Journal Pre-proof

Using Augmented Reality with Speech Input for Non-Native Children's Language Learning

Che Samihah Che Dalim , Mohd Shahrizal Sunar , Arindam Dey , Mark Billinghurst

PII: S1071-5819(18)30316-1
DOI: https://doi.org/10.1016/j.ijhcs.2019.10.002
Reference: YIJHC 2365

To appear in: *International Journal of Human-Computer Studies*

Received date: 10 June 2018
Revised date: 25 September 2019
Accepted date: 8 October 2019

Please cite this article as: Che Samihah Che Dalim , Mohd Shahrizal Sunar , Arindam Dey , Mark Billinghurst , Using Augmented Reality with Speech Input for Non-Native Children's Language Learning, *International Journal of Human-Computer Studies* (2019), doi: https://doi.org/10.1016/j.ijhcs.2019.10.002

Highlights:

- Augmented Reality could be associated with theories of second language acquisition.
- Using speech input for the interaction, effective learning experience is predicted.
- 120 non-native children participated: learning gain and enjoyment measured.
- Results show higher learning gain of abstract words, children highly motivated.
- Speech-enabled Augmented Reality can be used as foreign language learning tool.

Using Augmented Reality with Speech Input for Non-Native Children's Language Learning

***Che Samihah Che Dalim*[a,c,d] *, Mohd Shahrizal Sunar*[a,b] *, Arindam Dey*[d] *, Mark Billinghurst*[d]**

[a]*UTM-IRDA Digital Media and Game Innovation Centre of Excellence, Institute of Human Centred Engineering, Universiti Teknologi Malaysia, 81310 Skudai Johor, Malaysia.*

[b]*Department of Software Engineering, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor Malaysia*

[c]*Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat Johor, Malaysia*

[d] *Emphatic Computing Lab, University of South Australia, Mawson Lakes, SA 5095, Australia*

Declarations of Interest: none

ARTICLE INFO

Article history:
Received
Received in revised form
Accepted

Keywords:
Human-computer interface
Improving classroom teaching
Interactive Learning Environments
Teaching/learning strategies

ABSTRACT

Augmented Reality (AR) offers an enhanced learning environment which could potentially influence children's experience and knowledge gain during the language learning process. Teaching English or other foreign languages to children with different native language can be difficult and requires an effective strategy to avoid boredom and detachment from the learning activities. With the growing numbers of AR education applications and the increasing pervasiveness of speech recognition, we are keen to understand how these technologies benefit non-native young children in learning English. In this paper, we explore children's experience in terms of knowledge gain and enjoyment when learning through a combination of AR and speech recognition technologies. We developed a prototype AR interface called TeachAR, and ran two experiments to investigate how effective the combination of AR and speech recognition was towards the learning of 1) English terms for color and shapes, and 2) English words for spatial relationships. We found encouraging results by creating a novel teaching strategy using these two technologies, not only in terms of increase in knowledge gain and enjoyment when compared with traditional strategy but also enables young children to finish the certain task faster and easier.

1. Introduction

Learning foreign languages can be a challenging but rewarding process for young children. Longitudinal studies by Harvard University confirm that young children who learn foreign languages benefit from an increase in critical thinking skills, creativity and flexibility of the mind (Ford, 2014). Additionally, children who know more than one language are better prepared to take place in a global society and in later years, have wider career opportunities. However, learning a language in countries where that language is not generally spoken can be difficult. An example of such a situation is the teaching and learning of English in Malaysia. Many factors contribute to the difficulty of learning English in Malaysia, including lack of student motivation, which could result from ineffective teaching pedagogy or learning strategies and limited communicative use of English, even at the preschool level (Ansawi, 2017; Azman, H., 2016; Saadiah, 2013; Musa, N., Khoo, Y. L. & Hazita, A, 2012; Hiew, 2012).

Creating an impactful educational setting is crucial for increasing young children's motivation for learning (Weiland and Yoshikawa, 2013). Traditionally, English is taught using songs, textbook exercises, nursery rhymes, and storybook reading. However, today's children grow up in a world full of advanced technologies, which differentiate the way they learn compared to past generations. Technology-supported lessons are able to shape the teaching and learning process to become more innovative (Shapley et al., 2011), which otherwise would be less engaging (Saidin et al., 2015; Teoh, Belinda Soo-Phing, and Tse-Kian, 2007). In order to foster foreign language learning skills among non-native young children, it is important to explore which technology is appropriate to be applied in the classroom to engage young students with the learning materials and motivate them to practice speaking English.

In this paper, we explore the potential of Augmented Reality (AR) and speech recognition as a strategy for teaching basic English to non-native young learners. AR is often used to refer to computer interfaces in which two-dimensional (2D) and three-dimensional (3D) computer graphics are superimposed over real-world objects (Azuma, 1997). Godwin-Jones (2016) explained that AR could be associated with current theories of second language acquisition (SLA), which emphasize localized, contextual learning, and meaningful connections to the real world. Ariza and Hancock (2003) believed that comprehensible input is critical for second language acquisition, and that computer interaction can enhance second language acquisition and fluency. Earlier, handheld AR had been used for learning Japanese nouns by adults (Wagner and

Barakonyi, 2003), to teach Spanish vocabulary to college students (Beder, 2012), and teaching Japanese words for spatial relationships using immersive VR (Rose and Billinghurst, 1995). However, most of the current AR-based language learning applications are primarily for adult users. Limited studies exploring the pedagogical aspect of this technology on young children has been carried out.

Speaking in the targeted second language is a good way to get a grasp of the language. We believed that speech is good at traversing interaction barrier, especially to young children because it lets them cut through reading difficulties and diffidence with a natural or more familiar way of communicating in their age. However, very limited use of the language at home or in school has made non-native young children feel less motivated to use the language (Ansawi, 2017; Azman, H., 2016). Speech recognition can be useful for second language learning where it can teach proper pronunciation and help young language learners develop fluency with their speaking skills. This paper explores how AR and speech recognition technologies could be used together for English language learning with children that are not native English speakers. Our aim is to explore the potential of AR for teaching English names for colors, three-dimensional shapes and spatial relationships to children who do not speak English as their first language. The objectives were; (1) to assess the effectiveness of using AR with that of non-AR tool for teaching English, and (2) to discover if combining speech input with AR cues facilitates the non-English speaking children to learn English names for colors, shapes, and spatial relationships. Through this paper, we intend to answer the following research questions:

1. Is there a significant difference in the learning gain, task completion time, enjoyment and easiness level of children, due to the teaching platforms in which they experience second language learning?
2. Do young children experience higher learning and enjoyment, when they use speech-enabled AR to learn English than when they use the other teaching platforms?

The main contributions of the paper include insights into how speech recognition affects children's interaction with an AR application, and influence their learning experience. The more that is known about the effect of speech input and AR for non-native young children's English language learning, the more effective learning approaches can be developed. This paper is structured into six sections. Following the introduction (Section 1), we provide an overview of related work (Section 2), and then we present the general architecture and development of TeachAR with some guidelines on the interface design (Section 3). Next, we present and discuss the results from a user study (Sections 4 and 5) and finally we give a conclusion and present directions for future work (Section 6).

2. Previous Work

Language is one of the most essential components of a person's life. There are many languages spoken in the world, and second language acquisition (SLA) has become an active area of research. Technology has enabled various strategies for language learning to take place. In this section, we review previous efforts using AR and VR for language learning and other educational applications.

Augmented Reality has been used in many earlier educational applications. For example, traditional printed books which were enriched with AR has been helping children who had difficulties in reading as shown by Hornecker and Dunser (2007), and also able to teach geometry such as shown by Kaufmann et al. (2005). Billinghurst and Dunser provide a high-level overview of how AR could enhance traditional learning models (2012). Zhou et al., (2004) created a new storytelling concept known as the Magic Story Cube (MSCube) using AR technology. The MSCube is a tangible AR interactive interface using a physical cube to provide a storytelling activity for children. More recently, Ibáñez et al., (2014) created an application which enabled students to experience the effects of magnetic fields by overlaying AR information such as electromagnetic forces, while Norraji and Sunar (2016) developed an AR mobile application called "WARna" to enhance children's experience during coloring activities. These applications show that Augmented Reality is able to extend the limitations of traditional learning approaches by enhancing learners' visualization, especially on spatial relationships.

Virtual learning, environments such as AR and Virtual Reality (VR), have previously been used in language learning research (Mroz, 2014). Zengo Sayu (Rose and Billinghurst, 1995) is an example of an immersive interactive VR environment which was developed for non-native adult speakers to learn Japanese prepositions. The results from a user study show that teaching simple Japanese phrases was more effective by using the Zengo Sayu system compared to text-based and real-world approaches. Besides teaching Japanese phrases, researchers such as Wagner and Barakonyi (2003) also developed a handheld AR application for people who wants to learn Japanese Kanji. When the user flips a card which was written in Kanji and shows it to a camera, the virtual objects representing the nouns will be overlaid on the card. However there was no user evaluation conducted with the system, and the software was targeting adult users. Second Life (SL) (2003), a 3-D Multi-User virtual environment, is another virtual environment which was commonly used for foreign language learning. For example, Cheng et al., (2010) used SL to deliver field experiences to pre-service Mandarin teachers. Yoon (2014) found that almost all students showed significant interest and positive attitude when learning the English language in virtual environments, such as Second Life.

Lin and Lan (2015) conducted a survey on the use of VR for teaching languages where they reported that the higher education groups received primarily more attention on the use of the technology between 2004 and 2008. However, between 2009 and 2013, junior and senior high school students dominate the attention, leaving behind the elementary groups. Therefore we are interested to study how children in the elementary group, whose age between 4 to 6 years old, learning a language using AR interface.

As an example of an application for young children, Juan et al., (2010) created a marker-based AR game which can be used by children of 5 to 6 years old to learn Spanish. The results gained from the study indicate a strong potential for using AR as a language learning tool for children. marker-based AR system for English vocabulary learning was developed by Chang et al., (2011) who conducted a user study finding that students liked the multimedia instruction that the AR learning system provided. Similarly, Santos et al., (2016) reported that use of handheld AR could potentially lead to improved retention of words and keep students motivated and satisfied with vocabulary learning. This is consistent with the study by Solak and Cakir (2016) whose post-test and retention test on fifth-grade Turkish students learning new English words at the elementary level shows that participants who did the learning using AR achieved a higher score and performed better in recalling the learned information. There is also AR web-based language learning websites such as LearnAR (2015) which offer quizzes for several choices of language vocabularies, such as English, French, and Spanish. Miyosawa et al., (2012) investigated brain activity while university student's learned languages using AR and the traditional printed methods, finding that using AR was less stressful than with printed material.

There are already some AR applications that aid in real-time language translation and learning. For example, Parhizkar et al., (2013) demonstrated an AR mobile translator which could recognize text in the Malay language and translate it into English in real-time to assist foreigners communicating with Malay people. Fragoso et al., (2011) described a fairly similar application called TranslateAR, a multimodal mobile AR translator developed for the Nokia N900 smartphone. A more recent application of this kind was developed by Meda (2014), who developed a system which could detect and translate English words into Telugu in real-time. However, none of the above systems were designed for young children or evaluated with actual users to identify their acceptance and usefulness.

In recent years, researchers and developers have begun to implement language learning aids in the form of games. Bereira et al., (2012) developed Matching Objects and Words, an AR game for learning words in Portuguese and English. This desktop-based game provided visual and auditory cues to motivate 7 to 9-year-old elementary school children to memorize how to write and pronounce the names of animals. Boonbrahm et al., (2015) developed an AR game for English language learning on a mobile platform, to encourage Thai primary school students to learn written and conversational English. In this game, virtual objects appear in response to written input by the students. The students could learn English conversation by watching and listening to virtual characters speaking to each other when two AR markers were positioned close to each other. However, these systems only provided recorded audio output in order to teach children the pronunciation of words.

Table 1 lists previous works using AR for language learning. Compared to this earlier research our work makes a number of important contributions. Kumar et al., (2012) reported that verbal activities such as recalling and vocalizing words enhance language learning skills. However, in the previous AR systems, none allowed the users to vocalize words for language learning. In our AR system, we use speech recognition, so that students have the experience of pronouncing words and not only listening to them. In addition, our system is the first AR language learning system to target children below 7 years old. Finally, our work is also the first to use AR for English language learning of spatial relationships and 3D shapes.

Teaching abstract concepts such as spatial relationships is an effective application of AR that should be used in education (Liarokapis et al., 2004; Shelton and Stevens, 2004; Bujak et al., 2013). Bujak et al., (2013) explained that AR environments allow students to make physical movements, such as moving their body to change perspective and have control over how they explore the virtual objects, which makes the learning of spatial contents easier. Furthermore, spatial behavior and physical movement are significant for memory process (Pacheco et al., 2015). The spatial association between words and objects could be efficient for vocabulary improvement, so it may be beneficial to teach English words of spatial relationship concepts, not only to improve language learning of non-native children but also to explore how effective AR technology is in creating spatial illusions for young children. To our knowledge, there has been no previous work done in teaching spatial relationships for young children using an AR environment. In our work, when children arrange AR markers in correct English sentence order, a virtual object will be shown in the correct position. Children can also control the view or change the position of 3D objects in real-time by rearranging the AR markers or using their voice. Instead of focusing fully on the screen, children construct the learning experience with the support of the physical world, which results in a more holistic understanding of the experience (Malinverni et al., 2018). These kinds of activities could provide exciting ways to enrich the educational experience and motivate children's learning.

In the next section, we introduce our AR language learning environment and provide a description of the user experience with the technology. We have developed a marker-based AR system that enables children to interact with simple virtual objects using speech and physical motion of real objects. Using this system we have also developed several language learning experiences. An evaluation of the educational experience with the system is described in Section 4.

