The Effect of Cognitive Load and Immersion in an Immersive Virtual Environment, on Vocabulary acquisition.

Finn Slots

June 2025

Bachelor Advanced Technology

dr. Mariët Theune
Human Media Interaction Group
University of Twente

dr. ir. Robby van Delden
Human Media Interaction Group
University of Twente

dr. ir. Michel de Jong
NanoElectronics Group
University of Twente

Abstract

Acquiring new vocabulary can be one of the most time-consuming and tedious aspects of learning a new language. Literature suggests that performing intuitive gestures and learning in an immersive virtual reality (IVR) environment is beneficial for vocabulary acquisition. In this study, I create an IVR environment where Japanese words can be learned in an engaging way by exploring the world and selecting target objects. 19 students of the University of Twente learned 54 Japanese words divided over three learning conditions: 1) an immersive world, where the participant performs intuitive gestures, 2) an immersive world, where only the target word is shown, and 3) an empty world, where the participant performs gestures as well. The study focuses on how these conditions affect immersion and cognitive load during learning, and how they influence short-term and long-term learning outcomes. The participants were tested on their recall memory of the words after each learning condition, and again after one week on all of the words. The results suggest that learning in a contextual world increases the immersion felt, but the performance of gestures does not significantly impact the immersion. There is no significant difference in the cognitive load experienced across all conditions. It appears that the increase in immersion due to the contextual world, with the seemingly limited effects of the cognitive load, affects both short-term acquisition as well as long-term word retention. The performance of gestures shows no significant difference compared to where no gestures are performed. This shows that an immersive environment aids language learning, which is promising for the potential application of IVR for language learning in classrooms.

Contents

1 Introduction

Learning a new language is something many people will have to do at some point in their lives. There are many different reasons to learn a new language, ranging from work in an international company, school, travel, to just for fun. On Duolingo alone, over 500 million people are studying a language [1]. One aspect of learning a language is vocabulary acquisition (VA), otherwise known as learning new words. Learning new words can be a tedious and time-intensive process for people [2]. To streamline these tedious tasks, there are many opportunities technology could help in [3]. In this report, I dive into the possibilities Immersive Virtual Reality (IVR) could provide.

Previous research has already proven that IVR can aid in learning [4–6], and especially long-term vocabulary retention [3]. Furthermore, VR could bring more accessibility to effective language learning, to people with attention difficulties [7], or ADHD [8]. However, many questions remain on how to implement this best. Bergsma [9] conducted a study to outline the effect of new or learned context on VA in IVR. Here, a new context means a new visual representation of an already encountered object-word combination, whereas a learned context means the same visual representation of the object of the combination. She did not find a significant difference between both conditions, but did show a promising 69% - 75% retention after one week in both conditions. Her research acts as a starting point for my research, as the VR environment can be changed efficiently to investigate related elements of learning.

Another way to aid VA is through embodied learning. Embodied learning emphasizes the importance of body movement while learning [10], which can make learning more involved and make abstract concepts more concrete [11]. Sun et al. [12] showed in their research that performing intuitive gestures in IVR can improve learning; however, more extensive body movements may impair learning instead, due to causing a higher cognitive load. An intuitive gesture is a hand movement related to a target word, which should be intuitive to the learner and aid in VA.

Ratcliffe et al. [13] research the difference between performing a gesture with object manipulation or without object manipulation. In their study to contextualize interactions in IVR, they tested the effect of object manipulation in aiding the VA of related verbs. They claim that performing object manipulation leads to better memorization than having no object manipulation. For example, grabbing a virtual cup and making a drinking movement, while learning the word for *to drink* as opposed to not grabbing the cup and only making the drinking motion.

In a different study, Ratcliffe et al. [14] researched the effect of the richness of virtual feedback on verb memorization. They concluded that if the feedback is too rich, it shows a negative impact on VA. This seems to be in line with the findings of Sun et al. [12], which state that the performed gestures should not be too extensive for learning in VR, due to an increase in the cognitive load. The study by Radcliffe et al. [13], suggesting that having object manipulation is superior to not having object manipulation, took place in an 'empty' IVR environment, where there were no distractions besides the target objects. So, we can assume that the cognitive load and immersion in this world were lower, than in a contextual environment.

Fuhrman et al. [4], in their research to aid VA, also combined both techniques, IVR and embodied learning. Their study investigated the effect of object manipulation in VR on long-term retention of nouns. They compared three conditions: seeing the object, performing an irrelevant gesture, and performing an intuitive gesture. Their research showed that short-term word acquisition improved by employing embodied learning in the context of picking up objects and performing intuitive gestures, while holding the objects in a contextual world in IVR. A contextual world, which is the opposite of a non-contextual or empty world, is a virtual world meant to represent a place or location, such as the kitchen in Fuhrman et al. [4] their study.

However, as indicated, the potential downside is that (complex) gestures might be adding too much cognitive load. Whereas high cognitive load impairs learning in VR, a high immersion aids it [12, 14]. A high cognitive load results in the brain focusing on many different things, resulting in less focus on the learning. If the immersion is higher, participants are more inclined to do the learning task and focus better on it. Because of this reason, I focus on the consideration between the cognitive load and immersion while learning in VR.

While Ratcliffe et al. [13] focused on VA of verbs, I want to focus my study on the VA of nouns, to keep in line with what was done by Fuhrman [4] and Bergsma [9]. In most literature, the second assessment, for testing long-term learning, takes place after one - three weeks [4, 9, 13, 14]. I will do this after one week, due to logistical reasons. This study thus investigates the effects of 1) performing intuitive gestures with object manipulation and 2) learning in a contextual IVR world, on whether or how we can balance the limited increase in cognitive load while maintaining a high level of immersion during vocabulary acquisition.

An existing IVR environment [9] was used. In this world, visual representations were used that depict target words (words to be learned). Adjustments were made to the IVR world to incorporate the use of gestures while learning. A user study was conducted, where word retention, immersion, and cognitive load were measured, and participants' gestures were checked. The research question for this study is as follows:

Research question:
To what extent can immersion be increased, while maintaining a reasonable cog-

nitive load, by 1) performing intuitive gestures and 2) learning in a contextual world, and does this enhance noun acquisition in immersive virtual reality?

The hypothesis is as follows:

The gained immersion due to both the contextual world and performance of intuitive gestures has a larger positive effect than the drawback of increased cognitive load. Thus, both the short-term acquisition of the words and long-term retention will be aided most if both are present.

This thesis has the following structure: In Section 2, the theoretical background is outlined in more detail. In Sections 3 and 4, the design process and the methodology are discussed, respectively. Section 5 contains the results, which are discussed in Section 6. The conclusions are drawn in Section 7.

2 Background Theory

To form a basis for this research, the first phase included a literature research. A foundation was made of 3 main papers, from Furhman et al. [4], Bergsma [9], and Ratcliffe et al. [13]. After collecting papers manually, an online tool was employed to streamline the process. PURE suggest [15] is a scientific literature search tool, in which you insert the foundational papers surrounding a research topic. Based on the citations and references of these seed papers, it will suggest additional papers for your study, resembling an automated way of a multi-level (reverse) snowball sampling method. The resulting collection is summarized and discussed below in relation to this current study.

Christou et al. [16] and Huang et al. [17] have conducted a systematic review study on the benefits of AR and VR technologies on language learning. While Christou focused their study on acquiring general knowledge in a foreign language, such as VA (28/88 papers). In Huang's study, this was almost half (16/33). Both studies suggest that VR increases motivation through its immersive environment, multi-modal (gestural, visual, textual, auditory) feedback, and interactive learning contexts. Huang found that learning in an explorable VR environment can also reduce anxiety by creating a safe space where students are allowed to make mistakes. Therefore, in my study, I will maintain the explorative character from the VR world of Bergsma [9], where participants had to walk around the world and search for the target words themselves.

There are some limitations to using VR. VR can only be employed for relatively short learning sessions, as extended exposure to VR can be uncomfortable and cause motion sickness [16]. Additionally, if there are too many distractions in the VR world, such as too many animations or sounds [16], or if a student is given too much freedom in the VR world [17], they can be demotivated to do the intended learning. One partial contributor could be cognitive overload, where the brain receives too many stimuli to be able to focus on any of them well [18].

In the following sections, I will dive deeper into the effect of the VR environment on immersion and cognitive load. I will take a deeper look at the effects that performing gestures in these contextual worlds has on motivation and cognitive load. I will also explain the effect of VR-assisted learning on the long-term VA of foreign words.

2.1 Immersion

Immersion[1] is described as the feeling of being completely immersed in a VR environment, according to Essoe et al. [3]. For example, if you are in a VR environment in a barn, you would almost feel you are actually in a barn, as opposed to just looking at a screen of an avatar in a barn. Essoe et al. [3]

[1] Some researchers define a difference between immersion and presence, and others use the terms interchangeably. I will use them interchangeably in this paper.

and Ratcliffe and Tokarchuk [19] showed in their studies that feeling present improves language learning, though Essoe [3] states the increased motivation of the participant has a positive role in this, Ratcliffe and Tokarchuk [19] claims the motivation does not affect the learning outcome.

In their study, Essoe et al. [3] compared the difference in learning gain between two groups learning words from 2 similar-sounding foreign languages, in a desktop-based non-immersive VR environment. The first group learned the words for both languages in distinct environments, while the second group learned the words in the same environment. The group that learned the languages in distinct VR environments showed better retention rates under three conditions: 1) The participants needed to feel a high presence. This increased their motivation for the learning exercises. 2) A unique VR environment per language was required. 3) Only the long-term retention rates were significantly better than those of the control group. For the short-term retention rates, this was not the case.

This is in line with the research conducted by Lamers and Lanen [20]. They compared the VA of participants when they learned words in real life or VR, and were examined in either the same or the other condition. They showed that when participants learned words in VR and were examined in real life, they had a 24% lower retention rate, when compared with participants who stayed in the same condition for both learning and examination. However, this was only tested for short-term retention rates, which cannot simply be extended to long-term retention.

The VR environment Lamers and Lanen [20] used was a VR replica of the desk set-up used for the learning and testing in real life, missing the integral engaging and unique environments VR can be used for in language learning. Wälti et al. [21] state that reinstating visual contexts such as VR does not promote recall memory significantly. Reinstatement means that the conditions in which words are learned are replicated for the posttest. Additionally, the effects on memory were only tested in the short term. Both these studies point to the shortcomings that come with employing VR, but the hypothesis, as stated in the introduction, remains worth investigating nevertheless. IVR offers unique learning advantages that are difficult to replicate through other methods.

Ratcliffe and Tokarchuk [19] conducted a study in a VR coffee shop to see the effects of motivation, presence, and embodiment on VA. The results showed a positive correlation between presence and learning gain, but showed no correlation between motivation and either presence or learning gain. This might be since all participants indicated a high level of motivation, so there is little variance in the levels of motivation between participants.

From these studies, we can infer that high levels of immersion result in better learning. In her study, Bergsma [9] indicates that it does not matter whether the same or a new virtual environment is used between multiple learning sessions. As long as these worlds are fully immersive, VA should improve. If participants do not feel enough presence, it can have no or even adverse effects on VA.

2.2 Gestures

Research has extensively shown that VA is aided by performing gestures related to the target words, both in VR [4, 22], as well as in real life [23]. Macedonia et al. [23] show that performing an intuitive gesture while learning a new word activates different brain areas than performing a non-intuitive gesture. The combination of behavioral, more effective learning, and neural evidence proves that performing intuitive gestures positively impacts VA.

Fuhrman et al. [4] investigated the difference between performing 1) no gesture, 2) an intuitive gesture, and 3) a non-intuitive gesture while learning nouns in a VR kitchen environment. They showed that performing an ituitive gesture results in the best short-term vocabulary acquisition. The long-term retention was also better, but not significantly. They reason that the increased cognitive load of performing a non-intuitive gesture, needing to memorize the movements, might have led to the worse results. An important distinction to make is that the intuitive gesture was with object manipulation, and the non-intuitive gesture was without object manipulation.

