition of these objects was a cognitive prerequisite to recalling the object's paired associate – i.e. its L1 translation equivalent. The results of the current study suggest that exposure to semantically congruent multisensory stimuli enhances not only subsequent unisensory object recognition, but also subsequent unisensory cued recall of an associated object (in this case, the L2 words' L1 translation equivalent). The present study contributes towards a growing body of literature that demonstrates that multimodal (or multimedia or multisensory) learning can be more effective than monomodal learning, and it addresses a gap in the literature by comparing the effectiveness of multimodal and monomodal flashcards for L2 vocabulary learning.

The main implication of this study for L2 teaching and learning is that creating and using multimodal flashcards of the kind used in this study is worthwhile because multimodal flashcards are significantly more effective than monomodal flashcards as a tool for learning the meanings of L2 concrete nouns. While this study used Esperanto as the target language, the mnemonic advantage of using multimodal flashcards for vocabulary learning may be generalised to the learning of any second language, including English. Indeed, flashcards that include both text and audio can be expected to especially benefit learners of languages with complex grapheme-phoneme correspondences, such as English.

5. CONCLUSION

The results of this study show that multimodal flashcards of the kind used in this study are significantly more effective than text-only monomodal flashcards as a tool for learning the meanings of L2 concrete nouns, even in response to monomodal test cues. The mnemonic advantage of multimodal learning over monomodal learning is due to its greater effectiveness at facilitating the formation of interconnections between different symbolic (sub)systems, enabling retrieval to operate on a richer, more informative network of representations, and resulting in a larger number of possible retrieval routes for word pairs learned multimodally, thus improving recall accuracy.

Since participants provided feedback about their own recall accuracy, the main limitation of this study is that its validity depends upon participants' ability and willingness to follow the researcher's instructions. Participants could tap 'Good' even if they failed to recall the answer correctly; however, it is difficult to imagine a motive for doing so since participants were not aware of the research objectives, they knew that their performance data would be anonymised, and they had agreed to follow the researcher's instructions. Future research could replicate this

study but carry out delayed post-tests under controlled laboratory conditions in which answers are marked by the researcher and participants are given no feedback, and with larger intervals between post-tests, which would strengthen the study's validity, avoid the practice effects we observed in post-tests 2 and 3, and allow the researchers to observe the effect of card type on long-term vocabulary retention.

ACKNOWLEDGMENT

The first author would like to express his gratitude to Universitas Andalas for granting the scholarship for his postgraduate study in Linguistics at Universitas Andalas, Indonesia, and for supporting the completion of this study.

REFERENCES

Bakla, A., & Çekiç, A. (2017). Using an online vocabulary memorization tool versus traditional vocabulary exercises. *Ana Dili Eğitimi Dergisi*, *5*(4), 948-966. https://doi.org/10.16916/aded.339241

Boros, A. (n.d.). *Forgetting curve*. The Decision Lab. https://thedecisionlab.com/reference-guide/psychology/forgetting-curve

Carpenter, S. K., & Olson, K. M. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 92-101. https://doi.org/10.1037/a0024828

Carreker, S., & Birsh, J. R. (2018). *Multisensory teaching of basic language skills activity book* (4th ed.). Paul H. Brookes Publishing Co.

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal*, *80*(2), 183-198. https://doi.org/10.1111/j.1540-4781.1996.tb01159.x

Clark, J. M., & Paivio, A. (1987). A dual coding perspective on encoding processes. In M. A. McDaniel & M. Pressley (Eds.), *Imagery and related mnemonic processes* (pp. 5-33). Springer. https://doi.org/10.1007/978-1-4612-4676-3_1

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149-210. https://doi.org/10.1007/BF01320076

Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Kwantlen Polytechnic University.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966-968. https://doi.org/10.1126/science.1152408

Li, W., Yu, J., Zhang, Z., & Liu, X. (2022). Dual coding or cognitive load? Exploring the effect of multimodal input on English as a foreign language learners' vocabulary learning. *Frontiers in Psychology*, *13*, Article 834706. https://doi.org/10.3389/fpsyg.2022.834706

Lin, C.-C., & Yu, Y.-C. (2017). Effects of presentation modes on mobile-assisted vocabulary learning and cognitive load. *Interactive Learning Environments*, *25*(4), 528-542. https://doi.org/10.1080/10494820.2016.1155160

Lotto, L., & de Groot, A. M. B. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning*, *48*(1), 31-69. https://doi.org/10.1111/1467-9922.00032

Magrassi, L., Aromataris, G., Cabrini, A., Annovazzi-Lodi, V., & Moro, A. (2015). Sound representation in higher language areas during language generation. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(6), 1868-1873. https://doi.org/10.1073/pnas.1418162112

Mayer, K. M., Yildiz, I. B., Macedonia, M., & von Kriegstein, K. (2015). Visual and motor cortices differentially support the translation of foreign language words. *Current Biology*, *25*(4), 530-535. https://doi.org/10.1016/j.cub.2014.11.068

Mayer, R. E. (2002). Cognitive theory and the design of multimedia instruction: An example of the two-way street between cognition and instruction. *New Directions for Teaching and Learning*, *2002*(89), 55-71. https://doi.org/10.1002/tl.47

Moran, Z. D., Bachman, P., Pham, P., Hah Cho, S., Cannon, T. D., & Shams, L. (2013). Multisensory encoding improves auditory recognition. *Multisensory Research*, *26*(6), 581-592. https://doi.org/10.1163/22134808-00002436