Table 1. Previous works in AR for language learning

Author (year of publication)	Interaction Technique	Topic	Participant/Evaluation	Objectives	Lessons learned
Hornecker, E., & Dünser, A. (2007)	Marker-based (AR Book with paddle)	"Chick Story" and "Sun Story" storytelling	6 to 7 years old Observing children reading an AR book through video recording	How children aged 6 to 7 years old experience and interact with an AR book	Elements contributing to the user experience include: Story outline, purpose, and role of the interactive sequences, interaction design of interactive sequences, handling of AR elements, integration of paper and screen-based visualizations, handling of the flow between text pages and interactive sequences, use of physical metaphors.
Wagner, D., & Barakonyi, I. (2003, October)	Mobile (PDA), Marker-based	The meaning of Kanji symbols	Nil	Propose educational technology software that uses collaborative AR to teach users the meaning of Kanji symbols	Educational games keep the user focusing for longer periods. Augmented reality permits traditional games to be played to a new extent. The simplicity of the setup increases system performance
Beder, P. (2012)	Mobile-based	Foreign (Spanish) Vocabulary learning	20 students (20-26 years old) Flashcard vs AR group A statistical t-test Questionnaire: Vocabulary Knowledge test is given right after and one week later.	Is there any difference in vocabulary recall rate between AR and flashcard methods for short and long term memory	Display 3D virtual objects along with their spelling and audio pronunciation. Improvement in long term recall rate in the AR group. No difference in short term recall rate between both groups
Juan, C. M., Llop, E., Abad, F., & Lluch, J. (2010, July)	Marker-based	Nil	32 children, real vs AR game Usability (easiness), affective, cognitive, orientation, pedagogy	Evaluate user acceptance of the implementations of an AR-learning system for English vocabulary	No significant difference between both games, but 81% of the children liked the AR game most
Chang et al (2011, July)	Marker-based	English vocabulary learning	111 university students Assessed correlation (regression analysis) between variables in the TAM model Use questionnaires	Understand learner's attitude (satisfaction, behavioral intention, effectiveness) towards AR learning	System quality is a critical factor affecting perceived satisfaction (PS), perceived usefulness (PU) and AR-learning effectiveness. The learner's behavioral intention was affected by PS and PU of the AR learning system. The design of system functions must be straightforward when adopting new technology in learning
Miyosawa, T., Akahane, M., Hara, K., & Shinohara, K. (2012)	Marker-based	Teaching Japanese and Indonesian language (things around the house, foods)	30 university students (verification test), 10 students for brain activity Printed teaching material vs AR Use questionnaires seeing for thoughts and opinion Compare verification test result, and monitoring brain activity during the learning	To assess whether AR teaching tools are useful in learning foreign languages	No significant difference in test results between the two media, however, subject brains was more active while studying the non-AR, suggesting AR as more natural and less stressful.
Parhizkar et al (2013)	Mobile-based	AR word by word Mobile Translator	Nil	How the implementation of a real-time visual translator on AR mobile applications can help people in understanding/translating and give meanings to some writings displayed around them ubiquitously	The objects in the real and virtual world must be properly aligned with respect to each other, or the illusion that the two worlds coexist will be compromised.
Fragoso, V., Gauglitz, S., Zamora, S., Kleban, J., & Turk, M. (2011, January)	Mobile-based	Mobile Translation application	University students	Describing a multimodal AR translation system using a Nokia n900 camera combined with the Tesseract OCR engine and the Google translation API	The OCR engine is the most likely cause of failure. Instead of improving OCR (time-consuming), one can increase the image quality by fusing multiple images over time. Integrating a spell checker is also an option.
Meda, P., Kumar, M., & Parupalli, R. (2014, December)	Mobile-based	Mobile Translation application (English to Telugu)	System testing Evaluate accuracy and usefulness	To study the efficacy(accuracy) and usefulness of the developed prototype	OCR produces good results under normal lighting condition and degrades (the quality) if the lighting condition is poor or uneven. Low-quality image (blur, out

					of focus, improper image rotation) are among the reasons for low detection accuracy.
Barreira et al (2012, June)	Marker-based	3D Objects (animals) vocabulary English words learning (Portuguese and English)	26 children, 7-9 years old (Portuguese elementary school) Traditional method vs AR Pre-post test	To compare the impact of using AR with the traditional method	Provide audio output to teach oral pronunciation The AR game had superior English learning progress Children considered AR easy to use, and it had a positive pedagogical impact on young children's learning process
Mahadzir and Phung (2013)	Marker-based	Digital storytelling tool	Integrate AR pop-up book with Keller's ARCS model of motivation. Elementary school students (year 1). Interviews	Role of AR pop-up books to motivate and help increase English language proficiency	AR helped boost students' performance by providing a more motivating learning environment for students
Boonbrahm, S., Kaewrat, C., & Boonbrahm, P. (2015, August)	Marker-based Mobile AR (Frame marker & user-defined target)	Learn letters, words, and conversation	3 experiments, each having two groups of ten students (writing, reading, conversation) Primary school students (grade 1-5). Higher grades test the conversation experiment	To prove that AR can motivate children in learning English	Expt 1 and 2: AR group spent more time and the percentage of perfection is higher, showing that students enjoy taking more time to study the result. Expt 3: AR group was found to be less afraid of making a conversation in English Overall, students who have no experience in using AR can quickly learn the AR tool
Li, S., Chen, Y., Whittinghill, D. M., & Vorvoreanu, M. (2014)	Mobile-based	Vocabulary: 6 nouns (animals)	Use the ARCS model 5 graduate students who had been learning English for several years	To assess motivation	Learner's motivations were increased by AR in the beginning and decreased along with the disappearance of its novelty. The influence of AR technology on learning motivation varies by different vocabulary categories, such as concrete or abstract words. Learners indicated that their motivation to use AR to learn abstract words was higher than pictographic words.
Solak, E & Cakir, R. (2016)	Marker-based	New vocabulary	Quasi-experimental 61 fifth grade students, experimental vs control group Post-test and retention test	Inform current applications and literature on AR in education and present experimental data about the effectiveness of AR in language classroom at the elementary level	The experimental group achieved a higher score and better recall of the learned information. The use of AR increased learner's performances, made vocabulary learning more effective compared to the traditional method
Santos et al. (2016)	Mobile-based	30 Filipino and 10 German words Learning	31 graduates (26 male, 5 female, aged 23-42, different native languages, between-group study Usability (SUS, HARUS), Learning outcome (Post-test, delayed posttest)	To explore the design and evaluation of AR for educational settings.	AR may lead to better retention of words and improve student attention and satisfaction AR has positive effects on motivational factors of attention and confidence

3. Material and Method

The materials used for this study were developed in two versions: a desktop-based educational AR application namely TeachAR (Fig. 1), and it's equivalent non-AR version (Fig.2) both aims at teaching young children the English words for colors, shapes and spatial relationships. Both versions were compared to explore the effectiveness of AR combined with speech as input for children's English language learning. A desktop platform was chosen to provide a wider view of the AR scene to children. Our system uses the Microsoft Kinect's (2014) hardware and software for speech recognition, the ARToolkit plugin for the Unity game engine for square marker tracking (Kato and Billinghurst, 1999) and application development, a webcam with 1280 x 720 pixel resolution and 78 degree field of view for image capturing, and a desktop monitor for viewing the AR scene.

Table 2 shows the type of modes available in the application. Both the AR and non-AR modes support interaction with and without speech input. In the AR case, with speech disabled (SD), children can use the AR markers to initiate and interact with virtual objects. In the speech-enabled (SE) mode, AR markers are used with speech input to enable interaction with 3D virtual content in the live camera view. In the non-AR case, the 3D objects are shown on the screen and can be interacted with mouse-only input in the SD mode, or using speech-only input in the SE mode.

Fig. 1. AR Interface for color and shapes (top) and spatial relationship (bottom) learning. SD version (left), SE version (right)

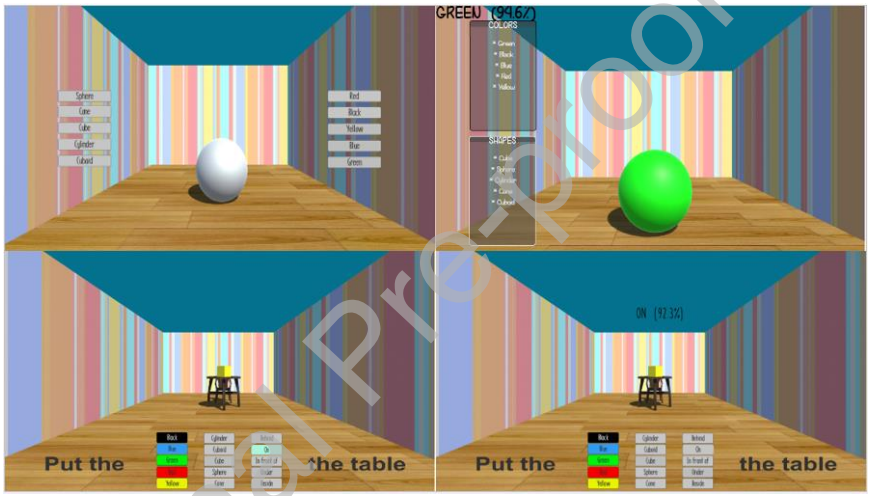


Fig. 2. Non-AR Interface for color and shapes (top) and spatial relationship (bottom) learning. SD version (Left), SE version (Right)

In the speech-disabled (SD), AR version of our application, ten different markers were used to teach five colors and five shapes, with one marker for each color or shape. When the AR marker is in view of the camera, the children will see either a virtual color square or a shape on the marker (see Figure 1). For the speech-enabled mode (SE), only one marker was used and children are required to say out loud a color, followed by shape to enable a virtual object to appear on the marker. To teach prepositions we changed the spatial location of a virtual object in relation to a virtual table shown on the screen. This was done by using an additional Scene marker to initiate the virtual table, and five preposition markers in the SD mode. For the SE mode, the Scene marker is used together with three other markers, each representing colors, shapes, and prepositions, which respond to voice input from children. All markers used in this study are as shown in Fig. 3.

For the non-AR version, similar changes in the virtual objects were achieved using a mouse click in the SD version and using voice commands in the SE version. For example, for the SD version, the user clicks on the green button on the screen to change the color of the virtual cone. For the SE version, we used virtual buttons on the screen and a list of colors, shapes, and prepositions that could be selected by voice to interact with the system.

3.1 Application Development

We developed our application using the Unity game engine (2005) with the ARToolkit for Unity plugin (2016). ARToolkit is a computer vision library used for tracking square image markers to create an AR view. A total of 31 square markers were created for the application. In the AR scene, the rendering is done in Unity. The application sets the background video to the ARToolkit controller, while the rest of the scene is set as an AR foreground layer in the Unity camera. The microphone array of the Kinect sensor is used for capturing the user's speech and passing it to the Microsoft Speech API (SAPI) for recognition based on a list of keywords stored in an XML document. The XML format is easy to author, which makes it easy to modify the grammar.

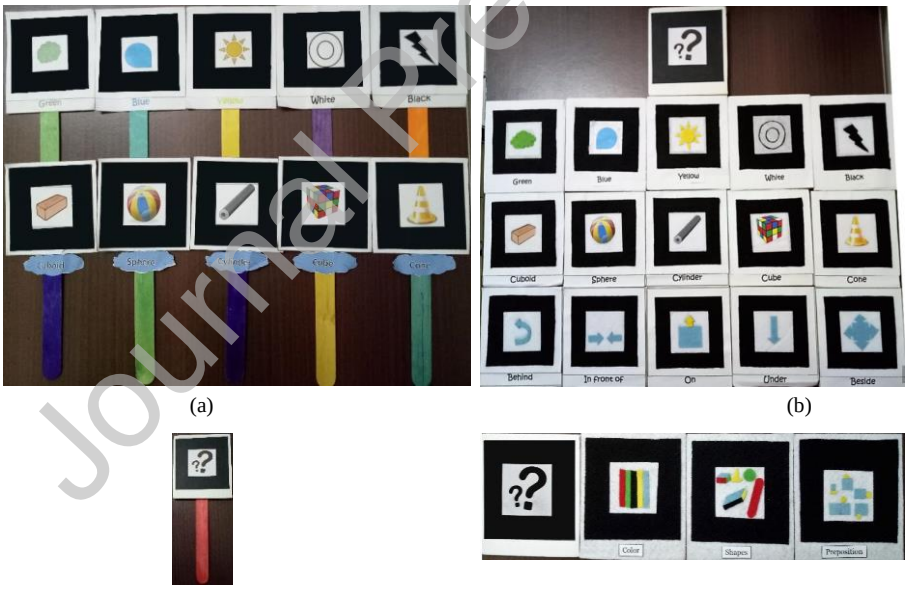
There are also challenges with speech recognition that we needed to design for. For example, speech recognition

accuracy decreases in a noisy environment and when there are many speakers at the same time. Furthermore, young children often have difficulty with word articulation. To avoid misrecognition due to these problems, we set the threshold value for the recognition confidence level to 0.4 and only had a single user at a time. The Kinect sensor was positioned 30 cm from the children.

The content for learning in the application consists of colors, 3-dimensional shapes, and prepositions. Color is taught in most preschool as a fundamental topic and included in preschool activity books. To take advantage of AR visualization, we decided to also include learning 3D shapes as a supplement to the current preschool teaching of 2D shapes. Comparison of 2D and 3D shapes promotes the understanding of transformations which is fundamental to developing visualization abilities and understanding geometry. We also include the learning of prepositions in order to explore the potential of AR for teaching spatial relationships. In the following section, we describe the AR marker and speech input interaction for our application in more detail.

3.1.1 AR Markers

In this application, 31 square AR markers (Fig. 3) were used to initiate the AR scene. These AR markers are simple black and white square printed objects with colored patterns which can be easily recognized and tracked by the camera. When the ARToolkit markers are detected by the camera, the system will draw 3D virtual objects on them by tracking their position and orientation. We used a big-sized marker (8cm on each side), as suggested by (Park and Sang-Jin 2014) because it is good for accurate marker recognition and tracking. ARToolkit must see the whole marker in order to recognize it, therefore, we attached a paddle to the color and shapes markers so that children did not have to touch the marker area while holding it. To avoid fingers covering the tracking area of the preposition markers, we used a white border of 0.5cm to provide an area for children to touch when arranging the markers next to each other. To minimize the detection imprecision due to light, we used a matte felt cloth to create markers that did not reflect light. During the design process, we had to ensure that the felt cloth was properly attached to avoid casting shadows which may affect the tracking accuracy.