Ratcliffe et al. [13] and Macedonia et al. [24] show the promising results of gestures in combination with object manipulation. Whereas Macedonia compared these gestures to having only audiovisual feedback outside of VR, without performing any embodied learning, showing that gestures lead to better long-term word retention, Ratcliffe shows that gestures with object manipulation work more effectively for VA than gestures without. He shows that participants feel a higher presence when object manipulation is incorporated, as opposed to when it's not, while they do not report a difference in the level of realism of the VR simulation. This might be due to the non-contextual world Ratcliffe used in his study, where the objects of target words were presented on a podium in an empty space one by one.

2.3 Cognitive Load

Makransky et al. [18] describe cognitive load of a task as the amount of load the brain demands to perform this task. You experience cognitive overload when the cognitive load required to perform a certain task exceeds the capacity of your working memory. In their research, Makransky [18] states that VR environments cause a higher cognitive load than non-VR learning environments. This is especially true of certain features he calls 'seductive features'; these are features designed to grab your attention. Blinking lights, for example, or high-pitched sounds.

Ratcliffe et al. [14] conducted a study on the effectiveness of feedback in VR while learning foreign words. They made the distinction between low interaction feedback (an abstract green tick) and high interaction feedback (audiovisual feedback based on the gesture performed, as well as the green tick). They found that the high interaction feedback impaired the learning gain, which is in line with Makransky's claim.

Markansky et al. [18] and Sun et al. [12] state that it is not the medium of IVR itself that is responsible for either positive or negative outcomes on learning, but the way it is employed. While Markansky [18] believes a balance should be found between cognitive engagement and cognitive load, Sun [12] shows that encouraging exploration might be more beneficial. In their study, Sun [12] describes how participants who explored the VR environment more had better learning outcomes than those who did not explore much. They do note that participants with more video game experience explored more than participants with less video game experience, due to their familiarity with virtual environments.

2.4 Conclusion

Considering all factors from previous research, some important requirements arise for VR and embodiment to have a positive result on VA. Firstly, a high feeling of presence is beneficial, which could be achieved by a contextual world that encourages exploration. This world should stay away from having too many distractions, outside of the target learning goals, to avoid distracting the learner. Secondly, the gestures performed need to be intuitive to aid the learning process. If this is not the case, it will have adverse effects on VA. Thirdly, gestures that incorporate object manipulation have been shown to work better than gestures that do not, therefore, object manipulation should be taken into account when building the system.

With all decisions, both the impact on the immersion and the impact on the cognitive load of the learner need to be considered. Since raising the presence also raises the cognitive load, smart decisions should be made to raise the presence, while trying to avoid introducing distractions.

3 Design

I conducted an experiment on the campus of the University of Twente to prove the hypothesis as stated in the introduction. The experiment was designed as a within-group design, where there were three different conditions. All participants went through all of these conditions in the experiment. The three conditions were: 1) Learning the words without performing gestures, in a contextual VR environment, 2) learning the words while performing gestures, in a contextual world in VR, and 3) learning the words while performing gestures, in an empty world in VR. All learning conditions featured different words to be learned in the target language, Japanese. Each condition was followed immediately by a posttest to test the short-term VA, and a week later, a second posttest on all the words was conducted to test the long-term VA.

The main research question had been split into eight separate sub-questions. Once these sub-questions are answered, the main research question can be answered. The sub-questions are:

1) To what extent does performing intuitive gestures influence the immersion?
2) To what extent does performing intuitive gestures influence the cognitive load?
3) To what extent does a contextual VR world influence the immersion?
4) To what extent does a contextual VR world influence the cognitive load?
5) To what extent does performing intuitive gestures influence the short-term vocabulary acquisition?
6) To what extent does a contextual VR world influence the short-term vocabulary acquisition?
7) To what extent does performing intuitive gestures influence the long-term word retention?
8) To what extent does a contextual VR world influence the long-term word retention?

With the hypotheses:

1) Performing intuitive gestures positively affects the immersion.
2) Performing intuitive gestures positively affects the cognitive load.
3) A contextual VR environment positively affects the immersion.
4) A contextual VR environment positively affects the cognitive load.
5) Performing intuitive gestures positively affects short-term vocabulary acquisition.
6) A contextual VR environment positively affects short-term vocabulary acquisition.
7) Performing intuitive gestures positively affects long-term word retention.
8) A contextual VR environment positively affects long-term word retention.

3.1 VR World and Word Selection

Bergsma [9] has previously built a contextual world in VR in the social VR program Neos VR, which was ported to the strongly related platform Resonite. Resonite, too, is a social VR platform that provides infrastructure to create interactive worlds for VR applications. One of the worlds in this experiment was a low-poly representation of a bedroom with a garden, which functioned as the contextual world for this experiment. Figure 1 shows this world, where in Figure 2, a close-up of the bedroom can be seen. The decision was made to choose this world as opposed to the others used by Bersma [9], due to the low-poly nature of the environment. It proved much easier to find low-poly 3D representations of target words than objects in other art styles. The non-contextual world was a brown disk of roughly the same size as the contextual world, blocked off by invisible walls. Around it, there is a basic skybox and a grid for the ground, as can be seen in Figure 3. All three worlds allowed object manipulation of the 18 objects, representing target words, scattered around the world.

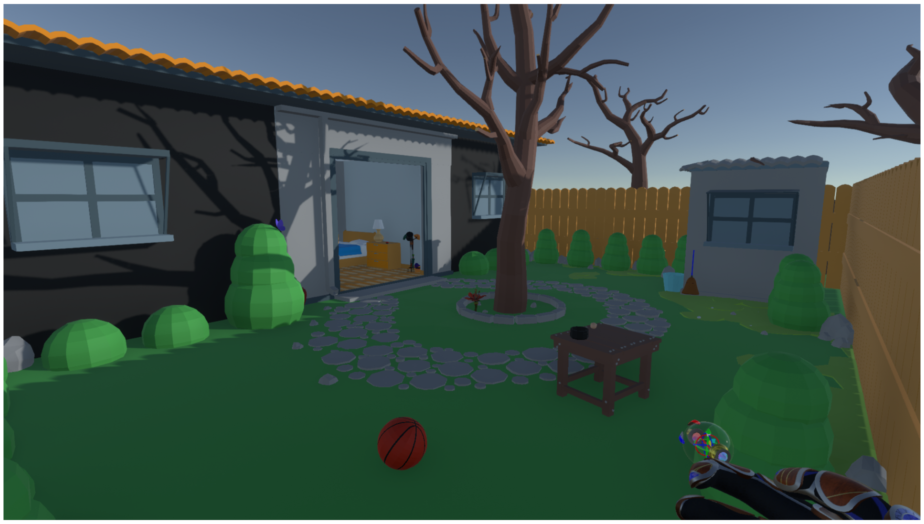

Figure 1: Contextual world (A & B): the garden with objects, representing the target words scattered through it.

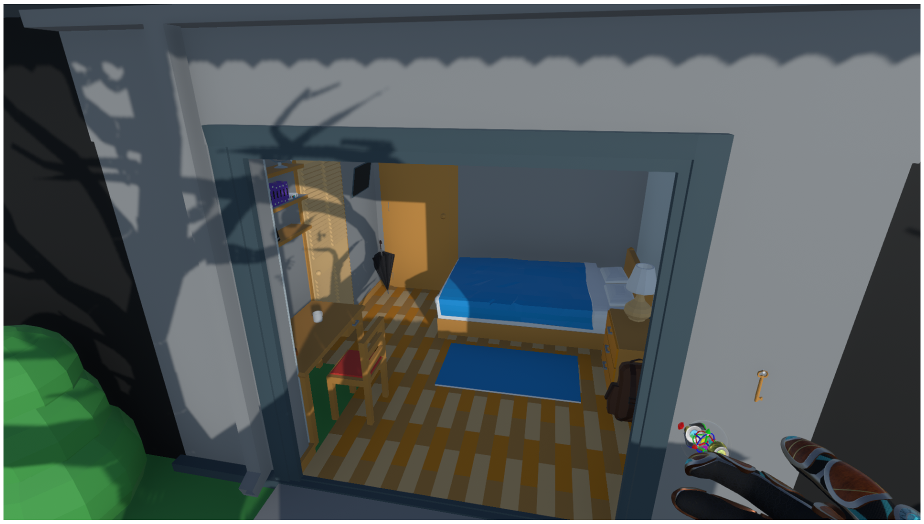

Figure 2: Contextual world (A & B): the bedroom adjacent to the garden.

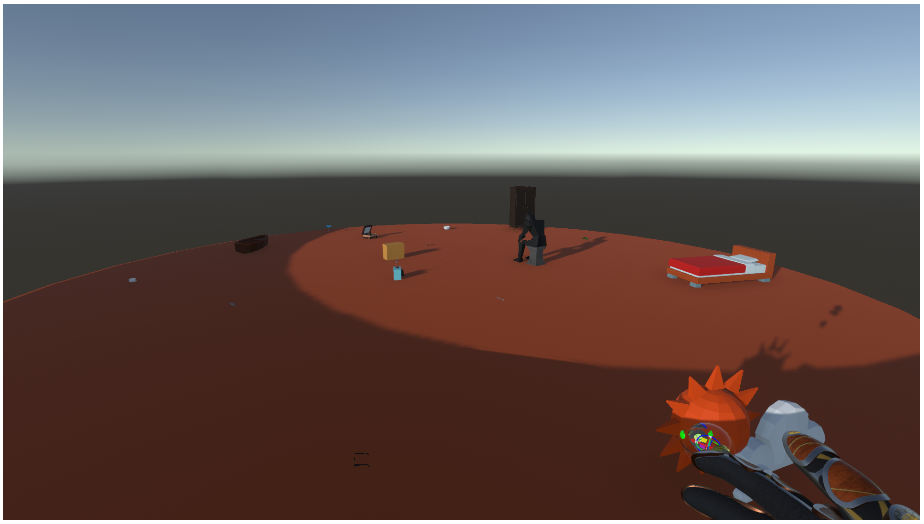

Figure 3: Non-contextual world (C), with objects, representing the target words scattered on the disk.

For her study, Bergsma [9] compiled a list of potential words that should be learned in her experiment. This list was used as a basis, since the objects representing these were already integrated into the VR environment. Seven words were left out, as it was not possible to find a gesture corresponding to this word that was both distinct and intuitive. For example, the word *kinoko* –mushroom was left out, as the gesture, picking it, was already used for the word *hana*

–flower. On the other hand, the word *tsukue* –desk, was excluded since no intuitive gesture could be found that would refer to desk specifically.

Bergsma, in her study, had a longer list of initially considered words, which she did not include in her final study. From this list, I selected the other words for this study. She did not choose these words since they did not fulfill the requirement of having five distinct 3D representations of these objects. Since I only needed a single, low-poly representation, this was not a consideration for me. Words were selected based on two factors: 1) if a low-poly representation of the word could be found, and 2) if an intuitive gesture could be made regarding this word. The low-poly objects for the words were found and downloaded via Google's Poly Pizza [25].

One word was picked, which was not found in the in the initial list of Bergsma. This was the word *boushi* –hat. In the wordlist, the word *boushikake* –hatstand was found, which was also selected. Since the word for hatstand was deemed recognizable enough, it is reasonable to assume the word for hat should also be recognizable enough to the average person. This decision was made so that in all three environments, there is a word that contains a different word. These word pairs being: *hon* –book & *hondana* –bookcase, *boushi* –hat & *boushikake* –hatstand, and *ocha* –tea & *chabin* –teapot. Bergsma had very consciously done this in her study to boost the confidence of the participant, who could recognize part of a word inside the other word. I wanted to keep this little confidence boost for participants. These related words are deliberately put close together in the three VR environments.

3.2 Word division

A total of 18 words per learning environment have been selected, for a total of 54 words. As many words as possible were selected from the list made by Bergsma. This amount is in line with related research that used a wide range of words, between 15 to 92 words [3, 4, 9, 13, 23]. The words that did not have a direct translation yet were translated by Google Translate.

