Game-Based Augmented Reality versus an Interactive 2D App: A Comparative Study on Learning and Interest in Late Primary and Early Secondary Education

Teun Boekel - 6949355

6514 words

Faculty of Social and Behavioral Sciences, Utrecht University

Master Thesis

First assessor: Michaela Arztmann

Second assessor: Jacqueline Wong

June 12th, 2023

Abstract

The use of immersive technologies, such as augmented reality (AR), in the classroom is becoming increasingly popular as a means to engage students and visualize abstract concepts. However, the effectiveness of AR in late primary and early secondary education, specifically for subjects like chemistry, is still unclear. This study aimed to contribute to the ongoing discussion about the impact of AR on learning and interest by comparing the effects of a game-based AR condition with an interactive 2D environment in a real classroom setting. The research question is: "Is there a difference in learning outcomes and interest between the use of game-based augmented reality and an interactive 2D environment for chemistry lessons for children in late primary and early secondary education." The game-based AR group outperformed the 2D group in terms of learning, with a small to medium effect ($\eta^2_p = .06$). There was no difference in interest found between the two groups. However, the last conclusion should be interpreted carefully, since the power was not met, and glitches during the gameplay might have had an impact on the participants' interest.

Keywords: Augmented reality, game-based learning, interest, motivation, presence, agency, early secondary education, late primary education, chemistry education, media comparison.

Table of contents

Introduction

Because of the digitalization of the classroom, immersive learning in virtual and augmented reality is not science-fiction technology anymore. Instead, immersive learning is a great opportunity for teachers who want to engage students and visualize otherwise abstract concepts (Garzón & Acevedo, 2019). From all immersive learning technologies, augmented reality (AR) is the easiest to use in the classroom this day. This AR technology is typically used with smartphones and tablets, which are already widely available in the classroom.

However, due to the novelty of the technology, research focusing on the effectiveness of AR is still scarce. A recent meta-analysis has shown that (game-based) AR apps are in general more effective than traditional lectures, but also more effective than technologies like games, simulations, and virtual worlds (Garzón & Acevedo, 2019). However, creating these is labor-intensive and learning is not always more effective than less complex teaching methods (e.g., Hung et al., 2016; Cai et al., 2017). Especially for students in primary and early secondary education it is still unclear whether immersive environments, and in particular AR environments, are more effective for supporting knowledge acquisition and motivation compared to a less immersive learning environment (Garzón & Acevedo, 2019). Nevertheless, using immersive learning technologies could help to increase students' motivation for subjects like chemistry (Garzón et al., 2019; Radu, 2012). This could be particularly relevant for younger students, since students motivation for science subjects is already decreasing in early secondary education (Höft et al., 2017).

What is immersive game-based learning?

Immersive learning is a promising new technology in education because the learners' presence and agency are higher than with other instructional methods (Makransky & Petersen, 2021). Presence is defined as the experience of 'being there' Sheridan, 1992). Agency is the feeling of generating and controlling actions (Moore & Fletcher, 2012). According to Makransky and Petersen (2021), this high presence and agency can facilitate affective and cognitive factors (such as motivation and self-regulation) that foster learning outcomes. Immersive learning usually refers to virtual reality or augmented reality environments.

Augmented reality (AR) as an immersive learning environment can provide presence and agency to a certain extent (Rosa et al., 2019). Akçayır and Akçayır (2017) define AR as a technology which overlays virtual objects (augmented components) into the real world. In practice, AR technology is used on devices

with cameras and touchscreens like smartphones and tablets. Learners in AR environments can alter digital objects in the environment, engage in realistic simulations, and become 'immersed' in the augmented reality environment.

Nowadays, augmented reality is often combined with game-based learning (Alper et al., 2021). Game-based learning refers to a game whose primary objective is not solely entertainment, but rather the use of the game's interactive and engaging qualities to support training and education (Wouters et al., 2013). Just like with immersive learning, game-based learning provides high agency, which has benefits for motivation (Sitzmann, 2011). Several meta-analyses have shown that game-based learning is more effective for learning and retention than conventional teaching methods (Vogel et al., 2006; Wouters et al., 2013; Arztmann et al., 2021). However, this does not mean that games should be the solution to every learning problem. According to Gros (2007), games are useful way to enhance the understanding of complex concepts, but not so much when the goal is text comprehension. In that case, other media like textbooks or lectures are more suitable (Gros, 2007).