(a) (b)

(c) (d)

Fig. 3. AR markers available in the speech-disabled mode of (a) Color and Shapes learning and (b) Spatial relationship learning, the speech-enabled mode of (c) Color and Shapes learning, (d) Spatial relationship learning

3.1.2 Speech Input

In our prototype, speech recognition was used to give commands directly to the AR system. The Color, Shapes and Preposition markers in our system work in combination with speech input. Our system requires children to say a word associated with the right marker, from a list of words. We used the Kinect microphone and accessed its SDK for speech recognition. The microphone array of the Kinect sensor was used for capturing the children's speech and passing it to the Microsoft Speech API for recognition based on a list of keywords stored in an XML document. The ConfidenceLevel property in the RecognitionClass of the speech recognition engine determines how well the speech input has been recognized. We set the threshold value for the recognition confidence level to 0.4 to avoid misrecognition in noisy

environments. The Kinect microphone was positioned 30cm from the children and only used by a single user at a time. The Kinect was used because it enables children to move in a natural setting without having to wear a microphone.

The list of grammars used in this system was:

a. **colorGrammar**: "red", "green", "black", "yellow", "blue"
b. **shapesGrammar**: "cube", "cuboid", "cone", "sphere", and "cylinder"
c. **prepositionsGrammar**: "on", "beside", "in front of", "behind", and "under"

Figure 4 shows the complete speech interaction process for our system. The speech-based AR for spatial relationship module works by using voice commands to position a virtual object relative to a virtual table. Children have to first say a color, followed by shape and finally a preposition, in that order. For example, saying "Put the red cone under the table". A virtual object will be created only when all three words are recognized. Similarly, in the speech-enabled mode of Color and Shapes, children have to first say a color followed by shape in order for a virtual object appears on the screen. The order of the words that need to be spoken should be able to teach children how to build a simple prepositional phrase, for example, "Put the **yellow cube on** the table" in spatial relationship learning, and the noun phrase such as "**Yellow cone**" in the color and shape learning.

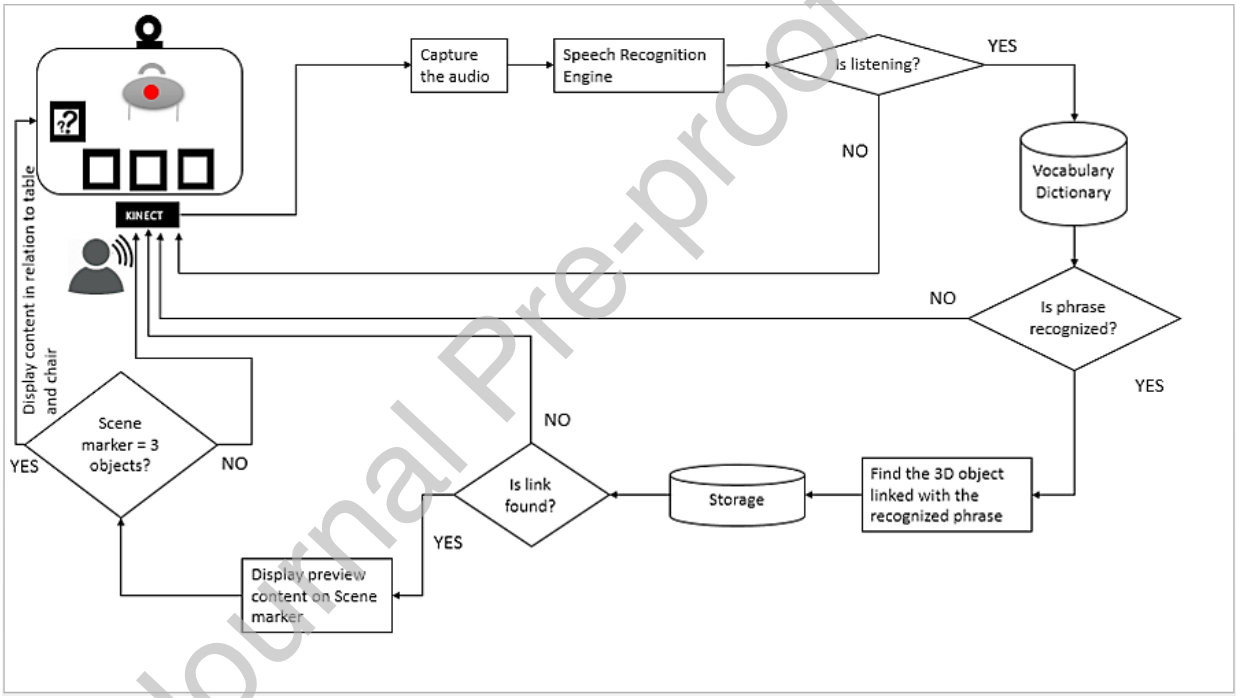


Fig. 4. Application workflow for Speech-based AR for Spatial Relationship (SR) module

3.2 Study Design

We conducted user studies at 6 preschools located in Johor, Malaysia. A total of 120 Malaysian preschool children (62 females, 58 males) aged between 4 to 6 years old, whose native language was Malay, participated in the study (M =5.36 years old, SD =0.658). The study was designed as a between-subjects design with one factor that has four levels:

Teaching Platform: 1) Non-AR Non-Speech, 2) Non-AR Speech, 3) AR Non-Speech and 4) AR Speech

In this study, there were 8 treatments to be tested: a) non-AR non-Speech for color and shapes learning, b) non-AR Speech for color and shapes learning, c) AR non-Speech for color and shapes learning, d) AR Speech for color and shapes learning, e) non-AR non-Speech for spatial relationship learning, f) non-AR Speech for spatial relationship learning, g) AR non-Speech for spatial relationship learning and h) AR Speech for spatial relationship learning, so the participants were divided into 8 groups, with 15 students each. This between-subject study is done to reduce the chances of young children getting bored or distracted after a long series of tests.

With the permission of the teachers of each preschool, the researcher made a few informal 'ice-breaking sessions' with the participants in order to build trust with the students prior to the experiment day. This process was important to ensure that children felt comfortable during the experiment and were willing to cooperate (Read and Fine, 2005). The children

had already attended preschool for at least a year and we chose children who had the capacity to verbalize (vocally or in writing) and to think aloud. This was to ensure that children were able to translate their experiences into verbal statements. The selected children also had a high degree of extroversion, which was assessed by observing how chatty and socially confident they were during the ice-breaking sessions. We also consulted the class teacher during the selection process. This was to ensure positive impact and cooperation on the outcome of the usability test. The chosen children also had no experience using AR or speech recognition applications before the study. Children with learning disabilities or severe behavior difficulties were not included in the study.

The participants were selected after the researcher obtained informed consent from the Director of the Department of Community Development (KEMAS) and teachers from all participating schools. All of the participants had very basic knowledge of English but do not use the language to communicate at home, or at school. The study was conducted at the children's own school to ensure that the children felt comfortable in the test atmosphere, paid more attention and cooperated well with the experimenter, as reported by Mispa et al. (2016) and Razak et al (2010). All of the tools required for the study were set up, including a table, desktop monitor, web camera, Kinect and a play board (see figure 5). The study took place in either a private room or a dedicated space provided by the school so that we could control the noise level which might affect the speech recognition and the lighting which could affect the marker tracking.

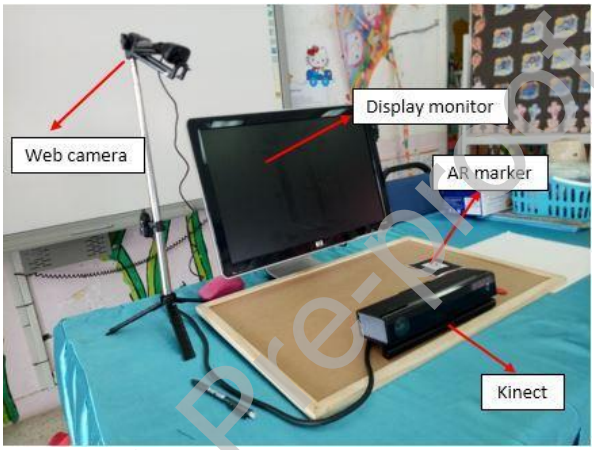


Fig. 5. The workspace setup

The primary purpose of the study was to investigate the effectiveness of the system in terms of learning gain, task completion time, enjoyment and perceived ease of use. Subjective feedback and behavioral cues were set as secondary dependent variables. Table 2 shows the types of modes available in the application and the pre-test scores of the students in different modules is shown in Fig. 6.

Table 2. Application modules

Learning	Technology	Speech
Color & Shape	AR	Enabled
		Disabled
	non-AR	Enabled
		Disabled
Spatial Relationship	AR	Enabled
		Disabled
	non-AR	Enabled
		Disabled

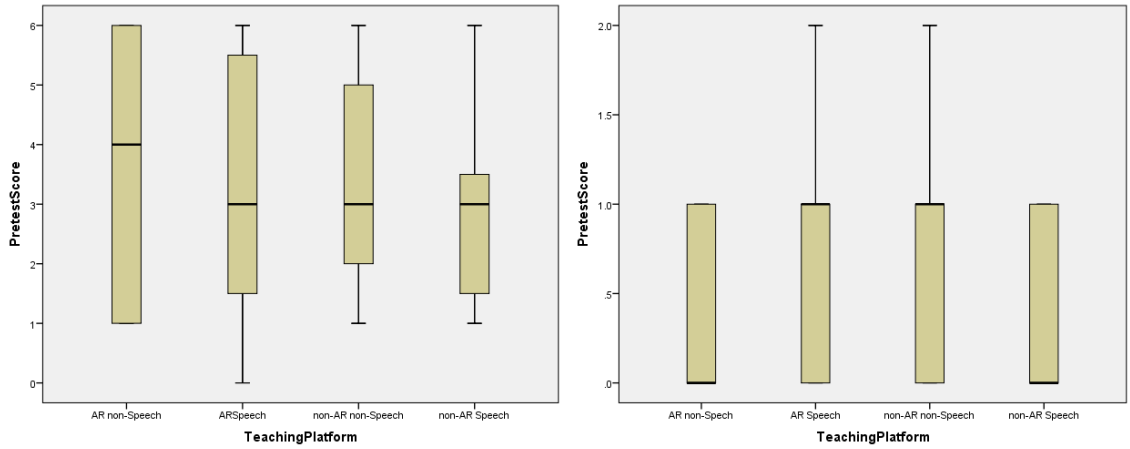


Fig. 6. Pretest score for the groups in Color and Shapes learning (left) and Spatial Relationship learning (right)

3.3 Data Analysis

To investigate how different modes of TeachAR shaped children's experience in language learning, we combined the analysis of data from a pre and post-test, task completion time, questionnaires from Intrinsic Motivation Inventory (IMI) (Plant and Ryan, 1985), and analyze respond on the perceived ease of use. As a pre-test, we first asked the children to fill out a short questionnaire before starting the experimental tasks. From the questionnaire, we collected information about their knowledge of colors, shapes, and spatial relationships, depending on which module they were in. We also had a post-test after each session to find whether the children gained any new knowledge. At the end of the session, we asked the children to fill out a short questionnaire which we adapted from the interest/enjoyment subscale of the Intrinsic Motivation Inventory (IMI) (Plant and Ryan, 1985) to analyze the enjoyment and interest of the students when using our system. In addition to the IMI questionnaire, we added a statement "This tool is easy to use" and base the response to the Smileyometer (Read, MacFarlane, and Casey, 2002) scale.

IMI is a multidimensional measurement device which is used to evaluate an individual's motivation when doing a specific activity. Originally there are 7 statements in this enjoyment/interest scale that have to be scored on a 7-point Likert scale, ranging from 1 (not at all true) to 7 (very true). Since the original statements and the answer categories were developed for adults, we have modified them and verbally translated them to the Malay language to make it more suitable for our participants. To measure enjoyment, we used the Smileyometer (Read, MacFarlane, and Casey, 2002) which is a pictorial representation of different kinds of happy faces to replace the answer category of Likert scale of IMI questionnaire. The Smileyometer has a 5-point scale with 'Strongly disagree' at the leftmost end and 'Strongly agree' at the rightmost end. To help children understand the different faces of Smileyometer, the researcher clarified the faces by mimicking the verbal and gestures of a local animation character that matches each face of the Smileyometer. The modified statement is shown in Table 3 while the answer category used is shown in Fig. 7. An assistant researcher read and explained all the questions and answers in the Malay language to ensure that the students understood them. To avoid the fear of being evaluated, the children were informed that the questions were just intended to help the researcher in her assignment and not to assess their intelligence. After the question was asked, children respond by pointing their finger to which emoticon on the Smileyometer that fits their answer. The assistant researcher will record the scale that represents the emoticon, on her copy of the answer sheet.

Table 3. Modified Enjoyment/Interest Subscale of IMI Statements adapted from PuppyIR (2009) project.

Original	Modified
I enjoyed doing this activity very much	I enjoyed doing this activity very much
This activity was fun to do	This activity was fun to do
I thought this was a boring activity (R)	**I thought this was an exciting activity**
This activity did not hold my attention at all (R)	**This activity held my attention very well**
I would describe this activity as very interesting	I would describe this activity as very interesting
I thought this activity was quite enjoyable	I thought this activity was quite enjoyable
While I was doing this activity, I was thinking about how much I enjoyed it	While I was doing this activity, I was thinking about how much I enjoyed it

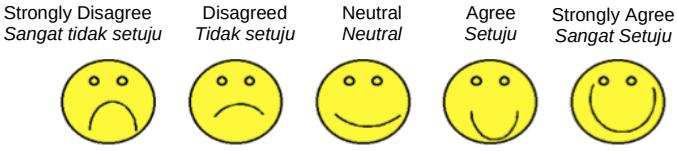


Fig. 7. The Smileyometer Scale used to represent answer categories

3.2.1 Procedure and Task

During the initial phase of the project, we had an interview session with the teachers of the preschools involved. The session aimed at getting their feedback and suggestion on the appropriate content for children's software and how the traditional teaching strategy could be enhanced. The information gathered during the session helped us plan our strategy and determine the content of our prototype. Following that, a low-fidelity prototype was developed and a pilot study was carried out with a small number of children (Dalim et. al, 2016). The current prototype used in this user study is a revised version of the previous prototype. The full user study was carried out as follows:

We conducted two experiments: (1) Color and Shape learning and (2) Spatial relationship (prepositions) learning. We taught the following words in these experiments:

Experiment 1:

Color: Red, Green, Blue, Yellow, and Black.