The words have been divided into the three environments based on the difficulty of the words, so that each list has as many 'difficult' words as it has 'easy' words. Similar to the reasoning of Bergsma [9], difficulty of a word was based on two factors: 1) the number of syllables in the word, and 2) the similarity it has to the English word for the object. An overview of the word division can be found in Table 1. Each row in the table should contain words of similar difficulty. Row 3 has, for example, only 2 syllabic words that are very similar to their English translations: *naifu* –knife, *potto* –pot, and *beddo* –bed.

A random sequence was picked to choose the order of the words for the immediate posttests, via the site random.org [26]. The same randomized sequence was picked for all three word lists, as can be seen in Table 1, to make sure they are as fairly divided as possible. The chance that participants find the pattern and exploit this is highly unlikely, since the objects are not placed in the same order as in the list and therefore cannot be logically linked to the order of the list.

The word list of the delayed posttest is semi-random. First, the lists A, B, and C have been randomized separately. The first thirds of each randomized list were combined and re-randomized to form the first third of the final list. The same was done for the second and final thirds of the lists to form the second and final thirds of the complete list. This was done to make sure the three lists are represented equally in the starting, middle, and final parts of the delayed posttest.

Table 1: Word list for all environments, with English words and their Japanese translation, written in Latin characters (romaji).

list A	list B	list C
Book – hon	Hat - boushi	Tea - ocha
Bookcase - hondana	Hatstand - boushikake	Teapot - chabin
Knife - naifu	Pot - potto	Bed - beddo
Lamp – ranpu	Curtain - kaaten	Bowl – bouru
Cup - kappu	Beer - biiru	Clock - kurokko
Hand - te	Eye - me	Statue - zou
Eiffel tower – efferutou	Computer – konpyuutaa	Closet – kurouzetto
Bag - kaban	Ball - tama	Boat - fune
Car - kuruma	Fish - sakana	Glasses – megane
Key - kagi	Button - botan	Spoon - supuun
Phone -denwa	Pillow - makura	Music - ongaku
Chair/stool - isu	Flower - hana	Chest of drawers - tansa
Earth – chikyuu	Butterfly – chouchou	Frog – kaeru
Screwdriver – doraibaa	Garbage bin – gomibako	Telescope – bouenkyou
Camera – shashinki	Outlet – deguchi	Sun – taiyou
Watering can - jouro	Vase - kabin	Shoe - kutsu
Umbrella - kasa	Rainbow - niji	Cloud - kumo
Rug - juutan	Broom - houki	Television - terebi

3.3 Gestures

I define a gesture as a motion made with one or both hands that involves object manipulation. The gesture may involve translational and/or rotational movements of the entire hand and can also include finger movements. The finger movements are not very precise, as they are difficult to mimic with the VR controllers in hand. An important distinction to make is that what I define as a gesture in this study involves object manipulation, but not object interaction. The objects can only be moved in the translational and rotational domains, but cannot be transformed (for example, in size or shape) or interacted with in any other manner. This object manipulation is possible for all target objects (also in the condition without gestures), as well as for most objects scattered throughout the world. When an object that is not a target object is grabbed. relocated and let go, it snaps back into its place, while the target object does not.

The words have been selected based on whether an intuitive and unique gesture could be selected for each word. I came up with all of the gestures. A sanity check has been performed with colleagues, where they had to link the gestures to the words for all 3 sets. They took about the same time per set and had no mistakes in any of them. Therefore, it can assumed that the gestures in the sets are equally clear. In Table 2, all words and their intuitive gestures are shown.

I have assigned all of the gestures to two major categories based on how the gesture is linked to the word it represents: 1) Action gestures and 2) Descriptive gestures, where action gestures can describe an action done with/to the entire object, or an integral part of it. For example, for the word *houki* –broom, the gesture is to act out sweeping the floor. For the word *konpyuutaa* –computer, the action is to act out typing on a keyboard, where the keyboard is seen as an integral part of the computer, even though it is not the entire computer, but just a part of it. A descriptive gesture describes the object, as opposed to performing an action on the object or part of it. An example is the word *kabin* –vase. Here, the gesture is to make the outline of a vase in the air. I preferred to come up with action gestures, as object manipulation is more logical when an action can be done to an object, as opposed to 'describing' the object while holding it. Descriptive gestures were only used when I could not think of an action gesture that I deemed intuitive. A total of 11 out of 36 gestures were descriptive, with 5 in condition B and 6 in condition C.

3.4 Data collection

For this study, the data that were collected are: 1) the performance on the immediate and delayed posttests, 2) the level of immersion felt, 3) the amount of cognitive load experienced, 4) the total time spent in each learning session environment, and 5) demographic information.

The data on immersion were collected to answer sub-questions (as given in Section 3) 1 and 3, and the data on cognitive load were collected to answer sub-questions 2 and 4. These data on immersion and cognitive load, as well as the results of the immediate posttest, were used to answer sub-questions 5 and 6. To answer questions 7 and 8, the data on immersion and cognitive load, as well as the results of the delayed posttest.

The posttest and delayed posttest consisted of a written recall test of the target words. The English word was given, and the Japanese representation of this word should be given in romaji. Romaji is the standardized spelling of Japanese

Table 2: List of all the words and their intuitive gestures. Descriptive gestures are indicated with a 'D', all other gestures are action gestures.

Condition B		Condition C	
Hat	tip hat	Tea	drink with pinky out
Hatstand	put hat on hatstand	Teapot	pour
Pot	open, close lid	Bed	tilt head, sleep on hands
Curtain	close curtains	Bowl	form bowl in air (D)
Beer	cheers	Clock	point at wristwatch (D)
Eye	point at eye (D)	Statue	mimic the pose (D)
Computer	type on keyboard	Closet	open closet doors
Ball	throw	Boat	row
Fish	swimming motion with hands (D)	Glasses	take off and put back on
Button	press	Spoon	eat
Pillow	shake	Music	dance
Flower	pick	Chest of drawers	open drawer
Butterfly	make butterfly in air (D)	Frog	jump like frog (D)
Garbage bin	open lid, throw garbage in	Telescope	look through telescope
Outlet	plug in cable	Sun	put hand above eyes to see (D)
Vase	form vase in air (D)	Shoe	put on
Rainbow	form rainbow in air (D)	Cloud	act out rain in air (D)
Broom	sweep	Television	change channel

words using the Latin alphabet [27]. The participants only learned the words in romaji as well, since they are more familiar with the Latin Alphabet than any of the Japanese scripts. All posttests were done on paper to not put the participants in VR longer than necessary to avoid the health risk related to using VR. The scores were judged by the main researcher based on their phonetic correctness. An answer can receive *0*, *0.5*, or *1* point.

Even though research suggests changing scenery between the learning and test environments harms learning results [20], this should not influence my results, as this is the case for all participants. It also resembled the real-life application of VA more closely. Words should be remembered in all contexts, not only in VR.

The level of immersion and the cognitive load experienced were both assessed with a questionnaire. For the level of immersion, an existing and validated questionnaire was used, the Igroup presence questionnaire [28]. It was conducted via a Microsoft Form, which can be found in Appendix C. On their website, or in the paper by Tran et al. [29], Igroup did not specify whether the questionnaire should be randomized or a specific order should be used. They did present an online tool to take the questionnaire. Here, the questions were given in what seemed like a specific order, different from the order given in the general overview of questions on their site. The order given in this online tool has been used in this study as well. For the amount of cognitive load experienced, the NASA task load index (NASA-TLX) questionnaire was used [30], which was conducted via the NASA-TLX app, available in the Apple App Store.

The total time in each learning environment was timed using a stopwatch by the researcher, where the time measured was rounded to the nearest minute.

The time per learning condition was recorded to explain possible outliers in the target data.

Finally, the following demographic data were collected by self-reporting of the participant. The participants provided their biological sex and age, and their familiarity with Japanese was measured, which was graded from *no proficiency* to *complete fluency*. Finally, participants are grouped based on the number of VR experiences they have had: *0*, *1–4*, *5–9*, or *10+* times.

3.5 VR environment

As the VR environment used is a modified version of the one Bergsma used in her study [9], many of the design choices have already been made by her and have been reused for this research. Only the additions or changes made specifically for this study are outlined in this section.

Before the participants could enter the learning areas, they first entered a tutorial area. This area contained some text boxes explaining how to interact with the object and giving directions for how to proceed in the study. There were two objects, a lighthouse and an apple, with which the user could practice interacting. These objects did not appear in any of the learning conditions. The text in these boxes was changed to reflect the different procedure in this study. To the apple, a 3D animation was attached so the participant could see what it would look like in the conditions where gestures needed to be done. .

In the learning environments, the objects representing the target words were located throughout the location. In the contextual world, the objects were spread out to encourage exploration of the entire world. In the non-contextual world, the objects were also spread around. The decision was made to do this, as opposed to having the objects lined up or appear in the world one by one, to mimic the contextual world as much as possible.

In the condition where no gestures needed to be performed, similar information appeared as in the experiment of Bergsma [9], a pop-up with the English and Japanese words of the object. In her study, Bersgma also played an audio recording of the pronunciation of the word. I do not play this as in combination with the gesture, this would cause too much of a mental load, as multi-modal feedback is not preferred as stated by Ratcliffe et al. [14]. In the conditions where gestures needed to be performed, an animation of one or two 3D modeled hands appears when the object is grabbed, depicting the to-be-performed gesture. These 3D animations were made in the 3D modeling program Blender. The text and animations that appear can be seen in Figures 4, 5, and 6.

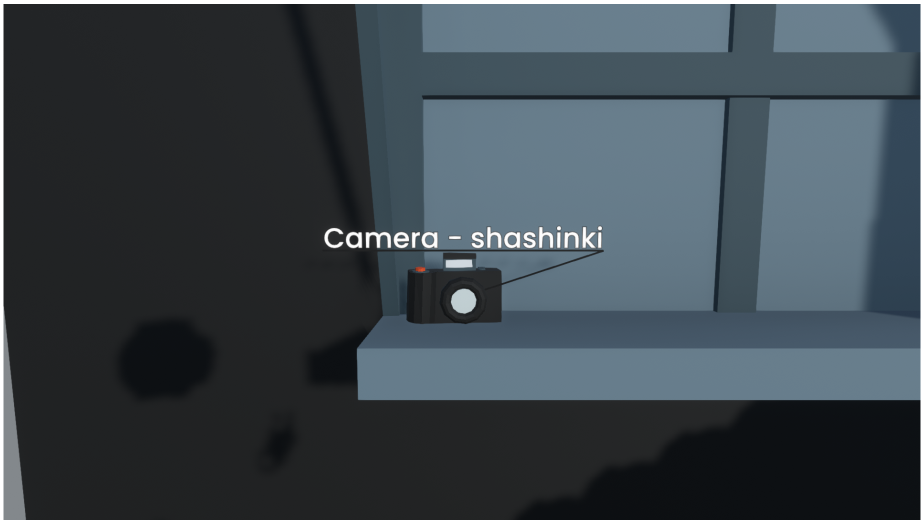

Figure 4: Depiction of an object, representing a word from the condition without gestures (A). When grabbed, it shows both the English and Japanese names of the object.

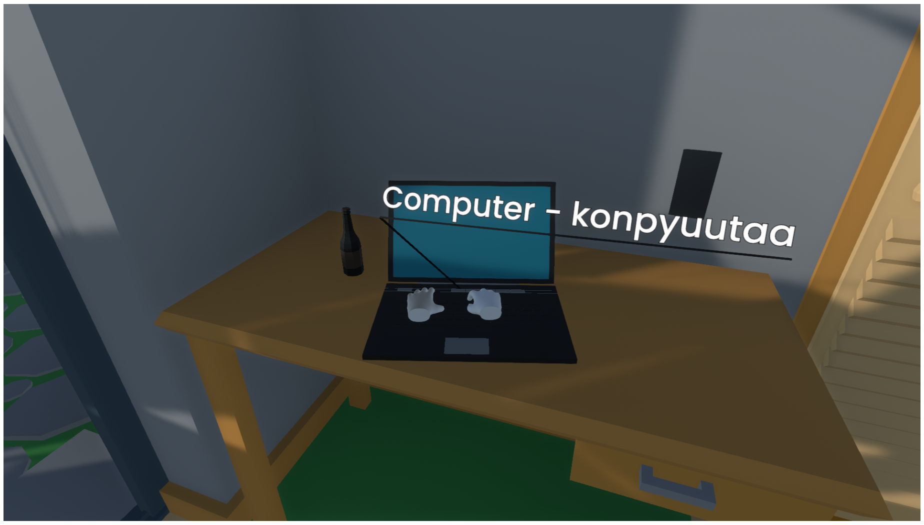

Figure 5: Depiction of an object, representing a word from the condition with gestures (B). When grabbed, it shows both the English and Japanese names of the object, as well as the gesture.