Learning outcomes of game-based AR

Although the meta-analysis by Garzón and Acevedo (2019) found a positive effect for (game-based) AR-learning, there was a high variability in effect sizes between studies. Garzón and Acevedo (2019) concluded that the positive impact of AR on learning is greater for mature students at universities and vocational colleges compared to younger students of late primary or secondary education. Challenges such as difficulty in navigating the systems (Garzón et al., 2019; Herpich et al., 2014), the ability to handle multiple tasks at once (Radu, 2014), and the amount of information presented (Dunlevy et al., 2009) can have an impact on the experience of young users of AR applications. Despite these results, Garzón and Acevedo still found a medium effect of AR on learning for primary and secondary education.

However, research focusing on the learning effects of AR games in late primary and early secondary education is scarce (Garzón & Acevedo, 2019). Thus far, AR has ben found to improve learning outcomes in this group (Cai et al., 2014; Chen & Wang, 2015; Cheng et al., 2015; Huang et al., 2016; Kamarainen et al., 2013; Karagozlu et al., 2018; Liu & Chu, 2010), but those studies differ in their design and generalizability. Huang et al. (2016) compared a guided tour, an AR tour and a guided AR tour for learning plant biology. They found that the guided AR tour scored significantly higher on learning than the two other groups, who did not differ significantly. This result is interesting, but might not be generalizable to a classroom setting. Liu and Chu (2010) compared an AR English learning trajectory with traditional lectures for 8 weeks, and

concluded that the AR group scored higher on learning. Furthermore, Karagozlu et al. (2018) also compared AR with traditional lectures and found that the AR condition not only had higher learning outcomes, but also increased confidence in problem-solving skills. However, these AR lessons from Karagozlu et al. were not game-based. Other studies also found that AR improved learning, but did not follow an experimental control group design (Cai et al., 2014; Chen & Wang, 2015; Cheng et al., 2015; Kamarainen et al., 2013).

Not all studies found that AR improved learning. Cai et al. (2016) compared AR-learning with traditional lectures and Hung et al. (2016) compared AR with a picture book. Both did not find a significant difference for learning between the two conditions, although it should be noted that Hung et al. did not use a game-based learning approach. An important note is that most of the mentioned studies were conducted with a small number of participants. Only Karagozlu and Chen & Wang reached substantial participant numbers. Thus, the effectiveness of AR in early secondary education remains still unclear.

The reason not all comparative studies showed that AR was significantly better for learning, might not only be the participants' age, but could also be due to the subject of the material. Garzón and Acevedo (2019) found variability in effect size for the subject, and concluded that AR was most effective in engineering, manufacturing, and construction. According to Makransky and Petersen (2021), this is because immersive learning is optimal for learning procedural knowledge[1]. Immersive learning provides the presence and agency to rehearse procedures in an environment that feels realistic, while the procedure can be broken down into comprehensible parts and learners can fail without consequences (Makransky & Petersen, 2021).

In late primary and early secondary education, declarative knowledge[2] seems to prevail over procedural knowledge in science subjects (Garzón & Acevedo, 2019). Research investigating the effects of immersive learning environments on declarative knowledge have shown mixed results, varying from no significant effect (Makransky et al., 2019a) to a significant effect for declarative knowledge gains (Webster 2016). Makransky et al. (2019b) even found a negative effect on declarative knowledge of immersive learning environments compared to a less immersive 2D simulation. Since immersive learning environments rely on visual and interactive elements to convey information, it may not be more effective at conveying abstract or conceptual knowledge (Makransky & Petersen, 2021) as other methods such as reading or

[1] Procedural knowledge is knowledge about how to do something (Anderson et al., 2001), for example flying a plane or working with industrial machines.

[2] Declarative knowledge is either conceptual or factual knowledge, for example knowing how to count to ten in Spanish or knowing what a democracy is.

watching lectures (Gros, 2007). In addition, the immersive aspect may also be less conducive to the kind of deep, reflective thinking that is often required to truly understand and retain declarative knowledge.

Motivation and interest effects of game-based AR

Besides learning, increasing motivation is also an important objective of game-based AR applications in education (i.e. Arztmann, 2022a; Huang et al., 2016). The specific application for this study, Marie's Chem Lab, is a game-based AR learning app designed to increase children's interest for learning chemistry (Arztmann et al., 2022). According to Deci (1992), interest is a strong motivator that encourages learners to take part in their education. Based on the self-determination theory, interest is part of intrinsic motivation and is defined as "the affect that changes oneself to activities that provide the novelty, challenge or aesthetic appeal that one desires at time" (Deci 1992, p. 45). Intrinsic motivation has shown to be highly effective for learning (e.g., Vansteenkiste et al., 2009), and therefore increasing students interest in the learning material could improve academic participation and performance.