Shape: Cube, Cuboid, Cone, Sphere, and Cylinder

Experiment 2:

Prepositions: On, Beside, In front of, Behind, and Under

We used two different presentation modes: AR and non-AR. In both of these modes, children could interact with the system using marker and speech (for AR) or mouse click and speech (for non-AR). Participants were first welcomed and positioned facing a desktop screen where the augmented scene was displayed. The selection of the participants for each group was done randomly to obtain a more representative result. In terms of the experimental setup, a webcam with an ultra-wide field-of-view (FOV) lens which provides 78 degrees of visibility was attached on top of the computer screen. The Kinect sensor was placed 30cm from the user. The entire duration of every session for each subject was between 30-35 minutes.

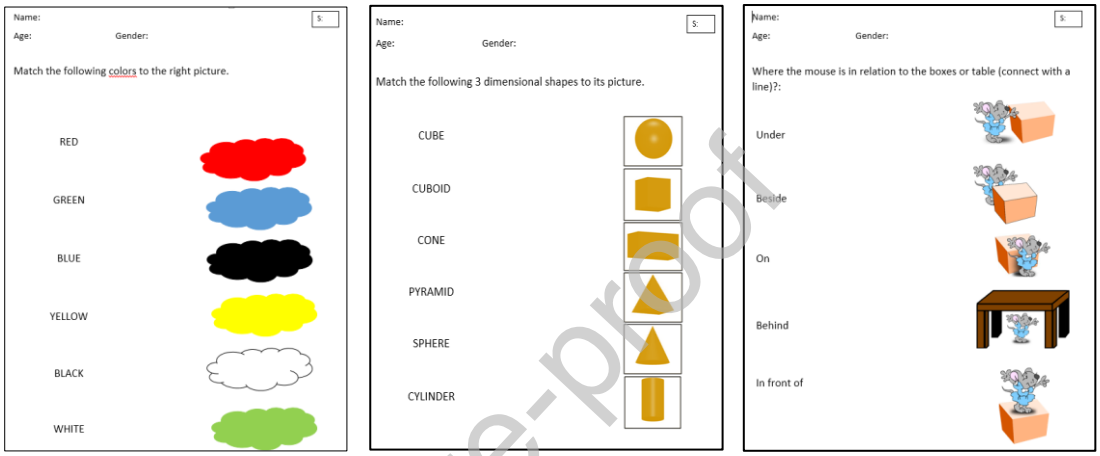
After getting their name and age, we gave a brief orientation to the participants. The pre and post-questionnaires for both experiments were in the form of a matching question (left) to the answer (right). Word to picture matching is a common type of preschool worksheet, therefore, our participants had no problem to understand the instruction. The session was administered verbally to enable the children to understand the questions. In the color and shapes learning, the questions were a list of five colors and five shapes printed on the left side of the question paper. On the right side were five clouds in different colors and five 3D shapes which represent the answers as shown in Fig.8a and 8b respectively. The researcher read color and a shape one at a time and the participant selected their answer from the list on the right side. We noticed that most of the participants knew all of the colors that we had in our system but could not pronounce their English names. However, almost all of the participants did not know the 3D shapes because they had not learned them in school. We demonstrated how the application worked by explaining what the markers were for and how they could be used to change shapes, colors, and position of the virtual objects in both ways, using markers and speech.

In the speech-disabled mode of Experiment 1, participants showed a shape marker to see the virtual shape appear on the marker and by presenting a color marker close to the shape marker, the virtual shape changed its default color to the color of the presented marker. Participants could change markers to learn different colors and shapes in real-time. In the speech-enabled mode, participants were given only one marker and they were required to say out loud a shape in order to see a virtual copy of it on the marker. Once the virtual shape appeared, they could change its color by saying out the color of their preference. The list of words that could be said was displayed on the screen. After each presentation mode participants filled out a post-test questionnaire where their newly acquired knowledge was measured.

In Experiment 2 participants tested the Spatial Relationship module. We asked them to answer a pre-questionnaire to measure their current knowledge about spatial prepositions. Five images of a cat in five different positions around a box were used as the answer list in spatial relationship learning as shown in Fig. 8c. Similar to the color and shape learning, the researcher will read a preposition listed on the left side of the question paper, and the participant will point to one of the cat's images on the right side. In this module, we grouped the markers into three groups; Colors, Shapes, and Prepositions, and specified three exact positions for those markers on the playing board so that when the participants placed the markers accordingly, it formed a short spatial relationship sentence structure. For example "Put the Yellow Sphere On the table". In the speech-disabled mode, participants selected a marker from each group and placed it correctly on the playing board. The virtual object associated with the markers first showed up on the preview marker (marker with a question mark) in real-

time. Once all three markers had been detected, our system rendered the virtual object on the preview marker in the correct position in relation to the table. In the speech-enabled mode, only one marker representing each group were used. Based on the marker group, participants had to say out loud a color, shape and followed by a preposition based on the list shown on the monitor screen in order for the virtual objects to appear on the preview marker. If one object from each of the three groups was detected by our system the virtual object was transferred to the correct table position. At the end of the session, we asked them post-questionnaires.

Throughout the sessions, we took observational notes focusing on identifying any sign of boredom, happiness, and how children interacted with the system. The researcher also took note of the verbal and non-verbal actions, facial expressions and gestures made to help with the interpretation and analysis of the children's feelings during the learning session. The whole session was video recorded using a camera placed in front of the child to be analyzed later to find their level of engagement and enjoyment (see Fig. 10). Fig. 9 shows the overall experiment workflow.



a b c

Fig.8. The question sheet

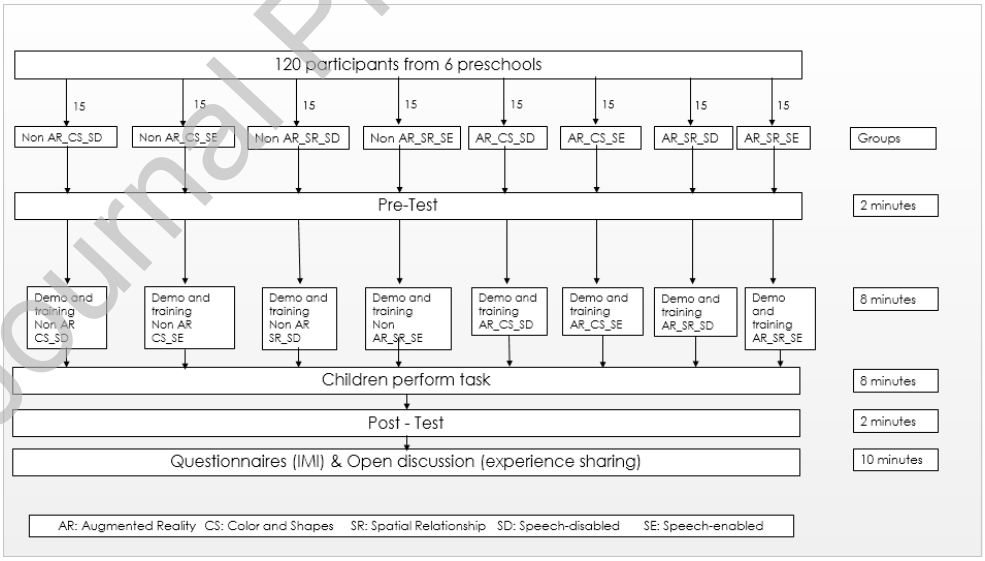


Fig. 9. The experiment workflow

Fig. 10. Children testing the speech-enabled, AR version of (a) Color and Shapes and (b) Spatial Relationship

4. Results

We collected both quantitative and qualitative data during these two experiments. In the first between-subjects experiment we evaluated the effectiveness of the AR-based and Speech-based teaching modules on Color and Shape learning and in the second between-subjects experiment we tested the use of AR-based and Speech-based teaching modules for Spatial Relationship learning. The independent variable was *Teaching Platform* (non-AR non-Speech, non-AR Speech, AR non-Speech, and AR Speech). As quantitative variables, we measured the knowledge gain (difference of scores in pre-test and post-test results) and the task completion time. As qualitative variables, we collected the IMI survey results (Interest and Enjoyment subscales) and the perceived ease of use through a subjective questionnaire. To statistically analyze whether the knowledge gain, task completion time, ease of use and enjoyment score differs significantly by different teaching platforms, we used the Kruskal-Wallis H Test for both quantitative and qualitative data, followed by Dunn-Bonferroni post hoc test on each pair of groups where the Kruskal Wallis test showed a significant difference.

Experiment 1: Color and Shape Learning

In the experiment for color and shapes learning, we did not find any significant difference between the different teaching platforms in the learning of colors as shown in Table 4. However, the Kruskal-Wallis Test showed a significant difference, $\chi^2(3) = 11.34$, $p = .002$ between the mean ranks of at least one pair of groups for the learning of shapes. Dunn's pairwise posthoc tests were carried out for all six pairs of groups. There was a significant difference ($p = .009$, adjusted using the Bonferroni correction) between the groups who used the non-AR_non-Speech teaching platform and those who used the AR_Speech. The Gain Shapes for the AR_Speech group (Mdn=3) is higher compared to the non-AR_non-Speech group (Mdn=2). There was no evidence of a significant difference between the other pairs for Gain Shapes. This finding is encouraging as it shows that the use of AR and speech input strategy was more effective than traditional methods in terms of knowledge gain. See Table 5 and Fig.11 for more details.

Table 4: Kruskal-Wallis Test Report of Knowledge Gain in color and shapes learning

Measures	***Group***	***N***	***Mean Rank***	***Median***	χ^2	***df***	***Asymp. Sig.***
Gain (Color)	***Teaching Platform***				2.17	3	.539
	non-AR_non Speech	15	28.80	1			
	non-AR_Speech	15	31.90	1			
	AR_non-Speech	15	29.80	1			
	AR_Speech	15	34.50	2			
Gain (Shapes)	***Teaching Platform***				11.34	3	.002
	non-AR_non-Speech	15	21.17	2			
	non-AR_Speech	15	33.20	2			
	AR_non-Speech	15	26.63	1			
	AR_Speech	15	41.00	3			

Table 5: Dunn's Pairwise test for Gain (Shapes) in color and shapes learning

Sample1-Sample2	**Sig.**	**Adj.Sig.**
non-AR_non-Speech-AR_non-Speech	.380	1.00
non-AR_non-Speech-non-AR_Speech	.053	.319
non-AR_non-Speech-AR_Speech	.001	.009
AR_non-Speech-non-AR_Speech	.291	1.000
AR_non-Speech-AR_Speech	.021	.126
non-AR_Speech-AR_Speech	.210	1.000

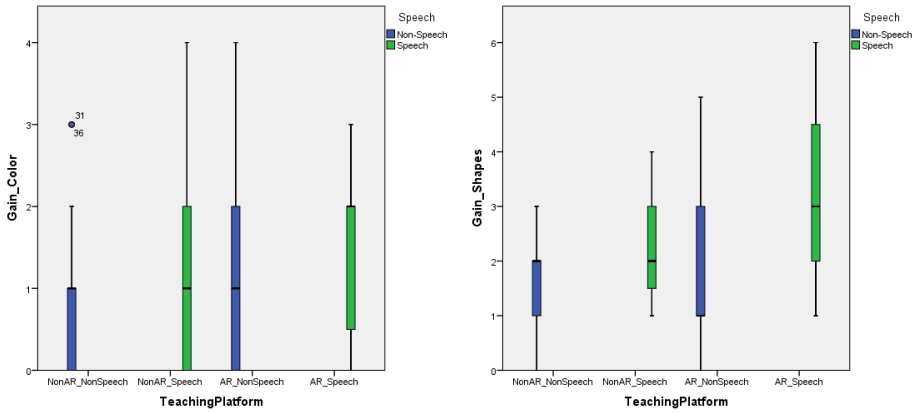


Fig. 11. Knowledge gain for color (left) and shape learning (right)

We found a significant difference in the Task Completion Time between the different teaching platforms, χ^2 (3) =41.35, p<.001 as shown in Table 6. Dunn's pairwise test revealed a significant difference in four out of six pairs of groups. There was a significant difference (p=.037, adjusted using the Bonferroni correction) between the groups who used AR_non-Speech and those who used the AR_Speech. Participants using the AR_Speech teaching style (Mdn=30 seconds) took a longer time to complete the task compared to the AR_non-Speech group (Mdn=14 seconds). Similarly, we found a significant difference (p<.001, adjusted using the Bonferroni correction) between the AR_non-Speech group and non-AR_non-Speech group. Children in the AR_non-Speech group (Mdn=14 seconds) completed the learning task significantly faster than the non-AR_non-Speech group (Mdn=45 seconds).

We also found a significant difference between the groups using AR_non-Speech and non-AR_Speech (p<.001). Participants in the AR_non-Speech group (Mdn=14) completed the task significantly faster than those in the non-AR_Speech group (Mdn= 58 seconds). A significant difference (p=.004) was found between the groups using AR_Speech and non-AR_Speech in which the non-AR_Speech group's median to complete the task is 58 seconds compared to 30 seconds for the AR_Speech group as shown in Table 7. Figure 12 shows that children who use the AR platform completed the learning task significantly faster compared to the non-AR platform. However, in both platform, children took more time to complete the task when speech is enabled.