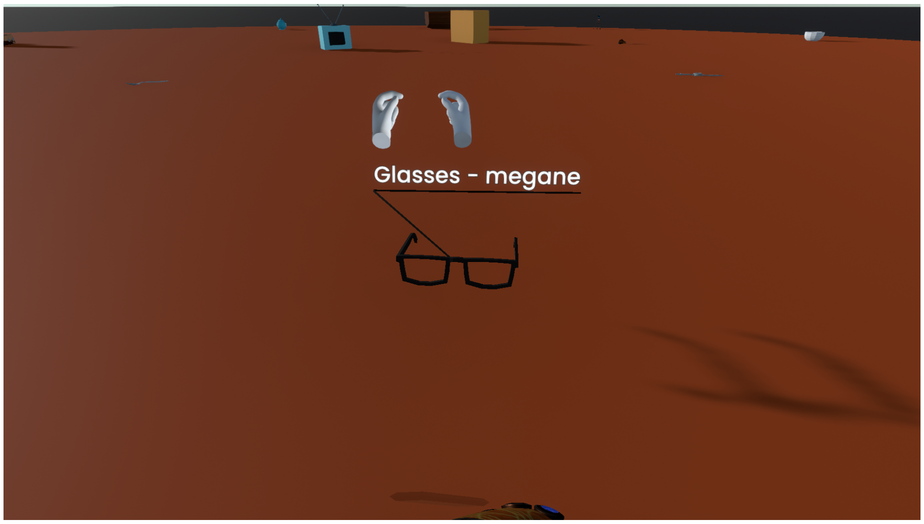

Figure 6: Depiction of an object, representing a word from the condition with gestures in the non-contextual world (C). When grabbed, it shows both the English and Japanese names of the object, as well as the gesture.

The system should be able to recognize the gestures performed by the participant via path-based collision detection. Due to the scope of the assignment and some setbacks with the setup, in the end, this detection system was wizarded. Based on observation, a button was pressed, after which the system played a voice recording informing the participant how the gesture was performed.

The voice recording either told the participants: 1) that the gesture was performed well and they can move on to the next object, or 2) that the gesture was not recognized, and should be repeated. If the system does not recognize the gesture 5 times, the system asks the participant to move on to the next word. This was done to make sure a participant would not get stuck on a certain word. Radcliffe et al. [14] reason it is better to give single-modal feedback, over something more multi-modal, to reduce the cognitive load. Once a gesture is recognized, the system does not play any additional voice recordings associated with that gesture. Since participants were expected to view the words multiple times, the voice recording was not played each time to confirm the gesture, as this could have been distracting and increased cognitive load.

Performing a gesture takes more time than just looking at the translation of a word. It is logical to assume that most participants are exposed longer to the words where they have to perform a gesture. This was counteracted by the fact that participants could come back to each word as much as they liked, which should result in them taking as much time as they need to learn each word. This was probably different for each word, based on the perceived difficulty of learning that word by the participant.

4 Methodology

Now that we have the design and differences clear between the conditions, I will explain the methodology for the user test.

The experiment consisted of two parts. In part 1, the participants ran through most of the experiment. They had three separate learning sessions that together spanned all of the words, immediately followed by a posttest after each condition on the words in that world. Part two only consisted of a delayed posttest, which was held one week after part one had taken place, on all words. Here, the words were mixed as explained in Section 3.2.

If a participant fails to finish the entire experiment for any reason, their data are removed. Participants are free to stop the experiment at any time without providing an explanation why. After a short explanation of the experiment, but before starting, participants were asked to read an information letter about the experiment and sign an informed consent form. This study has been approved by the Ethics Committee Computer & Information Science (EC-CIS) of the University of Twente under application number 250540.

In part one, there were three different conditions for the learning sessions that each participant ran through. This resulted in 6 different orders of the experiment. One of these orders was semi-randomly selected for each participant. The selection was semi-random since when one of the six was selected, this was eliminated for the next random draw, until only one option was left. Once that last order was also assigned, all orders would be possible again, this to ensure not one order was selected way more or less than others. In Table 3, one experiment is laid out in detail. The only difference between the rest is the order of sections A, B and C, which have the possible options: ABC, ACB, BAC, BCA, CAB, CBA.

During the experiment, the main researcher guided the participant through the experiment and made observations. A clear script of instructions to convey to the participant was followed, including, but not limited to: what to do, how to do it, and the time allowance. The full script is given in Appendix D. Note that the script does not contain fully written-out instructions, but it contains the key message. Participants had a maximum of 15 minutes in each learning session and 5 minutes per posttest. Both these times were chosen to allocate as much time as possible to these events while keeping the total experiment time between 1 and 1.5 hours.

Table 3: Outline of the different sessions the participant goes through, the conditions A, B, C were (semi-)randomized for order.

Part 1
Tutorial of Resonite
Learning session A
Objects are presented in a contextual world. No gestures need to be performed Questionnaires and written posttest
Learning session B
Objects are presented in a contextual world. Should perform gestures Questionnaires and written posttest
Leaning session C
Objects are presented in an empty world Should perform gestures Questionnaires and written posttest
Part 2 (1 week after Part 1)
Written posttest on all of the words

4.1 Participants

Participants were recruited through convenience sampling, primarily by approaching individuals within the researcher's personal network, and via the pre-university, a company at the University of Twente that gives many different kinds of workshops and masterclasses to students of all education levels before university. For this study, students from high school or university were selected from the age of 16+, as second language acquisition seems to be most prominent in high school. However, this is balanced with convenience by selecting participants who are 16+, so no parents need to be involved, which speeds up the process and likelihood of participation.

Participants were excluded from this study if they met one or more of the following criteria: 1) they are younger than 16, 2) they do not have full range of motion in their arms, or 3) they have extensive prior knowledge of the Japanese language. Convenience sampling was used to recruit participants, with the experiment running over a two-week period during which the setup was available in a lab setting for 8 hours per day. Based on previous studies, a maximum of 36 participants had been set, after which the study was concluded. Participation in this study was on a voluntary basis.

4.2 Equipment Set Up

The experiment took place in a room on the campus of the University of Twente. Here, a computer was set up with Resonite running. The language learning world was built in Resonite. The participant uses a VR Meta Quest 3 headset to navigate in Resonite.

In Resonite, the participant first entered a tutorial world, where the participant could familiarize themselves with the controls and the tasks in the experiment. After this, they went to the designated world for the experiment. This was either a 3D model of a bedroom with a garden for the contextual learning experience, or an empty world with only the objects relating to the to-be-learned words in it. A depiction of these worlds can be found in Appendix B.

In the given time frame for this research, it was more efficient not to implement the gesture recognition into the program. During the usertest, this functionality of the system was wizarded by the researcher. If a gesture was performed, the researcher could press a button on the keyboard to play an audio file from the system. This either said that the gesture was completed successfully or that the gesture should be tried again. The performance of the gestures was observed and judged by the researcher. Notes were made based on direct observation of the gestures performed. This was recorded to make sure the participant performed every gesture correctly at least once. to make sure the link between the gesture was formed.

4.3 Design and Analysis

Similarly to Bergsma [9], the posttests and the delayed posttest were graded with the following measures: A phonetically correct answer was given 1 point, for example, for the word for *juutan* –rug, both *juutan* and *jutan* would be counted as correct. The goal was not to try to measure how well they remembered the words, nor to check the Japanese spelling precisely. 0.5 points were awarded to a partially correct word, so if at most 1 syllable was wrong[2], or 2 letters had been switched. An example of such a mistake would be if *gomibako* –garbage bin, would be denoted as *gamiboko*. Wrong or blank answers received 0 points. A maximum of 18 points, one per word, could be scored for the normal posttests and 54 points for the delayed posttest. If, at the delayed posttest, the same incorrect answer was given as during the earlier posttest, this was still counted as incorrect. Scores and/or information about which words were remembered correctly were not shared with participants between sessions.

To compare the three conditions, a one-way repeated ANOVA was used for the immersion and cognitive load. After all, the test was repeated multiple times by one participant, so the data sets could not be assumed to be independent. The p-value that follows from this test only holds if there are equal variances of differences between the conditions. If this is the case, the normal p-value can be used; if not, a corrected p-value should be used. To test this, Mauchly's

[2] If a one-syllable long word was denoted incorrectly, it was awarded 0 points.

sphericity test was used. If sphericity was not met, the p-value was corrected using the Greenhouse-Geisser correction.

The language-learning scores of the posttest were compared using Friedman's ANOVA, since there were more than 2 conditions for which one entity (in this case, participant) provided data, and the data cannot be assumed to be continuous. The Friedman test counteracts unusual cases, like outliers in the data, so they do not need to be taken out of the dataset. The scores of the immediate posttest were able to tell us the effect on the short-term retention of the words. To compare the long-term retention of the participants, I was interested in the retention rate between the immediate and delayed posttest. Building on Bergsma [9], I define the retention rate as: $retention\ rate = delayed\ posttest\ score/immediate\ posttest\ score$.

5 Results

A total of 22 participants started the experiment, of which 16 were male and 6 were female. Due to dizziness or nausea, 3 experiments had to be cut short. Thus, a total of 19 (14m/5f) participants' data are used, as the data for participants 2, 7, and 18 were incomplete. All participants were between the ages of 18-27, and all had no proficiency in Japanese, excluding some very basic words, such as *sushi*, *konnichiwa*, or *arigatou*, words which were not part of this test.

There were 7 participants in the age range *18-20*, 6 were in the range *21-23*, and 6 were in the range *24-27*. In Table 4, a distribution of the number of people who did the order in a particular order, divided into the age ranges, is shown. Here, A refers to the contextual world, where no gestures are performed, B refers to the contextual world where gestures are performed, and C refers to the non-contextual world where gestures are performed. Of all the participants, 8 reported having had *0* VR experiences before, 7 reported *1–4*, 0 reported *5–9*, and 4 reported *10+* experiences. An overview of their distribution over the different condition orders is shown in Table 5.

Table 4: Division of participant ages over the conditions

ages	condition order of the experiment					
	ABC	ACB	BAC	BCA	CAB	CBA
18-20	1	1	2	1	1	1
21-23	0	2	0	2	1	1
24-27	2	0	1	0	2	1
total	3	3	3	3	4	3

Table 5: Division of participants' previous VR experience over the conditions

VR experience	order of the experiment					
	ABC	ACB	BAC	BCA	CAB	CBA
0	3	0	0	2	1	2
1–4	0	2	1	1	2	1
10+	0	1	2	0	1	0

The immersion of all participants was measured on a scale of -3 to 3. After admitting, this scale has been offset by 3, as intended by the creators [28]. The new scale is thus from 0 to 6, with 3 as the middle value. To put the results of the Igroup presence questionnaire into numbers, the average is taken from the answers to all questions. This average value is interpreted as the immersion felt by a participant for a certain condition. Table 6 shows the immersion of all participants per condition, as well as the mean immersion experienced. The immersion experienced per condition is significantly different, $F(2, 36) = 16.04$, $p < 0.0001$, $\omega = 0.66$. This p-value can be used as sphericity is met, since the

Mauchly test showed $W = 0.79$, $p > 0.05$. Pairwise comparisons were executed as follow-up tests, showing there is no significant difference between conditions A and B, $t(18) = -0.06$, $p > 0.05$ (one-tailed), $r = 0.01$. There is a significant difference between conditions A ($mean = 3.3$, $std = 0.63$) and C ($mean = 2.5$, $std = 0.84$), $t(18) = 4.16$, $p < 0.001$ (one-tailed), $r = 0.70$, and between B ($mean = 3.3$, $std = 0.77$) and C ($mean = 2.5$, $std = 0.84$), $t(18) = 4.84$, $p < 0.001$ (one-tailed), $r = 0.75$.