A meta analysis by Garzón et al. (2019) has shown that students reported feeling more motivated by using AR applications compared to other pedagogical tools. Additionally, Radu's (2012) comparative review found that the use of AR increased students' motivation, as they had fun while learning and were eager to repeat the AR experience. Makransky and Lilleholt (2018) explain these findings by stating that the high presence students experience from immersive technologies influences motivation and enjoyment. When looking at studies within the age group of this study, Di Serio et al. (2013) demonstrated that including AR in learning environments acted as a motivating factor for students in a visual art course. Also, Kamarainen et al. (2013) found that AR increased motivation for biology, but did not compare their environment with a control condition. With only one study that compares an AR application for late primary and early secondary education, it remains somewhat unclear whether game-based AR is also increasing young students' motivation compared to traditional learning environment.

This becomes evident when looking at the results of a meta-analysis on game-based learning in general. Wouters et al. (2013), found that game-based learning does not always seem to be motivating for students. One explanation could be due to the increased focus on instructional design quality in the game design compared to the more engaging aspects of games, which makes the game less engaging and motivating (Wouters et al., 2013). Second, the setting of a game-based learning intervention could influence the feeling of autonomy (Wouters et al., 2013). Although the learners play a game, this game is selected by

the teacher, they are still at school, and play the game involuntarily at a fixed time-slot. This low autonomy decreases intrinsic motivation (Deci & Ryan, 2000). However, another meta-analysis by Arztmann et al. (2021) showed that game-based learning results in higher motivation than traditional classroom settings for STEM (science, technology, engineering, and mathematics) subjects. Game-based learning, which could be combined with AR-learning, is especially beneficial for children in primary school (Arztmann et al., 2021). Thus, the effectiveness of game-based learning on motivation is still an area of debate, with some studies finding no improvement and others finding benefits.

Present study

Thus far it is still unclear whether immersive learning environments actually benefit young students. Therefore, the aim of this study is to investigate the impact of game-based AR chemistry education on the learning outcomes and interest levels of late primary and early secondary education students. To determine whether the immersive learning environment provided by game-based AR benefits students, the game-based AR condition is compared to a 2D environment that includes the same information and interactive elements. With this comparison, the value of the AR affordances for learning and interest can be assessed.

For this objective, the following research questions were formulated, with the first of the two being: “is there a difference in learning outcomes between the use of game-based augmented reality (AR) and an interactive 2D environment for chemistry lessons for children in late primary and early secondary education?” Previous studies indicated that although AR is an effective method for learning, immersive learning might not be more effective at conveying declarative knowledge than a non-immersive 2D environment (Makransky et al., 2019b; 2020; Makransky & Petersen, 2021). Since the AR environment is mainly teaching declarative knowledge, it is not expected that the game-based AR environment will be more beneficial for learning than the 2D environment.

The second research question is: “is there a difference in interest between the use of game-based augmented reality (AR) and an interactive 2D environment for chemistry lessons for children in late primary and early secondary education?” According to Wouters et al. (2013), game-based learning is not necessarily more motivating than other types of lessons. However, since game-based learning is motivating for STEM subjects (Arztmann et al., 2021), and meta studies by Radu (2012) and Garzón et al. (2019) found that AR motivates students, it is hypothesized that interest is higher for children in the game-based AR condition.

Method

Design

The present study followed a quasi-experimental design using a pre- and post- knowledge test, and a pre- and post- interest questionnaire. Therefore, both research questions are a 2x2 mixed designs.

Participants

In total, 108 children participated in this study. When looking at the demographics, the participants were 12 years and 8 months old on average. The oldest was 15 and the youngest 11, and they were evenly distributed across both conditions $t(106) = 1.17$, $p = .244$. In total, 55 boys, 52 girls and one non-binary person participated. These were also evenly distributed across the conditions ($\chi^2(2) = 2.82$, $p = .275$). As for the education level, participants were not evenly distributed across the conditions ($\chi^2(3) = 12.34$, $p = .006$). In the Dutch education system, group 8 is the last year of primary school, and there is no differentiation yet. After group 8, children move to secondary school. Based on a centralized test performance and a teacher advice, they either go to (from low to high) 'VMBO basis', 'VMBO kader', 'VMBO theoretisch', 'HAVO' or 'VWO'. Table 1 shows the distribution of education level for this study. Adjusted standardized residuals were calculated as a post-hoc test. This showed that only the second year VWO group was not evenly distributed.

Table 1

Crosstabs for education level, with adjusted standardized residuals between brackets

Level	2D ($z_{adj.}$)	3D ($z_{adj.}$)	Total
Group 8	15 (-1.61)	23 (1.61)	38
1st year of HAVO	18 (1.29)	12 (-1.29)	30
2st year of HAVO	21 (1.88)	12 (-1.88)	33
2st year of VWO	0 (-2.74*)	7 (2.74*)	7
Total	54	54	108

**significant; exceeding 1.96 and -1.96 (Field, 2018)*

Procedure

Before the experiment started, the ethical committee of the University gave their permission for the study. Also, parents were informed about the study with an email. They were informed about the aims of the

experiment, as well as the option to withdraw their child from the study. Moreover, the participants themselves were also informed about the general aim of the experiment, their anonymity, and their right to withdraw.