Table 6: Kruskal-Wallis Test Report of task completion time in color and shapes learning

Measures	***Group***	***N***	***Mean Rank***	***Median***	χ^2	***df***	***Asymp. Sig.***
Task Completion Time	***Teaching Platform***				41.35	3	.000
	non-AR_non-Speech	15	37.83	45			
	non-AR_Speech	15	48.30	58			
	AR_non-Speech	15	9.20	14			
	AR_Speech	15	26.67	30			

Table 7: Dunn's Pairwise test for Task Completion Time in color and shapes learning

Sample1-Sample2	**Sig.**	**Adj.Sig.**
AR_non-Speech-AR_Speech	.006	.037
AR_non-Speech-non-AR_non-Speech	.000	.000
AR_non-Speech-non-AR_Speech	.000	.000
AR_Speech-non-AR_non-Speech	.080	.478
AR_Speech-non-AR_Speech	.001	.004
non-AR_non-Speech-non-AR_Speech	.100	.602

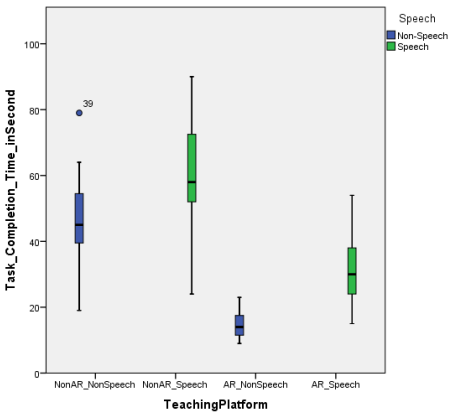


Fig. 12. Task completion time in color and shapes learning

To assess enjoyment, we use the interest/enjoyment subscale of the IMI questionnaire. We found a significant difference in the IMI(Enjoyment) score between the different teaching platforms, χ^2 (3) =37.00, p<.001 as shown in Table 8. Dunn's pairwise test (see Table 9) revealed a significant difference in four out of six pairs of groups. There was a significant difference (p<.001, adjusted using the Bonferroni correction) between the group who used non-AR_non-Speech and those who used the AR_Speech. The median IMI(Enjoyment) score for the groups using AR_Speech (Mdn=4.71) is higher compared to the non-AR_non-Speech group (Mdn=3.71). Similarly, there was a significant difference (p<.001, adjusted using the Bonferroni correction) between the groups who used Non-AR_Non-Speech and those who used AR_Non-Speech teaching platform. The AR_non-Speech group rated higher enjoyment score (Mdn=4.71) than the non-AR_non-Speech group (Mdn=3.71). We also found a significant difference between the non-AR_Speech group and AR_Speech group (p=.007) where the median score for the AR_Speech group (Mdn=4.71) was significantly higher compared to 3.86 in the

non-AR_Speech group. Lastly, a significant difference (p<.001) was found between the groups using non-AR_Speech (Mdn=3.86) and AR_non-Speech (Mdn=4.71). Children in the AR_non-Speech group rated higher enjoyment compared to the children in the non-AR_Speech group. Figure 13 shows the details.

Table 8: Kruskal-Wallis Test Report of enjoyment in color and shapes learning

Measures	*Group*	*N*	*Mean Rank*	*Median*	χ^2	*df*	*Asymp. Sig.*
Enjoyment	***Teaching Platform***				37.00	3	.000
	non-AR_non-Speech	15	14.87	3.71			
	non-AR_Speech	15	19.53	3.86			
	AR_non-Speech	15	47.47	4.71			
	AR_Speech	15	40.13	4.71			

Table 9: Dunn's Pairwise test for enjoyment score in color and shapes learning

Sample1-Sample2	Sig.	Adj.Sig.
non-AR_non-Speech-non-AR_Speech	.462	1.000
non-AR_non-Speech-AR_Speech	.000	.000
non-AR_non-Speech-AR_non-Speech	.000	.000
non-AR_Speech-AR_Speech	.001	.007
non-AR_Speech-AR_non-Speech	.000	.000
AR_Speech-AR_non-Speech	.248	1.000

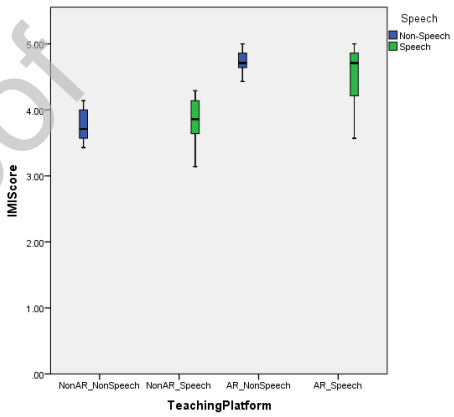


Fig. 13. Enjoyment in color and shape learning

We also asked participants about the perceived ease of use of these teaching platforms. We found a significant difference in the Ease of Use between the different teaching platforms, χ^2 (3) =21.58, p<.001 as shown in Table 10. Dunn's pairwise test revealed a significant difference in two out of six pairs of groups. The non-AR_non-Speech group's rating was significantly different than the AR_non-Speech group (p=.001, adjusted using the Bonferroni correction). Children rated AR_non-Speech easier to use (Mdn=5) compared to the non-AR_non-Speech (Mdn=4). Similarly, we also found a significant difference between the groups using AR_Speech and AR_non-Speech (p=.001, adjusted using the Bonferroni correction) where the median score for AR_non-Speech group (Mdn=5) is higher than the AR_Speech group (Mdn=3). This is understandable as speech input required some training, which made it harder for children to use. Table 11 and Figure 14 shows the details.

Table 10: Kruskal-Wallis Test Report of ease of use in color and shapes learning

Measures	*Group*	*N*	*Mean Rank*	*Median*	χ^2	*df*	*Asymp. Sig.*
Ease of Use	***Teaching Platform***				21.58	3	.000
	non-AR_non-Speech	15	20.60	4			
	non-AR_Speech	15	34.93	4			
	AR_non-Speech	15	43.77	5			
	AR_Speech	15	21.03	3			

Table 11: Dunn's Pairwise test for ease of use in color and shapes learning

Sample1-Sample2	Sig.	Adj.Sig.
non-AR_non-Speech-AR_Speech	.942	1.000
non-AR_non-Speech-non-AR_Speech	.018	.107
non-AR_non-Speech-AR_non-Speech	.000	.001
AR_Speech-non-AR_Speech	.022	.129
AR_Speech-AR_non-Speech	.000	.001
non-AR_Speech-AR_non-Speech	.144	.863

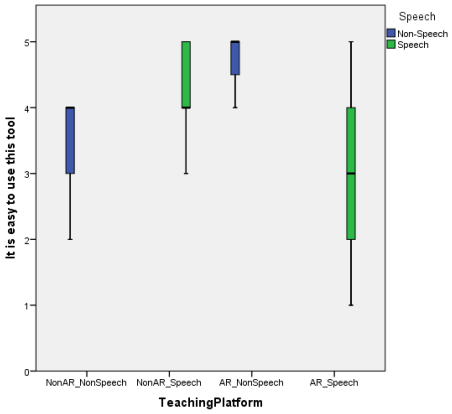


Fig. 14. Ease of use in color and shape learning

Experiment 2: Spatial Relationship (Preposition) Learning

We noticed a significant difference in the knowledge gain of the spatial relationship between the different teaching platforms, χ^2 (3) =17.36, p=.001 as shown in Table 12. Dunn's pairwise test (Table 13) revealed very strong evidence of a difference in two out of six pairs of groups. There was a significant difference (p<.001, adjusted using the Bonferroni correction) between the groups who used Non-AR_non-Speech and those who used the AR_Speech. Children in the AR_Speech group (Mdn=2) produced a significantly higher knowledge gain of the spatial relationship than children who use the non-AR_non-Speech teaching platform (Mdn=1). Similarly, children in the non-AR_Speech group (Mdn=2) produced a higher knowledge gain than the children in the non-AR_non-Speech group (Mdn=1), (p=.044, adjusted using the Bonferroni correction). Figure 15 shows the detail.

Table 12: Kruskal-Wallis Test Report of learning gain in Spatial Relationship learning

Measures	*Group*	*N*	*Mean Rank*	*Median*	χ^2	*df*	*Asymp. Sig.*
Gain	***Teaching Platform***						.001
	non-AR non-Speech	15	16.23	1	17.36	3	
	non-AR_Speech	15	32.70	2			
	AR_non-Speech	15	31.70	2			
	AR_Speech	15	41.37	2			

Table 13: Dunn's Pairwise test for learning gain in Spatial Relationship learning

Sample1-Sample2	Sig.	Adj.Sig.
non-AR_non-Speech-AR_non-Speech	.012	.071
non-AR_non-Speech-non-AR_Speech	.007	.044
non-AR_non-Speech-AR_Speech	.000	.000
AR_non-Speech-non-AR_Speech	.871	1.000
AR_non-Speech-AR_Speech	.116	.695
non-AR_Speech-AR_Speech	.159	.951

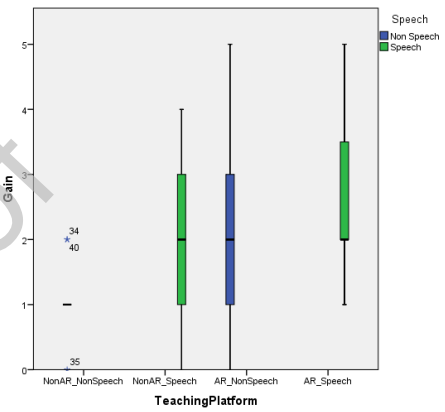


Fig. 15. Knowledge gain for spatial relationship learning

Table 14 shows that the Kruskal-Wallis Test yielded a significant difference in the Task Completion Time between the different teaching platforms, χ^2 (3) =44.75, p<.001. Dunn's pairwise test revealed a significant difference in four out of six pairs of groups as indicated in Table 15. There was a significant difference (p<.001, adjusted using the Bonferroni correction) between the groups who used non-AR_non-Speech (Mdn=10) and those who used the AR_Speech (Mdn=25). Children in the AR_Speech group took a longer time to complete the task compared to those in the non-AR_non-Speech group. Similarly, there is a significant difference (p<.001, adjusted using the Bonferroni correction) between the groups who used non-AR_non-Speech and those who used non-AR_Speech. The non-AR_Speech group took longer time (Mdn=36) to complete the task than the Non-AR_Non-Speech group (Mdn=10). We also found a significant difference between the groups using AR_non-Speech and AR_Speech (p=.001) where the median for the AR_Speech group is 25 seconds compared to 10 seconds in AR non-Speech group. Finally, a significant difference (p<.001) was found between the groups using AR_non-Speech and non-AR_Speech in which the non-AR_Speech group's median is 36 seconds compared to 10 seconds for the AR_non-Speech group. Figure 16 indicates that in both AR and non-AR teaching platform, task completion time is longer when speech is enabled.

Table 14: Kruskal-Wallis Test Report of task completion time in Spatial Relationship learning

Measures	*Group*	*N*	*Mean Rank*	*Median*	χ^2	*df*	*Asymp. Sig.*
Task Completion Time	***Teaching Platform***				44.75	3	.000
	non-AR_non-Speech	15	15.53	10			
	non-AR_Speech	15	49.57	36			
	AR_non-Speech	15	15.97	10			
	AR_Speech	15	40.93	25			

Table 15: Dunn's Pairwise test for task completion time in Spatial Relationship learning

Sample1-Sample2	Sig.	Adj.Sig.
non-AR_non-Speech-AR_non-Speech	.946	1.000
non-AR_non-Speech-AR_Speech	.000	.000
non-AR_non-Speech-non-AR_Speech	.000	.000
AR_non-Speech-AR_Speech	.000	.001
AR_non-Speech-non-AR_Speech	.000	.000
AR_Speech-non-AR_Speech	.175	1.000

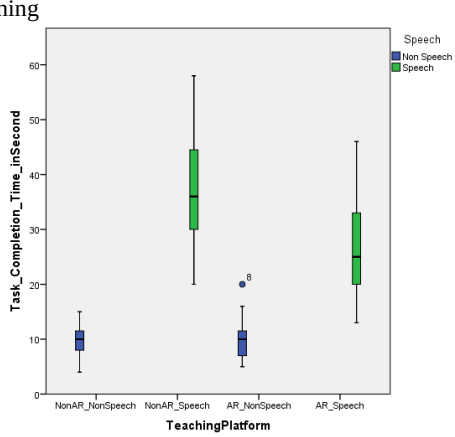


Fig. 16. Task completion time in spatial relationship learning

Our results in Table 16 show a statistically significant difference in the IMI(Enjoyment) score between the different teaching platforms, χ^2 (3) =44.14, p<.001. Dunn's pairwise test revealed a significant difference in four out of six pairs of groups as shown in Table 17. There was a significant difference (p<.001, adjusted using the Bonferroni correction) between the groups who used non-AR_Speech and those who used the AR_Speech. As detailed in Figure 17, the median IMI(Enjoyment) score is higher for the groups using AR_Speech (Mdn=4.71) compared to 3.71 in the group using non-AR_Speech. Similarly, there is a significant difference (p<.001, adjusted using the Bonferroni correction) between the groups who used non-AR_Speech and those who used AR_non-Speech. The AR_non-Speech group rated higher enjoyment score (Mdn=4.71) than the non-AR_Speech group (Mdn=3.71). We also found a significant difference between the groups using non-AR_non-Speech and AR_Speech (p=.009) where the median for the AR_Speech group is higher (Mdn=4.71) compared to 4. 14 in the non-AR_non-Speech group. Lastly, a significant difference (p=.001) was found between the groups using non-AR_non-Speech and AR_non-Speech in which the AR_non-Speech group's median is higher (Mdn=4.71) compared to 4. 14 for the non-AR_non-Speech group.