Table 6: Immersion felt per participant per condition, as well as the mean immersion per condition.

participant	condition			participant	condition		
	A	B	C		A	B	C
1	3.0	3.0	2.8	13	2.9	3.7	2.1
3	3.9	3.9	2.6	14	4.8	3.9	2.4
4	3.2	3.4	3.5	15	3.7	3.7	2.2
5	3.6	2.6	1.8	16	3.5	3.6	2.7
6	3.7	5.4	4.4	17	4.1	3.8	3.1
8	1.8	1.9	1.4	19	3.2	3.3	2.7
9	3.5	3.4	1.9	20	2.8	2.3	2.4
10	3.6	3.8	2.8	21	2.9	2.3	2.4
11	3.0	3.3	2.9	22	3.1	3.4	0.3
12	2.9	2.8	2.6	mean	3.3	3.3	2.5
				std	0.63	0.77	0.84

The NASA-TLX determines the cognitive load of the participant as a number between 0 and 100, where 50 is the median. This number thus represents the cognitive load experienced by a participant during each condition. Table 7 shows the cognitive load experienced by all participants per condition, as well as the mean cognitive load experienced. As sphericity is not met, $W = 0.47$, $p < 0.05$, the p-value of the ANOVA needs to be corrected. The Greenhouse-Geisser correction is used for this, which gives $F(2, 36) = 0.91$, $p > 0.05$. This shows that there is no significant difference in the cognitive load experienced between conditions. The means and standard deviations of the three conditions are: A, $mean = 56$, $std = 16$; B, $mean = 56$, $std = 13$; C, $mean = 59$, $std = 13$.

Table 7: Cognitive load experienced per participant per condition, as well as the mean cognitive load per condition.

participant	condition			participant	condition		
	A	B	C		A	B	C
1	70	61	64	13	47	52	58
3	63	59	44	14	33	40	59
4	70	61	54	15	83	62	54
5	50	65	67	16	48	54	57
6	69	75	63	17	47	51	53
8	52	49	67	19	52	50	66
9	71	60	64	20	22	23	33
10	43	37	34	21	78	74	69
11	53	60	57	22	69	77	91
12	45	52	70	mean	56	56	59
				std	16	13	13

Table 8 shows the scores each participant had for the immediate posttest. The maximum score that could have been gotten was 18 points. The results of the posttests differed significantly across conditions, $\chi^2(2) = 10.6$, $p < 0.01$. A Wilcoxon test was used for the follow-up tests. These tests showed that between conditions A and B, there is no significant difference, $r = -0.27$, $p > 0.05$. There is a significant difference between conditions A ($mean = 10.9$, $std = 3.9$) and C ($mean = 8.4$, $std = 4.6$), $r = 0.57$, $p < 0.05$, and between conditions B ($mean = 11.5$, $std = 4.6$) and C ($mean = 8.4$, $std = 4.6$), $r = 0.66$, $p < 0.01$.

Table 8: Results of the immediate posttest per participant per condition, as well as the average score per condition.

participant	condition			participant	condition		
	A	B	C		A	B	C
1	4	7	3	13	18	14.5	17
3	4	6	1.5	14	11.5	10.5	9
4	12	14.5	8	15	5.5	11	8.5
5	12	14.5	9	16	10.5	11.5	5
6	10.5	15.5	17	17	7.5	1	5.5
8	13.5	16.5	15.5	19	12	18	13
9	16.5	14.5	8	20	12.5	16	11
10	15.5	5	8.5	21	7	8.5	2.5
11	9.5	8	5.5	22	13.5	15.5	8.5
12	11	11	4	average	10.9	11.5	8.42
				std	3.92	4.60	4.63

Except for 3 participants, all participants took the delayed posttest on pen and paper on the campus of the University of Twente. Due to logistical reasons, including a strike of the national railway, participants 3, 4, and 12 were unable to return to the campus. The same document that was printed for the delayed posttest was sent as a Microsoft Word document to these participants, to still include their data in the results of the study. A read-me file was sent alongside this document, with the request to read it before starting and adhere to the instructions stated in the document, which all participants indicated having done. The read-me file can be found in Appendix E. The data of these participants seems to be in line with the other data collected, and thus will be used.

Table 9 shows the results of the delayed posttest. The calculated retention rate is given in Table 10. We can see that participant 17 has a retention rate of 4.00 or 400%, for condition B. In the initial test, they only had one correct answer, while in the delayed test, they had 4 correct answers. They noted that during the immediate posttest, they were a bit nervous and chaotic, and therefore did not remember more words. They started the experiment with condition B as well. During the delayed posttest, they felt more at ease, so they spent more time trying to remember the words. The average retention rate for condition B drops from *0.69* to *0.50* if this outlier is not taken into account. The standard deviation drops from *0.83* to *0.22*.

The retention rates between the 3 conditions differ significantly, $\chi^2(2) = 8.84$, $p < 0.05$. The Wilcoxon follow-up tests showed that between conditions B ($mean = 0.69$, $std = 0.83$) and C ($mean = 0.39$, $std = 0.31$), the results were significantly different, $r = 0.59$, $p < 0.01$. Between conditions A and B, or between conditions A and C, no significant difference was found, $r = -0.42$, $p > 0.05$, and $r = 0.21$, $p > 0.05$, respectively.

Table 9: Results of the delayed posttest per participant per condition, as well as the average score per condition.

participant	condition			participant	condition		
	A	B	C		A	B	C
1	4	6.5	2	13	10.5	8	5
3	0.5	2	0	14	4.5	7.5	7.5
4	3.5	10	6	15	3	4.5	2.5
5	3.5	2.5	1	16	5	5	3
6	5.5	9	7.5	17	2.5	4	5.5
8	7.5	9.5	5.5	19	5	5	2
9	9	11	3	20	5.5	11	4.5
10	7.5	2.5	0	21	1	3	2
11	4	4.5	0.5	22	2.5	0.5	0
12	3	5.5	1	average	4.61	5.87	3.08
				std	2.51	3.11	2.43

Table 10: Retention rate per participant per condition, as well as the average retention per condition.

participant	condition			participant	condition		
	A	B	C		A	B	C
1	1.00	0.93	0.67	13	0.58	0.55	0.29
3	0.13	0.33	0.00	14	0.39	0.71	0.83
4	0.29	0.69	0.75	15	0.55	0.41	0.29
5	0.29	0.17	0.11	16	0.48	0.43	0.60
6	0.52	0.58	0.44	17	0.33	4.00	1.00
8	0.56	0.58	0.35	19	0.42	0.28	0.15
9	0.55	0.76	0.38	20	0.44	0.69	0.41
10	0.48	0.50	0.00	21	0.14	0.35	0.80
11	0.42	0.56	0.09	22	0.19	0.03	0.00
12	0.27	0.50	0.25	average	0.42	0.69	0.39
				std	0.20	0.83	0.31

In Table 11, the time spent per condition per participant is shown in minutes. A repeated ANOVA[3] test was run between the conditions, where no significant difference was found, $F(2, 36) = 1.36$, $p > 0.05$. The normal p-value cannot be used as sphericity is met, since the Mauchly test showed $W = 0.62$, $p < 0.05$, so the Greenhouse-Geisser correction has been used.

Table 11: Time spent per participant per condition in minutes, as well as the average time spent per condition.

participant	condition			participant	condition		
	A	B	C		A	B	C
1	6	10	7	13	6	9	7
3	9	8	12	14	9	15	9
4	15	14	15	15	13	11	13
5	15	13	14	16	9	12	10
6	10	11	13	17	8	9	11
8	15	13	15	19	15	15	15
9	15	15	15	20	11	15	14
10	13	8	14	21	15	15	15
11	6	11	9	22	14	13	12
12	12	15	11	average	11.4	12.2	12.2
				std	3.4	2.6	2.7

[3] Even though the data is not continuous, an ANOVA test was used. Since no significant difference was found with the assumption of continuous data, it is very unlikely that a significant difference would be found by another test. It is advised not to use an ANOVA test for future applications.

6 Discussion

In this research, I built an interactive IVR environment. There were 19 participants who studied 54 Japanese words over three conditions to answer the research question of this study:

To what extent can immersion be increased, while maintaining a reasonable cognitive load, by 1) performing intuitive gestures and 2) learning in a contextual world, and does this enhance noun acquisition in immersive virtual reality?

With the following hypotheses:

H_0 The effects of immersion in the environment and gesture performance are insignificant between conditions. Thus, no significant difference is found between the three conditions in short-term word acquisition or long-term retention.

H_a The gained immersion due to both the contextual world and performance of intuitive gestures has a larger positive effect than the drawback of increased cognitive load. Thus, both the short-term acquisition of the words and long-term retention will be aided most if both are present.

The results show that for the short-term VA, the performance of the condition in the contextual world (B) is better than for the non-contextual condition (C). The results suggest that the increased immersion due to the contextual world positively affects learning foreign words in VR. For long-term retention, we see he same. Condition B shows better results when comparing the conditions in a contextual world (B) or a non-contextual world (C). This suggests that the increased immersion due to the contextual world positively affects learning foreign words in VR. No significant difference could be found between conditions A and B, either for the short-term VA or long-term retention.

The results do not show a significant difference in experienced cognitive load between the 3 conditions. Therefore, we cannot conclusively say that immersion has a larger positive effect than the negative effect of the cognitive load, since no difference was found between the cognitive load of the conditions at all, and the immersion was the same for both conditions in the contextual world (A and B). We also cannot state that the performance of gestures significantly increases learning, based on the results. So, H_a is rejected, and H_0 is accepted, even though condition B performed better than C in both the short and long term, which implies that a contextual world has positive effects on both short-term VA and long-term retention.

6.1 Performance of gestures

Before the main research question of the study can be answered, the 6 subquestions have been answered first. The answers to these questions are outlined in the first 3 sections, starting with sub-questions 1 and 2:

1) To what extent does performing intuitive gestures influence the immersion?

With hypotheses:

H_0 Performing intuitive gestures does not influence the immersion.

H_1 Performing intuitive gestures positively affects the immersion.

Between conditions A and B, where in condition A, no gestures were performed, and in condition B, gestures were performed, there was no significant difference measured in the immersion experienced, $p = 1.0$. Therefore, H_0 is accepted, and H_1 is rejected.

2) To what extent does performing intuitive gestures influence the cognitive load?

H_0 Performing intuitive gestures does not influence the cognitive load.

H_2 Performing intuitive gestures positively affects the cognitive load.

For the cognitive load, no significant difference was found for any of the conditions, $p = 0.38$, so H_0 is accepted and H_2 is rejected.

For both sub-research questions, the null hypothesis was accepted. Therefore, we can say that either 1) performing gestures or 2) not performing gestures does not influence the cognitive load or immersion in the VR environment. It can be easily explained that the immersion is not as dependent on the performance of gestures as it is on the surrounding contextual world. However, the fact that the performance of gestures does not increase the cognitive load significantly is unexpected. Some observations from the user tests are presented that could help explain these results.

Not only were the objects that should have been interacted with, representing the target words, movable in the VR world, but everything else was too. In other studies [4, 13], only the words that should be learned could be manipulated, or nothing at all [9]. Due to the way I set up the worlds in Resonite, this could not be changed without making the entire world even more breakable. Within the time-scope of the study, it was not possible to adjust this for someone new to Resonite. What set the target words apart from other objects was the fact that when a target word was moved, it would stay in the new position, while other objects would snap back into place.