The lessons took place in a classroom setting during regular school hours. The participating classes were randomly assigned to a condition. This also explains why it was difficult to distribute the education levels equally over the conditions. Both the game-based AR group and the 2D group worked with iPads. These iPads were provided by the researchers.

The participants started with the pre-test to assess their knowledge. They were also assessed on their interest before and after the experiment. The interest questionnaires were integrated in the AR game and the 2D lesson. The knowledge questions were collected with Qualtrics (2023). The participants went to the posttest in Qualtrics when they were finished with the game. Also, participants were directed to the posttest when 30 minutes of game time had passed to ensure the experiment would not exceed one lesson hour and therefore would not interfere with the schools' schedule.

Instruments

Maries ChemLab

Marie's ChemLab, is a single-player AR game designed for tablets, specifically targeted at grade 11 to 15 year olds (Arztmann et al., 2022). The game has been iterated and tested for usability (Arztmann et al., 2022). Screenshots of Marie's ChemLab can be seen in Figure 1.

The aim of Marie's ChemLab is to trigger curiosity and intuitive acquaintance with chemistry. The game is divided into three levels, set in different locations (kitchen, forest, and lab). Each level teaches the player about a different chemical concept through various tasks. There is also an intelligent agent (Marie) who provides helpful advice and reminders throughout the game. At the end of each level, a summary of key concepts is displayed.

Each level has its own specific learning objective (Arztmann et al., 2022a). In level one, the goal is for students to learn about sourness and how it can be measured using pH-strips. Level two builds on this by introducing a natural indicator as another option for measurement. Finally, level three focuses on applying the knowledge and skills gained in the previous levels to different materials and tools, including liquids commonly found in the kitchen, to demonstrate that they can also be acidic.

Interactive 2D environment

The aim of the 2D control group (Appendix A, see Figure 2 for screenshots) is to provide the same information as the AR game. This means that the environment was designed with the same learning goals in mind, provided the same examples and images as the participants would see in the game. The 2D environment is less immersive, but still holds the interactive elements (i.e., ranking fruits in sourness and pH-testing products) of the game-based AR environment. All AR activities have an interactive 2D counterpart in the control group. The 2D environment is developed with Xerte (2022), which is community-backed software for creating (game-based) learning materials.

Figure 1

Screenshots from Marie's ChemLab

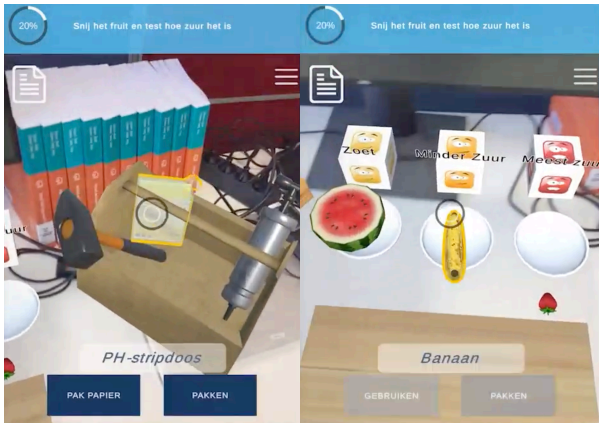


Figure 2

Screenshots from the interactive 2D environment

Interest

Situational interest was measured on a 5-point Likert scale which is based on Nuutila et al. (2021). This same questionnaire is used in the pretest and posttest. The original question "this task seems interesting" was adjusted to fit the environment and as such the term "task" was exchanged with "game". This

questionnaire contains a single item to not disrupt the gameplay. Short-scale questionnaires are believed to be less reliable due to their susceptibility to random measurement (Credé et al., 2012). Therefore, it is not possible to prove validity with a confirmatory factor analysis. Although this is a limitation, short scales are still reasonable alternatives to their corresponding long scales when study designs require brief measures (Gogol et al., 2014).

Learning outcomes

Learning was measured with a pretest and a posttest (Appendix B). The knowledge that is tested with these questions is factual and declarative, since they consist of facts about everyday chemistry. Participants got a multiple-choice test, and the distractors were formulated by combining common errors or misconceptions and plausible correct answers (Collins, 2006). These questions and their answer options were formed in English, and translated to Dutch using the back-translation method. This entails that the English questions were translated to Dutch by a native speaker, and translated back to English by a native speaker again. Both English versions were compared, and they did not differ in meaning.