Table 16: Kruskal-Wallis Test Report of enjoyment in Spatial Relationship learning

Measures	***Group***	***N***	***Mean Rank***	***Median***	χ^2	***df***	***Asymp. Sig.***
Enjoyment	***Teaching Platform***				44.14	3	.000
	non-AR_non-Speech	15	22.40	4.14			
	non-AR_Speech	15	10.30	3.71			
	AR_non-Speech	15	46.87	4.71			
	AR_Speech	15	42.43	4.71			

Table 17: Dunn's Pairwise test for enjoyment in Spatial Relationship learning

Sample1-Sample2	**Sig.**	**Adj.Sig.**
non-AR_Speech-non-AR_non-Speech	.056	.335
non-AR_Speech-AR_Speech	.000	.000
non-AR_Speech-AR_non-Speech	.000	.000
non-AR_non-Speech-AR_Speech	.002	.009
non-AR_non-Speech-AR_non-Speech	.000	.001
AR_Speech-AR_non-Speech	.484	1.000

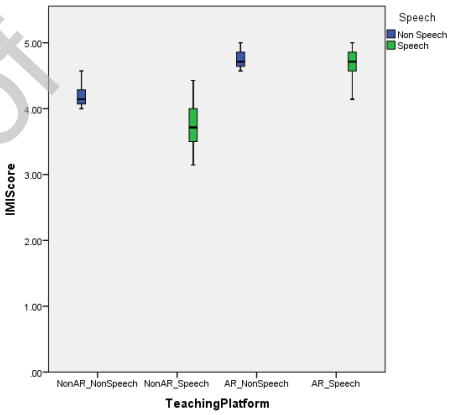


Fig.17. Results for enjoyment in spatial relationship learning

A Kruskal-Wallis Test (see Table 18) found a significant difference, χ^2(3) =9.59, p=.002 between the mean ranks of at least one pair of groups for Ease of Use. Dunn's pairwise tests were carried out for the six pairs of groups. As shown in Table 19 and Figure 18, there was a significant difference (p=.014, adjusted using the Bonferroni correction) between the groups who used the non-AR_Speech teaching style and those who used the AR_non-Speech. Children in the AR_non-Speech group rated the teaching style easier (Mdn=5) compared to the non-AR_Speech group (Mdn=3). There was no evidence of a difference between the other pairs for Ease of Use.

Table 18: Kruskal-Wallis Test Report of ease of use in Spatial Relationship learning

Measures	***Group***	***N***	***Mean Rank***	***Median***	χ^2	***df***	***Asymp. Sig.***
Ease of Use	***Teaching Platform***				9.59	3	.022
	non-AR_non Speech	15	32.70	4			
	non-AR_Speech	15	20.50	3			
	AR_non-Speech	15	39.13	5			
	AR_Speech	15	29.67	4			

Table 19: Dunn's Pairwise test for ease of use in Spatial Relationship learning

Sample1-Sample2	**Sig.**	**Adj.Sig.**
non-AR_Speech-AR_Speech	.135	.808
non-AR_Speech-non-AR_non-Speech	.046	.279
non-AR_Speech-AR_non-Speech	.002	.014
AR_Speech-non-AR_non-Speech	.621	1.000
AR_Speech-AR_non-Speech	.122	.734
non-AR_non-Speech-AR_non-Speech	.294	1.000

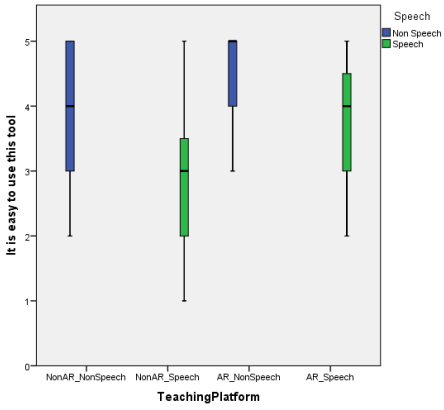


Fig. 18. Ease of use of spatial relationship learning

User Observation Results

In addition to the quantitative method, we also analyze the children's experience and feeling during the language learning activity based on observational notes and analysis of interview from the videos that we recorded throughout the session. Figure 19 shows the typical positive and negative gestures that were recorded during the session and the typical positive and negative verbal feedback is shown in Table 20. Figure 20 shows some difficulty that participants had with using a computer mouse. The researcher also had an interesting conversation (translated from the Malay language to English) with some of the young children who did the AR module, and the child responses are recorded in Table 21.

Fig. 19. Positive gestures of the children (top) vs negative gestures (bottom) in the AR system

Fig. 20. Some participant used both hands to move the mouse (left) and asked for the researcher's help (right)

Table 20. Positive and negative verbal feedback given from the speech-enabled AR platform (translated from Malay to English)

Positive verbal feedback	Negative verbal feedback
Enjoyable "I really enjoyed this game! Can I play again?" "I am so happy teacher."	**Frustrating** "Why the shape not showing?"
Engaging "I don't want to stop playing now" "Can I play again after my friend's turn?"	**Angry** "Why doesn't it work??!"
Exciting "Look! It can hear me!" "Wow. How can the sphere change color?" "I want to ask my mom to buy this game!"	

"Wow. I am in the game (screen)!"
Enthusiasm "Can you teach me how to pronounce it correctly?" "Why I need to say color before shape?"

Table 21. Conversation between researcher and children in the AR mode

Question (Researcher)	Response (Child)
"How did you figure out how to differentiate this shape with a rectangle? (Referring to cuboid)	Child 1: "I don't know. Because the rectangle only has lines. This one is solid. It is not empty."
"Why you did not use voice when playing games before?"	Child 2: "Because I don't have to. This game is different. I have to say the word to move the thing"
"Can you share with me what you feel after playing this?"	Child 3: "I feel like talking to a robot. It listens to my words."
"What do you like most about this game?"	Child 4: I could see it (the virtual object) moves when I move this card (the marker).
"What makes you so excited to play this game?"	Child 5: "I can see myself playing the game on the screen"
"Do you feel exhausted using your speech to move the object?"	Child 6: "No"
"Do you want to take a break?" (the question was asked when the child stopped his action for a while)	Child 7: "Wait, I am trying to figure out how to say the word correctly"

5. Discussion

According to Dual-Coding Theory (DCT) (Clark and Paivio, 1991), a cognition theory, a learner's memory consists of two separate but interrelated verbal and visual codes for processing information. Interestingly, there exists an interconnection between the two separate systems which facilitates dual coding of information if not activated independently. Psychologists have shown how our mind effectively responds to words that make a picture, and some researchers have shown the group means of those who read illustrated text outperformed those who read the text alone (Mayer & Sims, 1994). In this study, we examined the interconnection between the combination of AR which responsible for visual capability and speech recognition which provides the verbal capability. Based on DCT, the visualized spoken words may not only enhance learner's memory of English vocabulary but also stimulates other information processing such as recognition and association of features with previous knowledge. The detail examination of the four main sources (pre-post test result, task completion time, IMI score, perceived ease of use) mixed with the secondary variable (subjective feedback and behavioral cues) provided an insight of how the combination, actualize through different teaching platforms (AR and Non-AR with speech-enabled or disabled), scaffolds children's experiences and learning of second language.

1. Strong attention to details promotes the meaning-making process

The results of this study indicate that the combination of AR and speech recognition technology has a significant effect mainly on the knowledge gain of 3-dimensional shapes and spatial relationships. In both experiments, our participants who used the AR_Speech platform had a significantly higher knowledge gain than those who used the non-AR_non-Speech platform. Unlike the non-AR teaching platform, which provides limited ability to manipulate the visual feedback, AR visualization enabled our participants to examine the 3D shapes and its spatial relationship more interactively. The ability for children to hold the AR marker and see different angles of the virtual shapes on the screen while remaining in their real environment made the learning more interesting. While examining the different angles of the shapes, children started to relate the newly learned shapes with things that exist around them such as comparing a cone to an ice-cream cone, and cuboid to a tissue box. The affordance of the AR teaching platform, which facilitates the interaction between perception, action, the body and environment is aligned with the embodied cognition theory, which improves the cognitive function. Furthermore, this visuomotor involvement in the processing of information indirectly enriched their knowledge. The ability to provide a richer description of what they had learned about individual shapes shows that AR provides a better learning experience and make the newly acquired knowledge last longer. This is consistent with the findings of Chen and Su (2013) where they reported that students in the AR group had a better learning experience as they could give a richer explanation of the animal and sound feedback available in the AR learning version.

We also believed that the AR view that looks like a mirror to our participants where they can see themselves 'in' the screen together with the virtual objects makes them feel more motivated to learn. Mirror has been commonly used with infant and toddlers in a daycare and preschool setting as a developmental tool that is linked to promoting self-awareness

and pre-reading skills. Through AR 'mirror', our young participants were fascinated with their own reflection. For example, one participant said *"I can see myself playing the game on the screen"* in reply to the researcher's question about what makes him so excited to play the game. The feeling of being part of the scene helps them reduce anxiety and began exploring their learning environment carefully. It was also observed that on seeing their own reflection on the screen, our participant began to make more interaction with the virtual objects which eventually enriched the learning experience. Furthermore, the AR visualization also helped to simplify their thoughts of abstract words such as the preposition words that they learned in this study. As for the speech input part, those with speech-enabled teaching style learned more than those with speech disabled. This could be due to the active verbalization of the words and real-time visual feedback of the said object. This result supports a study by Kumar et al., (2012) which highlights the benefits of active verbalization for language learning. It is important to note that most of the words used in this learning were totally new to our participants. Saying these words during the test, along with the immediate 3D visualization of the spoken words, enhanced their cognition during the learning process, agreeing with the dual-coding theory. Furthermore, by seeing their expression of each word on the screen, children learn to imitate and improvise the way they pronounce the words.

Overall, through our research, we found several ways the combination of AR and speech recognition can improve cognitive function. First, we believe that this strategy potentially creates active learning, whereby learners are active in the learning process by constructing their own understanding. This teaching strategy fosters young children's greater involvement in the meaning-making process by forming synergy between physical, virtual and natural interaction. The manner of associating spoken words (natural) with visual information (virtual) strengthen the meaning-making process, while involving physical object from the surrounding promotes contextual learning. The language learning process that the children undergo using our teaching strategy follows what Stephen Krashen (1982) stated that the process of acquiring a language happens in a natural way when children are immersed into a language, reacting simultaneously and interacting subconsciously.

2. Engagement increases focus and reduce task completion time

The time to complete tasks may not be a useful measure for younger children as they are easily distracted, however, it is a concern at any age that children should not take an excessively long time to complete simple tasks (Hourcade, 2015). It was therefore important for us to ensure that young children did not have to spend too much time completing tasks using our learning tool. In both experiments, we found that in most cases, the participants in the AR group finished the task significantly faster than the participants in the non-AR group. It was observed that participants were more motivated to say words in the AR platform as they see their real environment appear on the screen supplemented with the virtual objects that they are speaking about. This has led the children to feel engaged with the activity and using their voice to interact with the system further increases the feeling of naturalness, although the speech recognition takes some time to recognize the children's voice. Without the augmented visual feedback, the engagement decreases.

In both AR and non-AR teaching platform, the children's hand-eye coordination skill is challenged hence affected the time taken to complete the task. In the non-AR platform, children had difficulty to coordinate their hand movement while dragging the mouse to meet the point on the screen. On the other hand, in the AR platform children need to select and place different markers in a specific space on the playboard within the field of view of the camera to ensure the markers were properly displayed on the screen. However, we noticed that children are more comfortable when playing around with the AR markers, most probably because it looks like the flashcards that they used in class. A student said, *"I could see it (the virtual object) moves when I move this card (the marker)".*

Speech input also had a significant effect on the task completion time where the participants complete the task significantly faster when the speech is disabled. This makes sense as saying the words required the device to recognize the participants' voice input, which was time-consuming. In the spatial relationship module, the most frequent word that was confused by the speech recognizer was 'On' and 'Under'. Every time their speech was not recognized by the system, most of the participants changed their pronunciation style which caused a delay in completing the task.

3. Sense of control and relatable immersive experience boost enjoyment

The connection between enjoyment and learning is a well-established premise. In both modules, significant differences were found in the IMI scores between the different teaching platforms. Overall, participants who used the AR version rated a higher score than the participants who used the non-AR version regardless of whether the speech is enabled or disabled. The user observation through video recording also confirmed that participants were very engaged and awestruck while using the AR interface. This could be due to the novelty of the technology, and also the play area setup. For example, in the AR platform of spatial relationship module, participants were provided with a physical play board and AR markers on the table. They were required to arrange the markers in a specific order to place a virtual object in the right position. Our participants were motivated and felt a sense of control when they are able to see how their marker arrangement on the physical playing board affects the position of the virtual objects on the screen. Similarly, in the color and shapes module, children seemed so excited to touch and pick up different AR markers, combining them, and wonder what would appear on the screen. Many participants thought that it was a magic trick as they saw the added graphics rendered on the marker and

some children tried to use their hand to touch the virtual objects. Most of the students turned the shape marker around to examine the different angles of the 3D shapes and started comparing it with real objects, saying things such as, *"this cone shape looks like an ice-cream cone!", "the cube looks like a tissue box".* Some of the participants even looked at the virtual objects as if they did exist around them. This shows the effectiveness of the technology in terms of creating a realistic spatial illusion. Almost all participants changed their facial expression from neutral at the beginning to surprised and happy when the AR system worked. In both modules, the three-dimensional effects provided through the AR version captured the attention of the young learners more easily. We also observed that children love the color markers, and can easily identify different markers by the colorful patterns on them. This result gives an insight that there is a complementary advantage of enjoyment for children who engaged in both physical and virtual interactions.