Mujahidah, M., Hasanah, N., Yusuf, M., Zulfah, Z., & Fatmasyamsinar, A. A. (2024). The implementation of Ankiapp to improve students' vocabulary mastery. *SALTeL Journal (Southeast Asia Language Teaching and Learning)*, *7*(1), 9-18. https://doi.org/10.35307/saltel.v7i1.115

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

Nguyen, B.-P. T. (2021). Mobile-assisted vocabulary learning: A review of Anki. *I-Manager's Journal of Educational Technology*, *18*(3), 16-21. https://doi.org/10.26634/jet.18.3.17917

Okray, Z., Jacob, P. F., Stern, C., Desmond, K., Otto, N., Talbot, C. B., Vargas-Gutierrez, P., & Waddell, S. (2023). Multisensory learning binds neurons into a cross-modal memory engram. *Nature*, *617*, 777-784. https://doi.org/10.1038/s41586-023-06013-8

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford University Press.

Paivio, A., & Csapo, K. (1973). Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology*, *5*(2), 176-206. https://doi.org/10.1016/0010-0285(73)90032-7

Paivio, A., & Desrochers, A. (1980). A dual-coding approach to bilingual memory. *Canadian Journal of Psychology / Revue Canadienne de Psychologie*, *34*(4), 388-399. https://doi.org/10.1037/h0081101

Rayner, J. C. W., & Livingston, G. C. (2023). *An introduction to Cochran–Mantel–Haenszel testing and nonparametric ANOVA*. Wiley.

Refold. (n.d.). *Basic Anki setup*. Refold Languages Inc. https://refold.la/roadmap/stage-1/a/anki-setup

Richardson, J. T. E. (2018). The use of Latin-square designs in educational and psychological research. *Educational Research Review, 24,* 84-97. https://doi.org/10.1016/j.edurev.2018.03.003

Scheff, S. C. (2016). *Fundamental statistic principles for the neurobiologist: A survival guide.* Elsevier. https://doi.org/10.1016/C2015-0-02471-6

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*(11), 411-417. https://doi.org/10.1016/j.tics.2008.07.006

Suen, L. J., Huang, H. M., & Lee, H. H. (2014). A comparison of convenience sampling and purposive sampling. *The Journal of Nursing*, *61*(3), 105-111. https://doi.org/10.6224/JN.61.3.105

The Editors of Encyclopaedia Britannica. (n.d.). Esperanto. In *Encyclopædia Britannica, Inc.* Retrieved https://britannica.com/topic/Esperanto

Thelen, A., & Murray, M. M. (2013). The efficacy of single-trial multisensory memories. *Multisensory Research*, *26*(5), 483-502. https://doi.org/10.1163/22134808-00002426

APPENDICES

APPENDIX A

The current study used Anki Version 2.1.54, with scheduler v2. The tables below show the Anki settings used. Any settings not listed below were left at their default values. These options were set in Anki for macOS and synced via AnkiWeb to participants' AnkiDroid applications on their smartphones.

Table A1. Anki Deck Options used for study decks.

Daily limits	
New cards/day	6
Maximum reviews/day	9999
New cards	
Learning steps	1m 10m
Graduating interval	1
Easy interval	1
Insertion order	Random
Lapses	
Relearning steps	10m
Minimum interval	1
Leech threshold	6
Leech action	Tag Only
Advanced	
Maximum interval	1

Table A2. Anki Deck Options used for post-test decks.

Daily limits	
New cards/day	30
New cards	
Learning steps	365d
Graduating interval	365
Easy interval	365
Insertion order	Random
Lapses	
Relearning steps	[BLANK]
Advanced	
Maximum interval	365

APPENDIX B

Table B1. Esperanto–Indonesian word pairs used in the current study.

English	Indonesian	Esperanto	Phonemic transcription (EO)	Syllable count	Sublist
house	rumah	domo	/domo/	2	1
gold	emas	omo	/oro/	2	1
paint	cat	farbo	/farbo/	2	1
water	air	akvo	/akvo/	2	1
frog	kodok	rano	/rano/	2	1
arrow	anak panah	sago	/sago/	2	1
cloud(s)	awan	nuboj	/nuboi/	2	1
lime	jeruk nipis	kalko	/kalko/	2	1

Table B1 continued…

snail	siput	heliko	/heliko/	3	1
cabbage	kol	brasiko	/brasiko/	3	1
tap	kran	frapeti	/frapeti/	3	1
lightbulb	bola lampu	ampolo	/ampolo/	3	1
ant	semut	formiko	/formiko/	3	1
scissors	gunting	tondilo	/tondilo/	3	1
donkey	keledai	azeno	/azeno/	3	1
hand	tangan	mano	/mano/	2	2
ball	bola	pilko	/pilko/	2	2
seed	biji	semo	/semo/	2	2
bed	tempat tidur	lito	/lito/	2	2
egg	telur	ovo	/ovo/	2	2
raft	rakit	floso	/floso/	2	2
garlic	bawang putih	ajlo	/ailo/	2	2
goat	kambing	kapro	/kapro/	2	2
duck	itik	anaso	/anaso/	3	2
lettuce	selada	laktuko	/laktuko/	3	2
roof	atap	tegmento	/tegmento/	3	2
ginger	jahe	zingibro	/zingibro/	3	2
monkey	monyet	simio	/simio/	3	2
rabbit	kelinci	kuniklo	/kuniklo/	3	2
parrot	Burung kakak tua	papago	/papago/	3	2