It was quite easy for participants to miss a target word and grab and move a normal object, which could be very distracting. Every participant accidentally grabbed and moved the entire floor in condition C, the non-contextual world, at least once. In the contextual words, participants could move the walls of the house, which each participant also did. The distractions caused by this could have had a big effect on the cognitive load. Which would, in turn, make the influence the gestures had on the cognitive load smaller.

Besides the participants for whom the experiment was cut short, several other participants mentioned feeling a bit dizzy or experiencing some nausea after the experiment. This was mostly due to the quick, snappy nature of avatar movement in Resonite. The participants did the experiment while standing up, using normal walking locomotion in Resonite. All participants used a combination of physically turning and joystick turning to look around. They did vary in how much they employed each tactic. Participants indicated that they had to concentrate on not feeling nauseous and trying to continue with the experiment. This could also explain a part of the cognitive load experienced, which decreases the influence the gestures had on the cognitive load.

6.2 Contextual VR environment

Sub-questions 3 and 4 revolve around the impact that conditions have on the immersion in the VR environment. They are shown here, alongside their respective hypotheses.

3) To what extent does a contextual VR world influence the immersion?

> H_0 A contextual VR environment does not influence the immersion.
>
> H_3 A contextual VR environment positively affects the immersion.

4) To what extent does a contextual VR world influence the cognitive load?

> H_0 A contextual VR environment does not influence the cognitive load.
>
> H_4 A contextual VR environment positively affects the cognitive load.

The results clearly show that condition C, the non-contextual world, led to a lower immersion than B, $p = 0.00037$. For sub-question 3, the null hypothesis is rejected, and H_3 is accepted. As has been stated in the previous paragraph, the cognitive load does not show any significant difference between any of the conditions, $p = 0.38$. Therefore, H_0, is accepted and H_4 is rejected for sub-RQ 4.

As was expected, the immersion is lowest for the non-contextual world. Interestingly, the performance of gestures appears to have no impact on the level of immersion felt in the VR environment, $p = 1.0$. In each condition, also the one where no gestures were needed, object manipulation was possible. This might have had a bigger impact on the immersion felt than performing the gestures. In other studies that measured the difference between learning with performing gestures or learning without gestures, object manipulation was only possible in the conditions where gestures were performed [4, 13]. This could explain why no difference was found in the immersion in the VR environments between conditions A and B.

The fact that no significant difference is found between the cognitive load experienced in the conditions is explained in the previous section. The same reasoning still holds for the difference between contextual and non-contextual worlds, as between either performing or not performing gestures.

From the first four sub-research questions, the null hypothesis was accepted for all except sub-question 3. So, the only significant difference found between the three conditions is that the immersion in the virtual world is higher if the participant is in a contextual world. The performance of gestures or not does not seem to influence the immersion experienced. The results of the experiment seem to suggest that the cognitive load is not affected by the performance of gestures or by being in a contextual world. However, due to the setup of the experiment and the clarifications provided in the previous section, this cannot be stated with a hundred percent certainty.

6.3 Vocabulary acquisition

The final sub-questions, 5 to 8, needed to answer the main research question revolve around the impact of the immersion and cognitive load on the short-term vocabulary acquisition, and long-term retention of the participants. They are shown here, alongside their respective hypotheses.

5) To what extent does performing intuitive gestures influence the short-term vocabulary acquisition?

> H_0 Performing intuitive gestures does not affect short-term vocabulary acquisition.
>
> H_5 Performing intuitive gestures positively affects short-term vocabulary acquisition.

6) To what extent does a contextual VR world influence the short-term vocabulary acquisition?

> H_0 A contextual VR environment does not affect short-term vocabulary acquisition.
>
> H_6 A contextual VR environment positively affects short-term vocabulary

acquisition.

7) To what extent does performing intuitive gestures influence the long-term word retention?

H_0 Performing intuitive gestures does not affect long-term word retention.

H_7 Performing intuitive gestures positively affects long-term word retention.

8) To what extent does a contextual VR world influence the long-term word retention?

H_0 A contextual VR environment does not affect long-term word retention.

H_8 A contextual VR environment positively affects long-term word retention.

The results of the immediate posttest also show slightly higher VA for B than for A, but the difference is not significant. This suggests that the performance of gestures does not affect vocabulary acquisition. H_0 is accepted, while H_5 is rejected. A significant difference is shown between conditions B and C, so H_6 is accepted and H_0 is rejected. The fact that no difference was found between conditions A and B, where the only difference is the performance of gestures, might be explained by the immersion and the cognitive load. Since neither of these shows a significant difference between the conditions. This can not be stated conclusively as the correlation was not investigated in this study.

The results show that the immediate learning effects of condition B are significantly better than those of conditions C. Xie et al. [5] show that in a VR environment, as opposed to audiovisual learning outside of VR, only the long-term retention is improved, while the immediate learning effects are worse. The negative impact of VR is still assumed to affect the results of my study. However, the results suggest that the increased immersion has a positive effect on immediate learning as well. If this effect is big enough to cancel out the negative effect of VR is unclear.

When comparing the retention rates, we see that condition B, the contextual world where gestures are performed, has a significantly higher retention rate than condition C, the non-contextual world. We also see that condition B has a higher retention score than condition A, the contextual world where no gestures were performed, but this is not a significant difference. Thus, hypothesis H_7 is accepted and H_0 is rejected. H_8 is rejected and H_0 is accepted.

There was no difference measured in the immersion between the two conditions in the contextual world: with (B) or without gestures (A). This implies that the impact of the performance of gestures on the immersion is not significant.

Interestingly, condition B did not show any significant increase in the short-term VA or the long-term retention, compared to A. Research suggests that the performance of gestures increases VA, in short-term VA [4, 13], which contradicts my results. In the section 6.4, I will outline some possible reasons for the contradictory results, as well as noteworthy observations and insights derived from the user study.

6.4 User study observations

Bergsma [9] showed a range 69% - 75% retention after one week. Comparing this rate to my data, we see that only condition B has such a high retention, of 69%. When taking the outlier from this data that retention also drops to 50%, meaning the average retention for all conditions is lower than her range. I identified two main reasons why this might be the case. The first is the number of words. Bergsma tested a total of 32 Japanese words, while I tested a total of 54 words. In the same time-frame, participants needed to remember more words, which is more difficult. The second reason is the fact that participants only had one learning session for the words in my study, and her participants were exposed to each word three times.

The participants had very different innate learning abilities, which could explain the differences in scores obtained on the posttest. Some participants only had a few points on a learning condition, while others had (almost) every answer right. The innate learning ability of the participants seemed to have a much higher influence on the results than the experienced cognitive load did. This might cause the cognitive load to be less significant to the overall learning process than previously assumed. Because I had a within-design user study, the big differences in innate learning ability were reduced. The downside of this design is that the cognitive load varies over time, which might have influenced the results.

Some participants took the full 15 minutes to learn all the words, while some indicated they knew all the words before 10 minutes had passed. A full overview of the time each participant spent in each learning session is presented in Table 11. Some participants who stayed relatively short in the learning conditions, for example, participant 1, scored quite low on the posttest. Others scored very high, such as participant 13. Others stayed relatively long and had high scores, participant 19, or scored low, participant 15. While there seems to be no correlation between the two variables, this relationship has not been investigated and therefore cannot be confirmed.

About 6 or 7 participants moved the objects around a lot when learning them, even collecting them all in one or a couple of locations. These 'collectors', as I call them, indicated that they liked to put all the objects together to go over them in a more structured way. Most participants kept the objects mostly in the place where they initially were in the world. Several participants even went out of their way to move objects back to their initial place if they accidentally moved them. A less common learning strategy was standing in a central location where all objects could be seen and studied, which 2 participants used. The advantage of having the objects be movable is that all participants can employ their own learning strategy and learn in a way that works best for them personally.

In some cases, it was impossible to let the objects be the trigger for the system to activate the learning words. An invisible box was placed over these objects to overcome this issue. When making the world, I did not expect the participants to move the objects around so much, as object manipulation was initially planned not to be a part of this study. Due to this, for some objects, object manipulation was not possible. A few participants were thrown off by this at first, but they all quickly changed strategies, such as moving all objects towards the one that could not be moved, or making two separate collections of the objects. However, some participants forgot about these words completely after seeing them initially, as they could not move them. This limitation of the system might have influenced the results in unexpected ways.

Besides these learning strategies, participants also had very different approaches to the gestures. Some did the gesture only when initially learning a word, and then stopped using it. While others repeated them each time when learning a word. Most participants fell somewhere in between these two extremes. Participants also repeated words while they did not pick up the object again, to check whether they were right or not. This resulted in them 1) not seeing the gesture every time and weakening the link between the word and gesture, and 2) sometimes they misremembered a word and continued learning the incorrect translation.

One participant even commented on the difference in learning between the two conditions with gestures. They (participant 11) stated that in condition C, the non-contextual world, they focused on remembering the gestures more than they focused on the words. In condition B, the contextual world, they used the gestures as a tool to remember the words. The unfamiliarity and maybe even discomfort of certain participants with performing gestures could be seen as they performed the gestures quite subtly. This was quite clear in participant 13's case. They grew more comfortable with the gestures during the experiment and began to do them more overtly.

While most gestures were immediately clear to participants, some were a bit more vague. This was due to a couple of different reasons. The gestures for vase, *tracing the outline of the vase in the air*, and for telescope, *taking the telescope and pointing it at the sky*, were relatively to the other gestures, quite long. Quite a few participants had to try to do these gestures several times. Since they did not watch the entire preview of the gesture, they oversimplified the gesture and did not perform the gesture in enough detail. The oversimplified gestures were not as clearly related to the objects, and therefore could weaken the connection between the gesture and the words.

Other gestures missed a frame of reference. A clear example of this is the gesture for bed, *tilting your hands to the side to lay your head upon them.* Since the animation consisted of only the hands and not the head, some participants interpreted the gesture differently than intended. Many participants interpreted the gesture as jumping into bed. The same can be said for the gestures for sun and eye. The gestures, *blocking the sun from your eyes with your hand* and *pointing at your eye*, respectively, also missed the head as a frame of reference. Participants found their own interpretations of these gestures. Participant 4 stated that the gesture for *sun* was logical: 'the sun setting on the horizon'. This is not seen as an issue as participants gave their own, new meanings to the gestures. Since for them, there was still a logical connection to the word, it should not impair the learning results.

Sun et al. [12] suggest that overly extensive gestures may increase cognitive load. However, since no significant differences in cognitive load were found between the gesture and no-gesture conditions, we can infer that the gestures used in this study were not overly demanding. While other factors, as discussed previously in the section, may also have played a role, their presence across all conditions makes it difficult to measure their direct impact. The findings suggest that the gestures were not too extensive.

The delayed posttest was planned to be 7 days after the original test for each participant. Due to the busy schedules of the participants, this was harder to achieve was anticipated. To stay as close as possible to the planned delay, only an offset of one day was allowed. For 10 participants, the delayed posttest did take place after *7* days. For 6 participants, it was after *6* days, and for 3 participants, it was after *8* days. By looking at the retention rates of these participants, no clear pattern emerges, suggesting the impact of this difference is negligible.

When the participants were learning the words, they saw the words represented by 3D objects, as well as the English and Japanese words for them. In the posttests they were only given the English word, and not a picture or other visual representation of the word. Participants noted that they knew what the objects representing the words looked like exactly, but not what the Japanese translation of the word was. Participant 22 stated they knew the exact location and colors of the objects, but not their names. Having visual representations of the object as well might result in having higher scores overall, as literature also suggests that staying in VR for the posttests has a positive effect on learning [20].

Many participants reported signs of motion sickness, and 3 participants dropped out for this reason. The contextual world was also used by Bergsma [9] in her study, where no participants needed to cut short the experiment due to motion sickness. In her study, she did not have a time limit per learning session, while I did set a time limit of 15 minutes. It is unclear where this difference in perceived motion sickness arises from. Perhaps the increased motion caused by performing the gestures resulted in more disorientation in participants. Participants noted that they experienced the dizziness most when walking through the world, but that standing still was fine. Though the reasons were very different, the dropout rates were quite similar: 4/26 for Bergsma and 3/22 for me.