With regards to speech input, in the non-AR version of the Color and Shapes module, we noticed that the group with speech-enabled teaching platform rated higher enjoyment than the speech-disabled group. This shows that young children were intrinsically motivated to learn to speak. All of them were mostly smiling and wanted to play more while using the AR system. We discovered that participants who used the speech-enabled version of the AR systems were more motivated and excited to speak out English words, as they wanted to see the virtual objects rendered on the marker. They also interact and communicate more with the researcher by asking questions and advice on their pronunciations. In an experience sharing an activity with the whole class, participants who did the speech-enabled version of both systems were observed to be more encouraged to share the new words and the pronunciation that they learned during the experiment with their classmates. This indicates that there could be a significant benefit of implementing speech as input in AR applications for language learning. Children put more effort to pronounce English words. Some participants tried to pronounce the words gently hoping that the speech recognizer recognized the word. Some pronounced it fast to get it recognized quickly. There were also children who thought that they should change positions, from leaning back in the chair to sitting up straight so that their voice would be recognized. This shows the effectiveness of the combination of speech recognition and AR in motivating children to get 'understood' and gain new knowledge. Furthermore, this self-motivated and interactive learning experience increased the enjoyment and willingness of children to learn. This indicated a potential intrinsic motivation to learn a foreign language. During the interview, one of the participants said that he felt like talking to a robot that responds to his speech. Additionally, our participants claimed that using speech to interact with computers is novel, less hassle, and more natural compared to using the keyboard or mouse. Novelty is one of the components of optimal experience (Csikszentmihalyi, 1990), which was described by Hourcade (2015) as a necessity for children to achieve optimal learning experience. When a novel technology is used by children, they may learn optimally. However, if the children find the technology is difficult, they tend to disengage easily (Hourcade, 2105). Compared to speech, children in the speech-disabled non-AR group used a mouse to interact which they reported as less natural and less enjoyable. We noted that a few students had difficulty synchronizing their hand movements to move the mouse while looking at the screen to see where the hand was going. As opposed to a limited number of words that need to be said at one go in the Color and Shapes module, children doing the spatial relationship module need to complete a sentence consisting of three words. This could be the reason for a higher IMI enjoyment score in the mean rank of the non-AR non-Speech compared to non-AR Speech. This indicates that children prefer a shorter sentence than longer ones.

Interestingly, in the AR version of both modules, the enjoyment score remains high for both speech and non-Speech condition compared to the non-AR version. This could be because in this version children could play with AR markers to create different combinations. In the non-AR system, most facial expressions were smiling and happy, but we also rarely observed amazed expressions, in contrast with the AR system. This could be due to the lack of new elements during playing. Based on our experience during the ice-breaking session, the children were quite shy and tended to be more cautious, however, they were amazed when they saw themselves appear on the screen together with virtual objects in the AR system. Furthermore, the excitement of playing with AR markers and motivation to see the virtual objects appearing on the screen can surpass the challenges of misrecognition by the speech recognizer. We identified some factors which could lead to lack of enjoyment during a certain condition. Some participants with lower voices were unhappy when they said a word repeatedly but obtained late or no feedback from the system. Furthermore, a participant who was still familiarizing themselves with the English words often did not pronounce the words correctly which caused the misrecognition by the speech recognizer.

4. The easy-to-use tool help lower children's anxiety in second language learning

From our previous study (Dalim et. al, 2016) we have learned that short participants could not always extend their hands far enough to make the marker stay within the camera field of view. That caused the camera to not properly recognize the marker and the virtual objects were not properly displayed, so the participants felt confused and disengaged. Therefore in this study, we have carefully set up space by using a 'kid-sized' table and a webcam with a wider field of view. As a result, we managed to reduce the misrecognition problem and avoided physical distress among the children. In the spatial relationship module, we attached the camera to a tripod and placed it facing the screen. This was done in order to give a perspective view, close to how the children's eyes were looking. This camera placement was important to avoid a mirror effect, which may have confused young children when they read the sentence on the screen. Children were observed not

sticking to one place or only sitting down. Some of them preferred moving around and tried to show the marker to the camera from a different angle. This shows that our system manages to create a low anxiety language-learning environment.

In the color and shapes module, we observed that when using the non-AR condition children found it hard to move the mouse to the desired point on the screen and to choose which button of the mouse to press on although the button interface was designed large enough to facilitate the children's on-screen navigation. On the other hand, using an AR marker is more practical as the whole marker functions as the 'button', which just needs to be shown to the camera. This explains the higher mean rank on the ease of use for the AR teaching style than its comparable non-AR version. However, enabling and disabling the speech input in the AR version does give a significant difference to the ease of use score. Children rated AR non-Speech easier to use than the AR Speech version because, in this version, children are only required to pick up a color marker, followed by a shape marker, and show it to the camera to see the immediate result on the screen. All participants did this faster compared to the participants who had to repeat a word a few times for it to be recognized in the speech mode Children also rated the AR non-Speech significantly easier to use in the spatial relationship module compared to the non-AR Speech. The simple interface design that we have in the AR platform could be one of the contributing factors to the easiness of using the platform. Similarly, the non-Speech input is easier to use than the Speech input, which can be related to recognition difficulty by the speech recognizer.

The greatest challenge observed in the non-AR system was the difficulty for children to synchronize their hand movements when moving the mouse while looking at the screen. The force needed to physically press the mouse button was also a concern for most of the participants. A few participants used both hands to move the mouse and for pressing the mouse button (see Figure 20 left), and some other participants asked for help from the researcher (Figure 20 right). In both interfaces, the speech input worked quite well after at least three trials. However, there were situations where they could not recognize the commands well. Contributing factors include unavoidable ambient noises such as the noise from the ceiling fan, inarticulate speech, participants' low speech volume and speaking with an accent.

From our experience of conducting this research, there are a number of additional lessons learned that can be a guide for future researcher or developer when developing an AR application for children.

1. **Colorful AR markers**. Use primarily recognizable and colorful picture or icons as the pattern for the AR markers to create awareness in children and keep them enjoying learning. Colorful markers enable a large number of unique marker designs, and young children prefer to have more marker selections that they can play around with. A wider border or paddle outside of the trackable area should be provided for ease of handling and avoiding occlusion.
2. **Simple textual information**. Although young children are mostly preliterate, for second language learning including fairly simple text on the marker is appropriate to indirectly promote the reading of new language, and to further distinguish between markers. Children can associate the icon/picture to its spelling and pronunciation.
3. **Wider display size and larger field of view.** Young children found that one of the compelling features of AR is that they can see themselves on the screen together with the virtual object. Young children show a strong liking for PC screen rather than a smaller laptop screen. The larger display size makes the learning material and the children appear close to actual size, and this somehow makes them feel more immersed in the activity. The larger field of view of the camera, on the other hand, facilitates the capture of AR markers.
4. **Appropriate workspace setup.** If the learning material such as the AR marker involves displaying text to the camera and has opted a PC screen, it is very important to ensure that the camera used for tracking the AR markers and the whole scene is placed at the right position. Placing the camera facing the PC display is preferred to prevent the text from being mirrored or flipped, which can cause serious confusion to young children. In order to facilitate marker detection in this camera position, the AR marker should be either laid flat on the plane or held straight.
5. **Simple verbal prompts.** While reading is typically a challenge to young children, verbal instruction is preferred for their learning tools. In our prototype, the instructions were only given at the start page to let children focus on the task. However, during the pilot test, we noticed that young children have trouble remembering what to do next and refer to a teacher for instructions. We could improve the current prototype by embedding verbal instruction in the new language on the marker itself. Whenever an AR marker is in view, the verbal instruction is played. For example, when a child shows a color marker, an instruction "Say a color" is played to guide the child.
6. **Medium level immersion.** We recommend that AR-based applications for young children are not designed for full immersion, because they may have difficulties distinguishing between the real and augmented environments. This might cause further issues such as safety and misconception.
7. **Foster collaboration or small group learning.** When young children learn collaboratively or in a group, the learning activity will become more fun and significant.
8. **Age-appropriate content, consider abstract concepts instead of concrete.** AR has the capability to simplify one's thinking of abstract concepts because technology gives the ability to envision details. This was shown in our study where young children understand the difference between two- and three-dimensional shapes and were able to relate one to another. Therefore we recommend designers make use of this affordance to develop applications that help children understand age-appropriate abstract concepts such as space or time.

9. **Allow interactivity or physical activities**: Young children have a very short attention span. From our study, we noticed that young children more actively participated when they could move during the session. Selecting and arranging AR markers is an example of an activity that keeps them moving and staying alert.

6. Conclusions and Future Work

In this paper, we presented an AR system for teaching young children who are non-native English speakers about English terms for basic colors, 3D shapes, and spatial relationships. Based on our review of previous studies, our system is the first AR language-learning tool attempting to teach young children, four to six years old, about spatial relationships and 3D shapes. A user study was done with 120 children from six preschools in Malaysia. The objectives of this study were to see how effective our AR teaching strategy was compared to a non-AR strategy and to explore if the use of speech input was able to further increase the effectiveness of the AR learning system.

Our findings show that our AR system could be effective as a teaching tool for young children as it enhances the engagement in learning and increases knowledge gain. Real-time interaction enhances the children's excitement to explore the learning materials more. All participants had no difficulty in interacting with the system, even with only one demonstration. Our finding concludes that the speech-enabled AR interface is usable for younger children even with little or no experience, and provides a motivating factor for their foreign language learning. The results show significant evidence of knowledge gain and a positive inclination towards using the AR interface over the non-AR interface. Language learning, especially to young children, is best learned in a conducive environment which children do not feel forced to learn. Although children have to say the new words repeatedly in order to get correct pronunciation in the speech-enabled teaching platform of TeachAR, they completed the task with fun and eagerness to see what will come out. This is proven through the high IMI's (Enjoyment) score of our study which shows that the learning experiences have increased the children's motivation for learning the new language.

However, as with the majority of studies, the empirical findings of this study have to be seen in the light of some limitations. The primary limitation of the generalization of these results is the sample size. Although it is advisable to always work with a larger sample, in our case we managed to gather only 15 children per group who is convenient to test each of the condition. Statistically, this might be an issue, however, most research that involved young children participant work with less than 30 participants (Masmuzidin et al., 2018) which can be due to the low cost of providing equipment to a larger sample (Bacca et al., 2015) or limited time for data collection. Additionally, in our case, we concern more about selecting a representative sample that can contribute to idea generation.

Whether for learning or enjoyment, there are a number of practical considerations to consider when using this teaching tool. For example, in terms of the feasibility of using the Kinect sensor for audio capturing. Typically, the prototype that was used in this study will operate correctly when running on the practitioner's development computer with a Kinect sensor for capturing speech input. However, for convenience purpose, future work could consider a setup without using a Kinect sensor. With a slight modification in the prototype, the use of Kinect sensor in this study is replaceable with any microphone including the built-in microphone on computers or laptops running Microsoft Speech API, without much difference in terms of speech recognition performance. This also allows wider implementation as no additional devices such as Kinect sensor will be required, this will be relatively easier for teachers to run the application as no complex configuration is needed. The AR marker patterns used in this prototype can be replaced with any design that best represents the topic of interest and only needs to be updated in the main code and the XML file.

Besides that, environmentally, although it can be tough to control the noise level in the classroom, this is something of practical consideration when using the speech-enabled mode of TeachAR because if the noise level is too high, the speech recognition performance could deteriorate and the children would get frustrated. The teacher might also want to consider using collar mic wherever possible to better capture voices that are too low.

In the aspect of familiarity, teachers need to consider ensuring close guidance is available all the time when children use this tool. This is because learning a new language using new technology at the same time might potentially create anxiety. It is important to guide the children with the opportunity to familiarise themselves with the enhanced learning environment especially during the first time by providing instant feedback to their query.

The proposed teaching strategy could also benefit other areas of language learning such as sentence construction, training of accent and pronunciation, and also intercultural simulation activity for communication practice. Similarly, in these areas, augmented reality functions as the visual feedback while speech input as the mean for interaction.

In this study, we showed that the children did the task individually. It can also be done in small groups of children with some challenges to tackle such as misrecognition as children might cause more noise when in the group. In this case, the speech recognition part requires more training to adapt to the noise level. Having a few small groups of children increases the cost of preparation as the school need to provide a few sets of equipment. An alternative to that is by having all the children gathered in front of a large screen and do the marker arrangement together while speech recognition is done by taking turns. Future promising work might also look into improving the speech recognition to identify children's speech better. Continuous speech recognition could also be implemented in the future instead of using isolated word recognition.

This will add to the variety of sentences available for language learning. The fusion of multimodal input (Ismail, 2014) could be also used to create a more natural interaction by supporting simultaneous speech and gesture input. To make learning easily accessible, it is also recommended that the application be extended to a mobile version to provide portability. This would allow children to use the application everywhere they go.

In this research, marker-based AR was chosen because it is much simpler to detect things that are hard-coded in the application and it takes less processing power. However, with the advances in mobile hardware and software technologies, markerless AR tracking, such as using the Vuforia library (2016) could be considered for future work. This would provide a larger number of images that could be recognized.

Another extension of this work is to implement the TeachAR system in head-mounted display-based AR, for example using Microsoft HoloLense. While it will enable more mobility and interaction with both hands, this type of display is not designed for young children. However, it will open new opportunities to extend TeachAR for adult users with more complex vocabulary.

Overall, more research is needed to discover the potential of AR for young children's activities, especially in the ways that it can be used for language learning. This includes research on the appropriate AR interfaces for young children and how to create new learning experience using AR that can be used in different fields of education.

References

Ansawi, B. 2017., Promoting the 3Es (Exposure , Experience, Engagement) in an English-Rich Rural Primary School Community. *The English Teacher*, 46(1), 30.

Ariza, E. N., & Hancock, S., 2003. Second language acquisition theories as a framework for creating distance learning courses. *International Review of Research in Open and Distance Learning*, *4*, 1–9. Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/142/710

Azman, H., 2012. Implementation and Challenges of English Language Education Reform in Malaysian Primary Schools, 3L: Language, Linguistics, Literature®, *22*(3), 65–78.

Azuma, R., 1997. A survey of augmented reality. *Presence: Teleoperators and Virtual Environments*, *6*(4), 355–385. http://doi.org/10.1.1.30.4999

Bacca, J., Baldiris, S., Fabregat, R., & Graf, S., 2014. Augmented Reality Trends in Education: A Systematic Review of Research and Applications. *Educational Technology & Society*, *17*(4), 133–149.

Bacca, J., Baldiris, S., Fabregat, R., & Graf, S. (2015). Mobile augmented reality in vocational education and training. Procedia Computer Science, 75, pp.49-58.