To establish validity of the instrument, a pilot was conducted. 38 children participated in the pilot, of which 26 completed the posttest. The pilot showed that one question (Q4) must be removed because there was more than one correct answer option. To decrease chance of guessing a correct answer, the questionnaire was changed from three answer options to four answer options (Dehnad et al., 2014; Bachman & Palmer, 1996). These later formulated answer options can be seen. By looking at the ‘option D’ in Appendix B. No items were removed based on their item-rest correlation due to the adjustments in the questionnaire. These adjustments might impact the reliability in such a way that questions could be unjustly removed. Additionally, considering the questionnaire's relatively short length of eight items, it was deemed more secure to retain the same number of questions and assess the reliability at a later stage. The item-rest correlations from the pilot can be seen in Appendix B, Table 5.

Besides establishing the validity of the questionnaire, the pilot was also used prevent ceiling- or floor effects (Uttl, 2005). This means that questions could be removed or adjusted if the test was too easy or too hard. Since mean scores were not noticeably high or low, the level of the test was considered as suitable for the participant’s knowledge level. Both the pre- and posttest consisted of the same questions, but these were asked in a different order and answer options were shuffled to prevent retention effects.

Data analysis

The statistical analyses were carried out with JASP (2022). The pretest functioned to detect high pre-knowledge, and to detect negative outliers on the posttest. Outliers were found by calculating *Z*-scores, and the threshold for exclusion from the study was 3.0 (Tabachnick & Fidell, 2013). No *z*-scores exceeded the threshold values in the pretest, and only one participant exceeded this in the posttest with a *z*-score of −3.01. This participant only answered one question correctly, while having three correct answers on the pretest. This pointed towards a deficiency in engaged contribution to the study. Reliability was measured using the Guttman's λ2 because it is more precise than the commonly used Cronbach's α (Hessen & Van Erp, 2020). Normality was assessed with three values, namely the Skewness, Kurtosis (Field, 2018) and Shapiro-Wilk. The Shapiro-Wilk was preferred over the Kolmogorov-Smirnov since Razali and Was (2011) state that it's the most powerful normality test. Also, a quality check was done to see if pre- and posttest outcomes for learning correlate. The absence of such correlations could indicate guessing, confusion or knowledge loss, meaning that either the questions or the material lack quality. Moreover, this quality check is used to see if interest explains knowledge scores. Dropouts on either the pretest or the posttest were excluded from the study. Because the knowledge data was collected with Qualtrics, and half of the interest data was collected within Marie's ChemLab, technical errors in both programs made it possible that learning data was stored and interest data not. Because of this, dropouts were excluded per research question, meaning that participant numbers slightly differ per research question (n = 97 for RQ1, *n* = 85 for RQ2). Moreover, in some cases participants had to restart Marie's ChemLab due to glitches and made the posttest twice. Therefore, for interest as well as knowledge, the first measure was chosen because it allowed for a more controlled and consistent assessment of the participants' performance, minimizing any potential confounding variables that could arise from conducting multiple posttests or seeing the learning material twice. To answer the first research question: "is there a difference in learning outcomes between the use of game-based augmented reality (AR) and an interactive 2D environment for chemistry lessons for children in late primary and early secondary education?", a 2x2 ANOVA was conducted. As for the assumption checks, the equality of variance was tested with the Levene's test, and the random distribution for the pretest was tested with a Mann-Whitney *U*-test. The second research question: "is there a difference in interest between the use of game-based augmented reality (AR) and an interactive 2D environment for chemistry lessons for children in late primary and early secondary education?", was answered with an ANCOVA. Random distribution for the pretest was also tested with a Mann-Whitney *U*-test, and homogenous regression was tested by adding an interaction term with the covariate and the outcome variable to the initial ANCOVA model.

Results

The assumption checks, the analysis for the effect of performance- and mastery-oriented conditions and the mediation analyses are described separately in this section. Due to dropout and data loss because of technical issues, 97 of the initial 108 children participated for research question one, and 85 participated for research question two. The desired power for both research questions was at least 80% since it represents a reasonable balance between Type-I and Type-II errors (Cohen, 1992). This was met for research question one, with a post-priori power of 83.27%, at a medium effect size of $f = .15$ (HHU, 2022). For research question two, the desired power was not met. With a medium effect size of $f = .15$, the power was 27.70%.