6.5 Limitations

In their study, Essoe et al. [3] compare the language learning of two separate languages, in either two different or the same virtual setting. Their results suggest that learning the two separate languages in two separate worlds is more effective, under the condition that the participant feels a high sense of presence in the world. It is unclear from their study how many of the participants reported a high presence. For this reason, it is unclear how accurate their conclusion is. Since their study centers on a different subject than what I recorded, it is unlikely to have a significant impact on the findings reported here.

In my design, I specifically did not use an audio recording of the word that should be learned to not have too much information at the same time: the English word, the Japanese word, the intuitive gesture, and the pronunciation of the Japanese word. This would have been unclear and increased the cognitive load. The disadvantage of this is that having to come up with a pronunciation on your own does increase the cognitive load [5]. To convey to the participant how the gesture was performed, a voice recording was used. This was deemed to be clearer and less invasive than a visual cue in the world, given that a visual cue would have needed to be quite big. Participants could approach the same word from any direction, so a visual cue would need to stand out from every angle, and be sure not to clip into a wall or other object.

The words have been divided into the 3 lists by the main researcher, based on their perceived difficulty of these words. These word lists have not been peer reviewed based on their relative difficulties. To lessen any subconscious biases that may have been introduced by this, the word lists have been randomly assigned to the three conditions. While some participants noted that the words in condition A were much more difficult than the other lists, others said the same thing about the words in conditions B or C. Based on this, it seems that the lists were of similar difficulty. One solution to negate the effect of these biases even further would have been to randomly assign each list to a condition on a participant basis, instead of having the same list per condition for all participants. This implementation would be very time-inefficient and was not possible in the scope of this study, as it would require a multitude of contextual worlds.

6.6 Future Research

The most interesting thing to further investigate is the influence of performing gestures on the learning effects. No significant difference in immersion, cognitive load, or learning effects was found between conditions A and B, where gestures either were not performed or were performed. This contradicts existing research that does suggest a significant short-term learning effect [4, 13], implying that something else caused the learning outcomes of A and B to not be significantly different. It would be interesting to measure brain activity during both conditions to identify which areas of the brain are most active, providing insight into potential reasons.

This study focused on the effects of performing intuitive gestures on the cognitive load and immersion. The result of this study suggests that the immersion and cognitive load do not differ between conditions where either a gesture is performed when learning (B) or where it is not performed (A). A plausible reason no such difference has been shown could be the fact that in both conditions, object manipulation was possible, which was not the case in previous research [4, 13]. It would be interesting to investigate the influence of object manipulation on language learning in IVR even more.

Another facet I did not investigate was the effect of a low-poly world (which I used) on the cognitive load, immersion, and language learning. A more realistic VR environment might cause a higher cognitive load, but also might lead to higher immersion. It would be interesting to investigate the effect of this on the VA.

Finally, there are some conditions for the posttest that might be interesting to adjust. I only looked at the cued recall from English to Japanese, where the cues were the English words. Other research also tested the cued recall from the foreign language to the native language [3, 23] and even recognition tests where word pairs need to be matched [4, 24]. Other research also tested the words with visual representations of the words, 2D pictures [4], or 3D objects [9]. It would be interesting to see the effects that an immersive world and the performance of gestures has on other learning metrics, such as recognition in stead of recall.

7 Conclusion

In this study, I looked at the effect of the performance of intuitive gestures, as well as learning in a contextual world, on the immersion in the virtual environment and experienced cognitive load. The goal is to look at the effect of the immersion and cognitive load on the short-term vocabulary acquisition and long-term word retention of Japanese words in IVR. A user study has been employed where participants went through 3 separate conditions: A) a contextual world where gestures did not need to be performed, B) a contextual world where gestures should be performed, and C) a non-contextual world where no gestures needed to be performed. Immediately after each condition, they made a posttest for the words they remembered. A week later, a delayed posttest was conducted.

An immersive VR world was built where the participants went through all the conditions. The order of these conditions was randomly assigned to each participant. Each environment contains 18 unique objects representing target words that the participant needs to study. Upon selection, objects showed the English and Japanese names of the object, as well as the gesture associated with the word (in conditions B and C). The participants had 15 minutes to explore the environment and learn the words.

The results of this study indicate that there is no significant difference in either short-term vocabulary acquisition or long-term retention between conditions A and B, where in condition A, no gestures were performed, and in condition B, gestures were performed. On the other hand, between conditions B and C, condition B seems to be significantly higher in both short-term and long-term learning. Condition B is a contextual world, while condition C is a non-contextual world. So, it seems that the immersion or other effects of the context have a positive effect on learning, as being in a contextual world, versus a non-contextual world, has a significant effect on the immersion.

The results suggest that between the conditions, the cognitive load does not differ significantly. This implies that the cognitive load has not significantly changed, whether a gesture is performed or not, or whether the environment is contextual or non-contextual. This seems to suggest that neither learning in a contextual world nor the performance of gestures significantly increases the cognitive load while learning in VR.

We can conclude that the contextual environment has a positive effect on learning in IVR, as opposed to the non-contextual environment. The performance of gestures does not seem to help in the short-term VA and long-term retention. Further research should be conducted to find what causes this.

References

[1] Duolingo, *Duolingo research, about us.* [Online]. Available: `https://research.duolingo.com/#:~:text=Careers-,About%20Us,at%20scales%20never%20before%20seen.`.

[2] H. P. Bahrick, L. E. Bahrick, A. S. Bahrick, and P. E. Bahrick, "Maintenance of foreign language vocabulary and the spacing effect," *Psychological Science*, vol. 4, no. 5, pp. 316–321, 1993. DOI: `10.1111/j.1467-9280.1993.tb00571.x`.

[3] J. K.-Y. Essoe, N. Reggente, A. A. Ohno, Y. H. Baek, J. Dell'Italia, and J. Rissman, "Enhancing learning and retention with distinctive virtual reality environments and mental context reinstatement," *npj Science of Learning*, vol. 7, no. 1, 2022. DOI: `10.1038/s41539-022-00147-6`.

[4] O. Fuhrman, A. Eckerling, N. Friedmann, R. Tarrasch, and G. Raz, "The moving learner: Object manipulation in virtual reality improves vocabulary learning," *Journal of Computer Assisted Learning*, vol. 37, no. 3, pp. 672–683, 2021. DOI: `10.1111/jcal.12515`.

[5] T. Xie, H. Zhang, and Y. Yang, "Effect of immersive virtual reality based upon input processing model for second language vocabulary retention," *Education and Information Technologies*, 2025. DOI: `10.1007/s10639-025-13333-x`.

[6] T. Jaganov, C. L. Nnadi, and Y. Watanobe, "The learning labyrinth: Integrating learning theories in vr," in *Proceeding of the 2024 5th Asia Service Sciences and Software Engineering Conference*, ser. ASSE '24, 2025, pp. 158–165. DOI: `10.1145/3702138.3702156`.

[7] A. Ferko, Z. Berger Haladova, and M. Batorova, "Enhancing accessibility with informed vr and ar authoring for hybrid geometry learning," ser. DSAI '22, 2023, pp. 67–72. DOI: `10.1145/3563137.3563166`.

[8] I. Cuber *et al.*, "Examining the use of vr as a study aid for university students with adhd," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ser. CHI '24, 2024. DOI: `10.1145/3613904.3643021`.

[9] T. Bergsma, *Vocabulary acquisition in new and learned contexts using immersive virtual reality*, Jan. 2022. [Online]. Available: `http://essay.utwente.nl/89400/`.

[10] L. W. Barsalou, "Grounded cognition," *Annual Review of Psychology*, vol. 59, pp. 617–645, 2008. DOI: `10.1146/annurev.psych.59.103006.093639`.

[11] X. Xu, J. Kang, and L. Yan, "Understanding embodied immersion in technology-enabled embodied learning environments," *Journal of Computer Assisted Learning*, vol. 38, no. 1, pp. 103–119, 2022. DOI: `10.1111/jcal.12594`.

[12] Y. Sun, S. Pandita, J. Madden, B. Kim, N. Holmes, and A. S. Won, "Exploring interaction, movement and video game experience in an educational vr experience," Association for Computing Machinery, 2023. DOI: 10.1145/3544549.3585882.

[13] J. Ratcliffe, N. Ballou, and L. Tokarchuk, "Actions, not gestures: Contextualising embodied controller interactions in immersive virtual reality," in *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, ser. VRST '21, 2021. DOI: 10.1145/3489849.3489892.

[14] J. Ratcliffe and L. Tokarchuk, "Rich virtual feedback from sensorimotor interaction may harm, not help, learning in immersive virtual reality," in *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*, ser. VRST '22, 2022. DOI: 10.1145/3562939.3565633.

[15] U. of Bamberg, *Puresuggest.* [Online]. Available: https://fabian-beck.github.io/pure-suggest/.

[16] E. Christou, A. Parmaxi, and M. Christoforou, "Implementation and application of extended reality in foreign language education for specific purposes: A systematic literature review," *Universal Access in the Information Society*, 2025. DOI: 10.1007/s10209-025-01191-w.

[17] X. Huang, D. Zou, G. Cheng, and H. Xie, "A systematic review of ar and vr enhanced language learning," *Sustainability*, vol. 13, no. 9, 2021. DOI: 10.3390/su13094639.

[18] G. Makransky and G. Petersen, "The cognitive affective model of immersive learning (camil): A theoretical research-based model of learning in immersive virtual reality," *Educational Psychology Review*, vol. 33, pp. 937–958, 2021. DOI: 10.1007/s10648-020-09586-2.

[19] J. Ratcliffe and L. Tokarchuk, "Presence, embodied interaction and motivation: Distinct learning phenomena in an immersive virtual environment," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20, 2020, pp. 3661–3668. DOI: 10.1145/3394171.3413520.

[20] M. H. Lamers and M. Lanen, "Changing between virtual reality and real-world adversely affects memory recall accuracy," *Frontiers in Virtual Reality*, vol. 2, 2021. DOI: 10.3389/frvir.2021.602087.

[21] M. J. Wälti, D. G. Woolley, and N. Wenderoth, "Reinstating verbal memories with virtual contexts: Myth or reality?" *PLOS ONE*, vol. 14, no. 3, pp. 1–20, Mar. 2019. DOI: 10.1371/journal.pone.0214540.

[22] C. Vázquez, L. Xia, T. Aikawa, and P. Maes, "Words in motion: Kinesthetic language learning in virtual reality," in *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, 2018, pp. 272–276. DOI: 10.1109/ICALT.2018.00069.

[23] M. Macedonia, K. Müller, and A. D. Friederici, "The impact of iconic gestures on foreign language word learning and its neural substrate," *Human Brain Mapping*, vol. 32, no. 6, pp. 982–998, 2011. DOI: `10.1002/hbm.21084`.

[24] M. Macedonia, A. Lehner, and C. Repetto, "Positive effects of grasping virtual objects on memory for novel words in a second language," *Scientific Reports*, vol. 10, 2020. DOI: `10.1038/s41598-020-67539-9`.

[25] Google, *Poly.pizza*. [Online]. Available: `https://poly.pizza/u/Poly%20by%20Google`.

[26] random.org contributors, *Random.org*. [Online]. Available: `https://www.random.org/sequences/`.

[27] Wikipedia contributors, *Romanization of japanese — Wikipedia, the free encyclopedia*, 2025. [Online]. Available: `https://en.wikipedia.org/w/index.php?title=Romanization_of_Japanese&oldid=1284933791`.

[28] Igroup, *Igroup presence questionnaire*. [Online]. Available: `https://www.igroup.org/pq/ipq/download.php#English`.

[29] T. Q. Tran, T. Langlotz, J. Young, T. W. Schubert, and H. Regenbrecht, "Classifying presence scores: Insights and analysis from two decades of the igroup presence questionnaire (ipq)," *ACM Trans. Comput.-Hum. Interact.*, vol. 31, no. 5, Nov. 2024, ISSN: 1073-0516. DOI: `10.1145/3689046`. [Online]. Available: `https://doi.org/10.1145/3689046`.