Barreira, J., Bessa, M., Pereira, L. C., Ado, T., Peres, E., and Magalhes, L., June 2012. Mow: Augmented reality game to learn words in different languages: Case study: Learning English names of animals in elementary school. In *Information Systems and Technologies (CISTI), 2012 7th Iberian Conference* on, pages 1–6.

Beder, P., 2012. Language Learning Via an Android Augmented Reality, (September).

Billinghurst, M., 2002. Augmented Reality and Education. *New Horizons for Learning*, (figure 1), 21(3) 195-209. http://doi.org/10.4018/jgcms.2011010108

Billinghurst, M., & Dünser, A., 2012. Augmented reality in the classroom. *Computer*, *45*(7), 56–63. http://doi.org/10.1109/MC.2012.111

Boonbrahm, S., Kaewrat, C., and Boonbrahm, P., 2015. Using Augmented Reality Technology in Assisting English Learning for Primary School Students, pages 24–32. *Springer International Publishing, Cham*.

Bujak, K. R., Radu, I., Catrambone, R., Macintyre, B., Zheng, R., & Golubski, G., 2013. Computers & Education A psychological perspective on augmented reality in the mathematics classroom. *Computers & Education*, *68*, 536–544. http://doi.org/10.1016/j.compedu.2013.02.017

Chang, Y.-J., Chen, C.-H., Huang, W.-T., & Huang, W.-S., 2011. Investigating students' perceived satisfaction, behavioral intention, and effectiveness of English learning using augmented reality. *2011 IEEE International Conference on Multimedia and Expo*, 1–6. http://doi.org/10.1109/ICME.2011.6012177

Chen, C., & Su, C. C., 2013. An Integrated Design Flow in Developing An Augmented Reality Game for Enhancing Children Chinese Learning Experience. *International Journal of Digital Content Technology and Its Applications*, *7*(4), 907–915. http://doi.org/10.4156/jdcta.vol7.issue4.109

Cheng, H. J., Zhan, H., & Tsai, A., 2010. Integrating Second Life into a Chinese language teacher training program: A pilot study. *Journal of Technology and Chinese Language Teaching*, 1(1), 31-58.

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

Csikszentmihalyi, M., 1991. Flow: The psychology of optimal experience: Steps toward enhancing the quality of life. *Design Issues*, *8*(1), 80. http://doi.org/10.2307/1511458

Ford, C., 2014. "Children should start learning languages at age three." *The Telegraph*.

Fragoso, V., Kleban, J., & Gauglitz, S. (n.d.). TranslatAR : A Mobile Augmented Reality Translator on the Nokia N900, *2*.

Godwin-jones, R., 2016. Emerging Technologies Augmented Reality and Language Learning : From Annotated Vocabulary To Place-Based Mobile Games. *Language Learning & Technology*, *20*(3), 9–19. Retrieved from http://llt.msu.edu/issues/october2016/emerging.pdf

Hiew, W., 2012. English Language Teaching and Learning Issues in Malaysia: Learners' Perceptions Via Facebook Dialogue Journal. *Journal of Arts, Science & Commerce*, *3*(1), 11–19.

Hourcade, J. P., 2015. Child-computer interaction. http://homepage.cs.uiowa.edu/~hourcade/book/index.php

Hornecker, E., & Dünser, A., 2007. Supporting Early Literacy with Augmented Books – Experiences with an Exploratory Study. In *Proceedings of the 1st international conference on Tangible and embedded interaction - TEI '07* (p. 179).

Ibáñez, M. B., Di Serio, A., Villarán, D., & Kloos, C. D., 2014. Experimenting with electromagnetism using augmented reality: Impact on flow student experience and educational effectiveness. *Computers & Education*, 71, 1-13.

Idris, I., Salam, M. S. H., & Sunar, M. S., 2016. Speech emotion classification using SVM and MLP on prosodic and voice quality features. *Jurnal Teknologi*, 78(2-2), 27-33. DOI: 10.11113/jt.v78.6925.

Ismail, A. W., & Sunar, M. S., 2014. Multimodal Fusion: Gesture and Speech Input in Augmented Reality Environment. In *Computational Intelligence in Information Systems: Proceedings of the Fourth INNS Symposia Series on Computational Intelligence in Informational Systems* (INNS-C11S

2014). 331, 245, Springer.
Juan, C. M., Llop, E., Abad, F., & Lluch, J., 2010. Learning Words Using Augmented Reality. *2010 10th IEEE International Conference on Advanced Learning Technologies*, 422–426. http://doi.org/10.1109/ICALT.2010.123
Kato, H., & Billinghurst, M., 1999. Marker tracking and HMD calibration for a video-based augmented reality conferencing system. *Proceedings 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR'99)*. http://doi.org/10.1109/IWAR.1999.803809
Kaufmann, H., Steinbügl, K., Dünser, A., & Glück, J., 2005. General training of spatial abilities by geometry education in augmented reality. *Annual Review of CyberTherapy and Telemedicine: A Decade of VR*, 3, 65-76.
Krashen, S. D. (1982). Principles and practice in second language acquisition. Oxford: Pergamon
Kumar, A., Reddy, P., Tewari, A., Agrawal, R., & Kam, M., 2012. Improving literacy in developing countries using speech recognition-supported games on mobile devices. *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems CHI 12*, 1149. http://doi.org/10.1145/2207676.2208564
Li, S., Chen, Y., Whittinghill, D., & Vorvoreanu, M., 2014. A Pilot Study Exploring Augmented Reality to Increase Motivation of Chinese College Students Learning English. *2014 ASEE Annual Conference*. Retrieved from https://peer.asee.org/19977
Liarokapis, F., Mourkoussis, N., White, M., Darcy, J., Sifniotis, M., Petridis, P., Basu, A. and Lister, P. F., 2004. Web3D and augmented reality to support engineering education. *World Transactions on Engineering and Technology Education*, *3*(1), 11–14.
Lin, T.-J. and Lan, Y.-J., 2015. Language learning in virtual reality environments: Past, present, and future. *Educational Technology & Society*, 18(4):486–497.
Mahadzir, N. N. N., & Phung, L. F., 2013. The Use of Augmented Reality Pop-Up Book to Increase Motivation in English Language Learning For National Primary. *IOSR Journal of Research & Method in Education (IOSR-JRME)*, *1*(1), 26–38. http://doi.org/doi:10.6084/m9.figshare.1176011
Malinverni, L., Valero, C., Schaper, M. M., & Pares, N. (2018, June). A conceptual framework to compare two paradigms of augmented and mixed reality experiences. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*(pp. 7-18). ACM.
Masmuzidin, M. Z., & Aziz, N. A. A., Dec 2018. The Current Trends of Augmented Reality In Early Childhood Education. The International Journal of Multimedia & Its Application (IJMA), 10(6), 47-58.
Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of educational psychology*, *86*(3), 389.
Meda, P., Kumar, M., and Parupalli, R., Dec 2014. Mobile augmented reality application for Telugu language learning. In *MOOC, Innovation and Technology in Education (MITE), 2014 IEEE International Conference* on, pages 183–186.
Mispa, K., Mansor, E. I., Kamaruddin, A., & Hinds, J., 2016. Measuring usability and children's enjoyment of virtual toy in an imaginative play setting: A preliminary study. *Journal of Theoretical and Applied Information Technology*, *89*(1), 45–52.
Miyosawa, T., Akahane, M., Hara, K., & Shinohara, K., 2012. Applying Augmented Reality to E-Learning for Foreign Language Study and its Evaluation. *... On E-Learning, E- ...*. Retrieved from http://world-comp.org/p2012/EEE3387.pdf
Mroz, A., 2014. 21st Century Virtual Language Learning Environments (VLLEs). *Language and Linguistics Compass*, 8(8), 330-343.
Musa, N., Khoo, Y. L. and Hazita, A. (2012). Exploring English language learning and teaching in Malaysia. *GEMA Online® Journal of Language Studies, Special Section*. Vol. 12(1), 35-51.
Nielsen, J., & Landauer, T. K., 1993, May. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 206-213). ACM.
Norraji, M. F., & Sunar, M. S., 2016. WARna Mobile-based augmented reality coloring book. In *Proceedings of the 2015 4th International Conference on Interactive Digital Media, ICIDM 2015* [7516324] Institute of Electrical and Electronics Engineers Inc.
Pacheco, D., Wierenga, S., Omedas, P., Oliva, L.S., Wilbricht, S., Billib, S., Knoch, H., Ver- schure, P., 2015. A location-based augmented reality system for the spatial interaction with historical datasets. Digit. Herit. 1, 393–396. doi: 10.13140/RG.2.1.3957.4487.
Parhizkar, B., Oteng, K., Ndaba, O., Lashkari, A. H., & Gebril, Z. M., 2013. Ubiquitous Mobile Real Time Visual Translator Using Augmented Reality for Bahasa Language. *International Journal of Information and Education Technology*, 124–128. http://doi.org/10.7763/IJIET.2013.V3.248
Park, H., Jung, H. K., & Park, S. J., 2014. Tangible AR interaction based on fingertip touch using small-sized nonsquare markers. *Journal of Computational Design and Engineering*, 1(4), 289-297.
Phon, D. N. E., Ali, M. B., & Halim, N. D. A., April 2014. Collaborative augmented reality in education: a review. In *Teaching and Learning in Computing and Engineering (LaTiCE), 2014 International Conference* on (pp. 78-83). IEEE.
Plant, R. W., & Ryan, R. M., 1985. Intrinsic motivation and the effects of self- consciousness, self- awareness, and ego- involvement: An investigation of internally controlling styles. *Journal of Personality*, *53*(3), 435-449.
Razak, F. H. A., Hafit, H., Sedi, N., Zubaid, N. A., & Haron, H., 2010.Usability testing with children: Laboratory vs field studies. In *Proceedings-2010 International Conference on User Science and Engineering, i-USEr*, pp.104-109.Read, J. C., MacFarlane, S. J., & Casey, C., 2002, August. Endurability, engagement, and expectations: Measuring children's fun. In *Interaction design and children* (Vol. 2, pp. 1-23). Shaker Publishing Eindhoven.
Read, J., and Fine, K., 2005. Using survey methods for design and evaluation in child computer interaction. In *Workshop on Child Computer Interaction: Methodological Research at Interact*.
Rose, H., & Billinghurst, M., 1995. Zengo Sayu: An immersive educational environment for learning Japanese. University of Washington, Human Interface Technology Laboratory, Report No. r-95-4.
Saadiyah Darus., 2013. The current situation and issues of the teaching of English in Malaysia. Retrieved from http://r-cube.ritsumei.ac.jp/bitstream/10367/4130/1/LCS_22_1pp19-27_DARUS.pdf. 2013.
Saidin, N. F., Halim, N. D. A., & Yahaya, N., 2015. A review of research on augmented reality in education: Advantages and applications. *International Education Studies*, *8*(13), 1–8. http://doi.org/10.5539/ies.v8n13p1
Samihah, C., Dalim, C., Dey, A., Piumsomboon, T., Billinghurst, M., & Sunar, S., 2016. TeachAR : An Interactive Augmented Reality Tool for Teaching Basic English to Non-Native Children, 0–4. http://doi.org/10.1109/ISMAR-Adjunct.2016.39
Santos, M. E. C., Lübke, W., Taketomi, T., Yamamoto, G., Rodrigo, M. M. T., Sandor, C., & Kato, H., 2016. Augmented reality as multimedia : the case for situated vocabulary learning. *Research and Practice in Technology Enhanced Learning*. http://doi.org/10.1186/s41039-016-0028-2
Shapley, K., Sheehan, D., Maloney, C., & Caranikas-Walker, F., 2011. Effects of Technology Immersion on Middle School Students' Learning Opportunities and Achievement. *Journal of Educational Research*, *104*(5), 299–315. http://doi.org/Doi 10.1080/00220671003767615
Shelton, B., & Stevens, R., 2004. Using coordination classes to interpret conceptual change in astronomical thinking. *... of the 6Th International Conference on ...*, 634. Retrieved from http://inst.usu.edu/~bshelton/resources/CCs-astro_shelton-stevens.pdf
Solak, E., & Cakir, R., 2016. Investigating the role of Augmented Reality technology in the language classroom. *Online Submission*, *18(4)*, *1067-1085*.
Teoh, B., & Neo, T., 2007. Interactive multimedia learning: Students' attitudes and learning impact in an animation course. *The Turkish Online Journal of Educational Technology*, *6*(4), 28–37. http://doi.org/Thesis Sarjana Muda

Wagner, D., & Barakonyi, I., 2003. Augmented reality kanji learning. In *Proceedings - 2nd IEEE and ACM International Symposium on Mixed and Augmented Reality, ISMAR 2003* (pp. 335–336). http://doi.org/10.1109/ISMAR.2003.1240747

Weiland, C., & Yoshikawa, H., 2013. Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, 84(6), 2112-2130.

Yoon, T., 2014. The Application of Virtual Simulations Using Second Life in a Foreign Language Classroom, *2*(1), 67–75.

Zhou, Z., Cheok, A. D., Pan, J., & Li, Y., 2004. Magic Story Cube : an Interactive Tangible Interface for Storytelling, 364–365.

Children IMI Interest/Enjoyment Scale., 2017. http://hmi.ewi.utwente.nl/puppyir/results/user-evaluation-toolkit/children-imi-interestenjoyment-scale/ Accessed 6 February 2017.

LearnAR, 2015. https://learninglovers.org/2015/05/06/learnar/ (Accessed 13 May 2018).

ARToolkit, 2017. https://github.com/artoolkit/arunity5/ (Accessed 6 February 2017).

Kinect for speech recognition, 2017. https://developer.microsoft.com/en-us/windows/kinect / (Accessed 6 February 2017).

Second Life.2017. http://secondlife.com/ (Accessed 6 February 2017).

Unity 3d.2017. https://unity3d.com/. (Accessed 6 February 2017).

Getting Started.2016. https://library.vuforia.com/getting-started.html/. (Accessed 25 September 2016).

(17095 words)