Table 2

Descriptive statistics (SD = standard deviation, N = participants) for learning (scale: 1-7) and interest (scale: 1-5) in the pretest and posttest

		2D condition			game-based AR condition		
		Mean	SD	N	Mean	SD	N
Learning	Pretest	3.8	1.0	54	3.5	1.0	54
	Posttest	4.1	1.1	52	4.3	1.1	45
Interest	Pretest	3.2	1.0	52	4.1	0.9	52
	Posttest	3.2	1.1	48	3.8	1.4	37

Note: these are the descriptives from the full sample. Descriptives slightly differ for the analyses.

Descriptive statistics and quality check

The descriptive statistics for both learning and interest are displayed per condition in Table 2. To assess the quality of the learning test and the interest questionnaire, correlations were calculated with Spearman's ρ. The knowledge pretest correlated significantly, but weakly with the knowledge posttest, indicating that participants with higher pre-knowledge also scored high on the posttest. Furthermore, the interest pretest correlated moderately with the interest posttest. There were no correlations found between interest and knowledge for both the pretests and the posttests, meaning that interest did not influence the knowledges test scores. The correlation values are displayed in Table 3.

Table 3

Spearman's ρ correlations between the dependent variables

			ρ	p
Learning pretest	-	Learning posttest	.24	.020
Interest pretest	-	Interest posttest	.30	.006
Learning pretest	-	Interest pretest	.02	.809
Learning posttest	-	Interest posttest	.16	.133

Preliminary analyses

Reliability

The reliability was low on the pretest λ2 = .22 and the posttest λ2 = .36. Item-rest correlations and scores per question can be seen in Appendix B. No item-rest correlations were higher than the threshold value of .30 (Henrysson, 1963). Although the low λ2 and item-rest correlations seem problematic, this is to be expected when testing a knowledge domain according to Taber (2018). Taber states that testing science knowledge usually means assessing a range of distinct knowledge facets, without focussing on internal consistency like a psychometric test would do. Therefore, the low reliability scores were not seen as a reason to eliminate questions. One exception was the question: "blackberries don't become more acidic once they ripen." It was removed from the data analysis since the item-rest correlation was negative on the posttest: *rir* = -.05, and because several participants raised their hands for further explanation during testing, indicating that the negative phrasing made it difficult to understand the question.

Normality, equality of variances, homogenous regression

The data for the learning and interest in the pretest and posttest were not normally distributed per condition. Although Kurtosis was between the cutoff values of -2 and 2 for all variables, the Skewness was between the cutoff values of -1 and 1 for all variables except the knowledge pretest in the 3D condition (Field, 2018). Moreover, the Shapiro-Wilk was significant for all variables (Razali & Was, 2011). According to Blancha et al. (2017) this is not problematic, as *F*-tests are robust regardless of their normality.

Regarding the first research question on learning outcomes, Levene's test showed that variances did not differ for the knowledge pretest: $F(1, 95) = 0.10$, $p = .757$, and the posttest: $F(1, 95) > 0.01$, $p = .957$. For the ANCOVA regarding the research question with interest as the dependent variable, variances were equal: $F(1, 83) = 2.05$, $p = .156$. Moreover, there was homogenous regression, meaning that there was no interaction between the outcome variable and the covariate: $F(1, 81) = 0.26$, $p = .610$.

Learning difference between game-based AR and interactive 2D condition

Pretest learning distribution

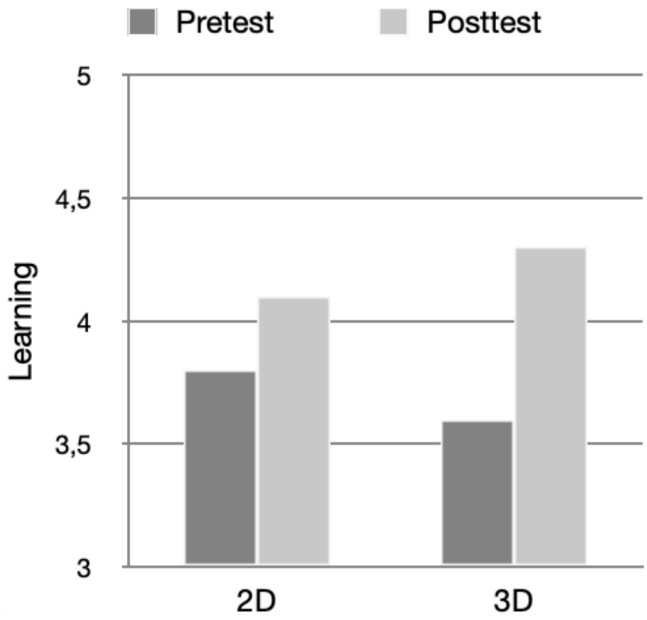
To check whether pretest knowledge was evenly distributed over the conditions, a Mann-Whitney *U* test was used, because the Shapiro-Wilk test showed that pretest scores were not normally distributed. This showed that pretest knowledge was evenly distributed between the 2D condition and the game-based AR condition: $U = 1756.5$, $p = .056$. Because of the unequal distribution of the education level, this random distribution analysis was also done without the second year VWO group. VWO is the highest education level in this sample, which could influence the results when not distributed equally. However, without the second year VWO group, distributions were still equal: $U = 1488.5$, $p = .120$.