[30] NASA, *Nasa task load index*. [Online]. Available: `chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://ntrs.nasa.gov/api/citations/20000021488/downloads/20000021488.pdf`.

Appendices

A Generative AI usage

During the preparation of this work, the author used ChatGPT & Grammarly to aid with wording, spelling, and grammar in the final paper. Additionally, PURE suggest was used to aid in finding related literature. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the work.

B VR environments

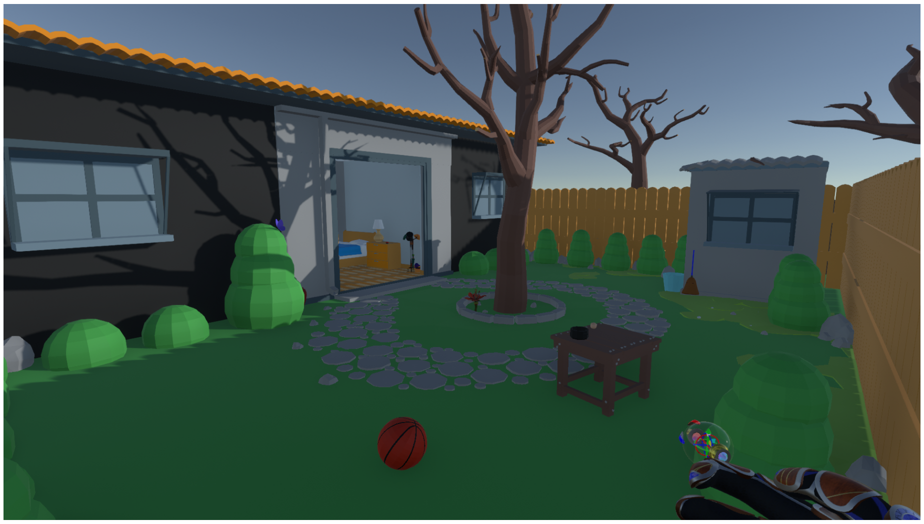

Figure 7: Contextual world: the garden with object words scattered through it.

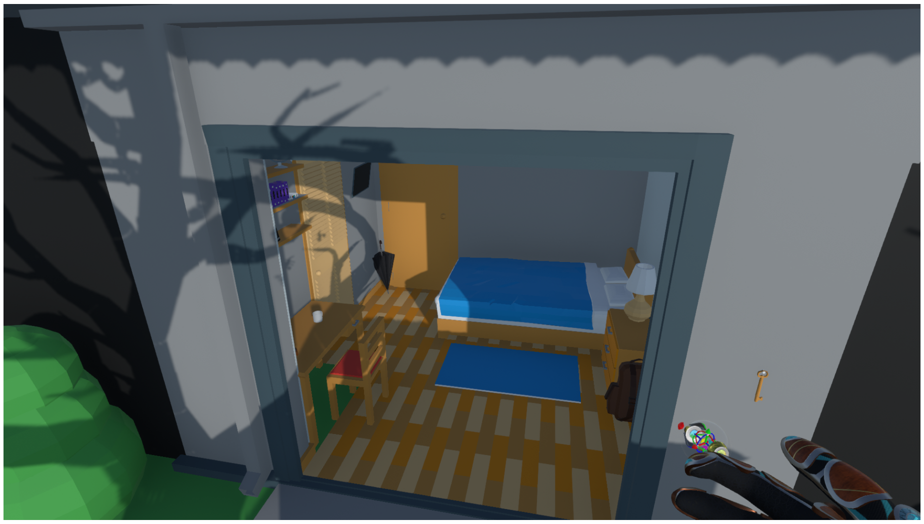

Figure 8: Contextual world: the bedroom adjacent to the garden.

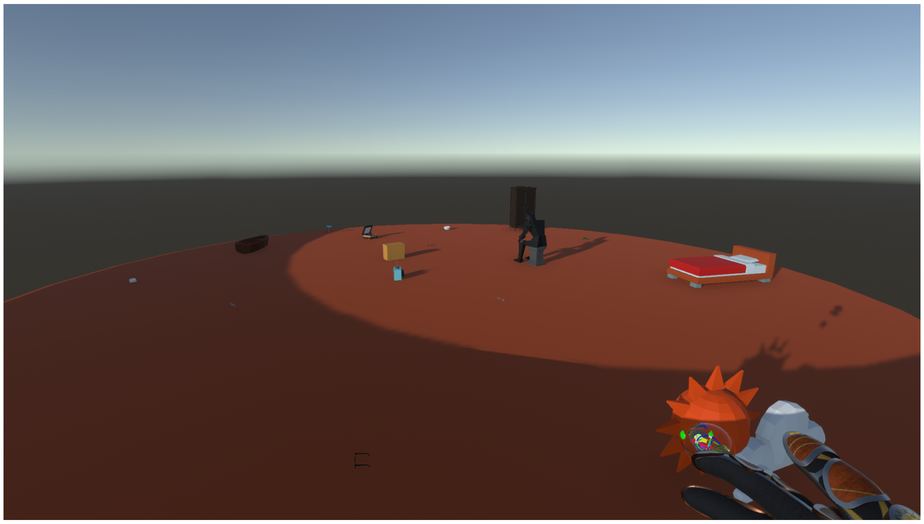

Figure 9: Empty world, with object words scattered on the disk.

C Questionnaire

Igroup Presence questionnaire

We understand the *sense of presence* as the *subjective* sense of being in a virtual environment. Importantly, the sense of presence can be separated from the ability of a technology to *immerse* a user. While this immersion is a variable of the technology and can be described objectively, presence is a variable of a user's experience. Therefore, we obtain measures of the sense of presence from subjective rating scales. This questionnaire contains 14 questions about the perceived presence. For each question, a scale is used from -3 to 3.

* Required

1. What is your participant number? *

2. Which world were you just in? *

○ A

○ B

○ C

3. Select one of the options ranging from -3 to 3, where -3 means extremely aware, and 3 means not aware at all. *

	-3 extremely aware	-2	-1	0 moderately aware	1	2	3 not aware at all
How aware were you of the real world surrounding while navigating in the virtual world? (i.e. sounds, room temperature, other people, etc.)?	○	○	○	○	○	○	○

4. Select one of the options ranging from -3 to 3, where -3 means completely real, and 3 means not real at all. *

	-3 completely real	-2	-1	0	1	2	3 not real at all
How real did the virtual world seem to you?	○	○	○	○	○	○	○

5. Select one of the options ranging from -3 to 3, where -3 means fully disagree, and 3 means fully agree. *

	-3 fully disagree	-2	-1	0	1	2	3 fully agree
I had a sense of acting in the virtual space, rather than operating something from outside.	○	○	○	○	○	○	○

6. Select one of the options ranging from -3 to 3, where -3 means not consistent, and 3 means very consistent. *

	-3 not consistent	-2	-1	0 moderately consistent	1	2	3 very consistent
How much did your experience in the virtual environment seem consistent with your real world experience?	○	○	○	○	○	○	○

7. Select one of the options ranging from -3 to 3, where -3 means about as real as an imagined world, and 3 means indistinguishable from the real world. *

	-3 about as real as an imagined world	-2	-1	0	1	2	3 indistinguishable from the real world
How real did the virtual world seem to you?	○	○	○	○	○	○	○

8. Select one of the options ranging from -3 to 3, where -3 means did not feel present, and 3 means felt present. *

	-3 did not feel present	-2	-1	0	1	2	3 felt present
I did not feel present in the virtual space.	○	○	○	○	○	○	○

9. Select one of the options ranging from -3 to 3, where -3 means fully disagree, and 3 means fully agree. *

	-3 fully disagree	-2	-1	0	1	2	3 fully agree
I was not aware of my real environment.	○	○	○	○	○	○	○

10. Select one of the options ranging from -3 to 3, where -3 means not at all, and 3 means very much. *

	-3 not at all	-2	-1	0	1	2	3 very much
In the computer generated world I had a sense of "being there".	○	○	○	○	○	○	○

11. Select one of the options ranging from -3 to 3, where -3 means fully disagree, and 3 means fully agree. *

	-3 fully disagree	-2	-1	0	1	2	3 fully agree
Somehow I felt that the virtual world surrounded me.	○	○	○	○	○	○	○

12. Select one of the options ranging from -3 to 3, where -3 means fully disagree, and 3 means fully agree. *

	-3 fully disagree	-2	-1	0	1	2	3 fully agree
I felt present in the virtual space.	○	○	○	○	○	○	○

13. Select one of the options ranging from -3 to 3, where -3 means fully disagree, and 3 means fully agree. *

	-3 fully disagree	-2	-1	0	1	2	3 fully agree
I still paid attention to the real environment.	○	○	○	○	○	○	○

14. Select one of the options ranging from -3 to 3, where -3 means fully disagree, and 3 means fully agree. *

	-3 fully disagree	-2	-1	0	1	2	3 fully agree
The virtual world seemed more realistic than the real world.	○	○	○	○	○	○	○

15. Select one of the options ranging from -3 to 3, where -3 means fully disagree, and 3 means fully agree. *

	-3 fully disagree	-2	-1	0	1	2	3 fully agree
I felt like I was just perceiving pictures.	○	○	○	○	○	○	○

16. Select one of the options ranging from -3 to 3, where -3 means fully disagree, and 3 means fully agree. *

	-3 fully disagree	-2	-1	0	1	2	3 fully agree
I was completely captivated in the virtual world.	○	○	○	○	○	○	○

D Script

Table 12: Script for usertests in key phrases.

when	**what to do and say**
they come in	give participant information letter & consent form, ask them to read and sign give short explanation of experiment: 3 conditions, then 2 questionnairs and a test, after week second test to test words ask: age, biological sex, proficiency japanese, VR experience
tutorial in resonite	please stand up during the VR sessions buttons: left joystick to move, right to turn, right index finger activate laser, middle finger to select object system will tell you if it is wrong or right, adjust volume if needed put on the headset comfortably, make sure everything is sharp there is a risk for motion sickness, if you get dizzy take off the headset and we take a break try to pick up the lighthouse and apple in the world and see what happens look at your hands
learning 1	explain condition (gesture or not, empty or contextual world) there are 18 words, you have 15 minutes to find them all say words aloud when learning you can keep learning the words until you think you know them all after 10 min I will give a hint for unfound words tell them if they found all words after 15 min ask to quit
cog load	first you will make a questionnaire about the cognitive load experienced, follow instructions on the tablet
presence	Now a questionnaire about the immersion in the VR world, fill in on computer, I have filled in participant # and world
post test 1	you get 5 min for this, so take your time try to fill in everything you know, even just a part of a word
learning 2 cog load presence post test 2	the same as in condition 1
learning 3 cog load presence post test 3	the same as in condition 1
appointment part 2	don't study the words between sessions thank you for participanting

E Remote delayed posttest Read-me file

Note: the read-me document was written in Dutch, as Dutch was the first language of all participants who did the posttest remotely, instead of English.

Aangezien je de posttest van een week later op afstand uitvoert gelden de volgende instructies hiervoor:

Ga in een rustige ruimte aan een bureau of tafel zitten, het liefst zodat je naar een muur kijkt.
Zorg dat je alleen bent in deze ruimte en je je goed kunt concentreren.
Je hebt 15 minuten voor deze test. Zorg dat je een timer zet zodat je niet over de tijd heen gaat.
Schrijf tijdens de toets zoveel mogelijk op als je je kunt herinneren. Dus ook als dit maar een deel van een woord is.
Zodra je de test begint open je het bijgevoegde bestand op je laptop, hierin kun je de antwoorden invullen.
Buiten het bestand met de Japanse woorden mag je geen andere hulpmiddelen gebruiken.
Na het afronden van de toets kun je het bestand naar mij sturen.

Je mag de toets maken ergens op de dag dat je deze email ontvangt, wanneer jou het het beste uitkomt.
Succes!