ANOVA

The first research question is answered with a 2x2 mixed ANOVA, with condition (2D or game-based AR) and testing moment (pretest and posttest) as independent variables, and learning as the dependent variable. The descriptive statistics can be seen in Table 2, these are visualized in Figure 3. There was a main effect for learning*:* $F(1, 95) = 22.64$*,* $p < .001$*,* $\eta^2_p = .19$*.* Participants scored higher in the posttest (M = 4.2, SE = 0.1) than in the pretest (M = 3.6, SE = 0.1). There was no main effect for condition*:* $F(1, 95) = 0.66$*,* $p = .419$*,* $\eta^2_p = .01$*.* Participants did not score higher in the game-based AR condition (M = 3.8, SE = 0.1) than in the 2D condition (M = 3.9, SE = 0.1). Lastly, there was an interaction effect between learning and condition: $F(1, 95) = 5.75$*,* $p = .018$*,* $\eta^2_p = .06$. These conclusions did not change when the second year VWO group was excluded from the study, meaning that the unequal distribution did not affect the results for research question one in a meaningful way.

Figure 3

Interaction between condition and testing moment: with the conditions on the x-axis, learning on the y-axis and testing moment displayed as bars



Note: actual learning scale is from 1 to 7.

Interest difference between game-based AR and interactive 2D condition

Pretest interest distribution

For the equal distribution check, a Mann-Whitney *U* test was used in the pretest for interest. This showed that interest was not evenly distributed between the 2D condition and the game-based AR condition: $U = 653.5$, $p < .001$. The 3D group scored higher on the pretest (Median = 4) than the 2D group (Median = 3, see Table 2 for the mean scores).

ANCOVA

Because of this randomization issue, research question two was answered with an ANCOVA, using the posttest as the dependent variable and the pretest as a covariate. The ANCOVA showed that there was no significant difference between the 2D condition and the game-based AR condition for interest: $F(1, 82) = 1.80$, $p = .183$, $\eta^2_p = .02$. The mean scores can be seen in Table 2.

Discussion

Theoretical contributions

The aim of this study was to investigate whether learning outcomes and interest differ between the use of game-based AR and an interactive 2D environment for basic chemistry lessons for children in late primary and early secondary education. As for the first research question, it showed that participants in the game-based AR condition learned more than participants in the 2D environment with a medium effect ($\eta^2_p = .06$). The finding that game-based AR improves learning more than the 2D environment is in line with the meta-analysis by Garzón and Acevedo (2019), and the comparative studies by Liu & Chu, 2010, Karagozlu (2018) and Huang et al (2016). It is also interesting that this study provided no content-related guidance from teachers or researchers, but only technical help during the experiment. Even with this limited assistance, the game-based AR group still managed to improve their learning more than the 2D group. This contradicts the findings of Huang et al (2016), who found that AR only worked better for learning with a guide.

With this finding, this study places a noteworthy question to the CAMIL model by Makransky and Petersen (2021). The aim of this model is to combine existing studies on immersive learning in a theoretical framework to describe the process of learning in immersive environments. Based on studies by for example Makransky et al. (2019b), they conclude that immersive environments are not the ideal medium for learning declarative knowledge. However, this study shows that the immersive game-based AR condition outperformed the interactive 2D condition on a declarative knowledge test. Makransky and Petersen do recognize that the design of an immersive environment plays a considerable role. Thus, there could be design elements from Marie's ChemLab that foster learning declarative knowledge in a way that other immersive learning applications do not achieve. It would be mere speculation to subtract these elements. An important note is that the CAMIL model is mainly based on virtual reality instead of augmented reality, which is another medium for immersive learning. Therefore, it might not be design elements, but the medium of game-based AR itself that fosters learning declarative knowledge.

As for the second research question, there seems no difference in interest between the use of game-based AR and an interactive 2D environment for chemistry lessons for children in late primary and early secondary education. This is peculiar, since meta studies (Radu, 2012; Garzón et al., 2019) and studies within the age group (Di Serio et al., 2013; Kamarainen et al., 2013) indicate otherwise. Therefore, the interest for digital learning apps might be explained better by interactivity (Mahle, 2011) that immersion, since both conditions in this study were interactive and interest did not differ significantly. This reasoning

follows from the lack of interactive elements in the control condition from Di Serio and colleagues' paper, and non-existence of a control condition from Kamarainen and colleagues. This conclusion should however not be drawn too rapidly, since more evidence from additional papers is necessary to support it.

When looking at the descriptive statistics, the interest levels for both conditions are relatively high on the scale (combined mean = 3.5), indicating that there is no clear indication which type of intervention, game-based AR or interactive 2D, teachers should use if they want to engage their classroom. While other papers (i.e., Garzón et al., 2019) would suggest using game-based AR, this study suggests that interactive 2D might be just as effective. An important note on the high interest is that this could be attributed to a novelty effect because there were iPads and headphones on the tables in the classrooms specifically for this intervention. The novelty effect is described as an increased motivation for something due to its newness (Koch et al., 2018), and can be created by innovative content or information technology (Huang, 2020). This increased motivation can wear off when people get familiar with this technology. If the aim is to increase interest over a longer period, a longitudinal study on game-based AR and interest is needed to examine how children's interest develops.

The finding that there is no interest difference between the two conditions contradicts the CAMIL model by Makransky and Petersen (2021), since they state that immersive environments with high presence and agency increase the interest of participants. Moreover, they state that, based on a study from Harackiewicz et al. (2016), interest promotes learning by increasing the learner's attention and engagement. This relationship is not observed in the present study, as there was no correlation between learning and interest. Based on the mentioned contradictions to the CAMIL model, two conclusions can be drawn. The first one is that the paths between the concepts in the CAMIL model need to be revised. This would not be unusual, since the model is still conceptual, and more research is needed to justify it (Makransky & Petersen, 2021). A second conclusion could be that game-based AR differs so substantially from other immersive learning techniques that it needs a different conceptual model to better understand how it works.

Limitations

Limitations of this study are firstly that the game was still in development during testing. There were several glitches that stopped the gameplay. In many cases, participants had to completely restart the game. Furthermore, in some cases the environment scanning did not work properly, which meant that participants had to wait a few minutes to play the game. Although it is not possible to state with certainty, this could have negatively influenced the interest in the game-based AR condition. This is backed by looking at the

dropout rates for both conditions. Two participants dropped out in the 2D condition, against nine in the game-based AR group.

Secondly, there was a problem with randomizing the participants based on their interest level. Although the instructions and other circumstances in the experiment were deliberately kept as identical as possible, the game-based AR group scored higher on the pretest for interest. Since whole classrooms were attributed to a condition, and group dynamics in classrooms influence motivation (Chang, 2012), it might have happened that less motivated classes accidentally were attributed the interactive 2D condition, and the more motivated classes were attributed the game-based AR condition. This problem was initially solved by using the pretest as a covariate. However, this lowered the power of the study and therefore increased the probability of Type-II errors. A follow-up study with an even distribution of interest on the pretest would be necessary to draw a more certain conclusion on interest differences between game-based AR and interactive 2D.

Thirdly, a more general limitation to the study is that time-on-task differed for both conditions. Although time-on-task was not measured in this experiment, it was clearly noticeable that participants in the game-based AR condition took longer to finish the game than participants in the interactive 2D condition. This influences the comparability of the two conditions. Because the content was deliberately kept as identical as possible, more time in the game-based AR condition went to less relevant tasks like learning the controls of the game or scanning the room for placing the AR environment. These activities could be described as extraneous cognitive load (or irrelevant/distracting activities for learning) and might have influenced learning and interest (Makransky et al., 2020). However, extrinsic load was not measured in the experiment and the game-based AR group still outperformed the interactive 2D group on learning.

Further studies

This study contributes to our knowledge of game-based AR in late primary- and early secondary education, yet it simultaneously generates numerous uncharted inquiries. Firstly, Maries ChemLab has no competitive or cooperative elements. Some scholars state that competitiveness is an important element of game-based learning (e.g., Rigby & Ryan, 2011), since competition and also cooperation are well-known gamification elements. Competition is related to motivation in game-based learning (Vandercruysse et al., 2013). Also, Ke and Grabowski (2007) found that cooperative game-playing was more effective at promoting positive attitudes than individual game-playing. Therefore, future research could investigate whether competition and cooperation mediate learning and motivation in game-based AR environments.

Secondly, the finding that there was no difference in interest might be attributed to the fact that both conditions were interactive. This could imply that in other studies comparing game-based augmented reality (AR) with a non-interactive environment, variations in interest levels might be incorrectly attributed to the level of immersion, when in fact, it is the interactivity that is responsible for the observed differences. Although interactivity and immersion are related concepts, interactivity refers to the degree of engagement and participation allowed or encouraged in an experience (Norman, 2013). It measures the level of interaction between the user and a system, typically through input and output mechanisms. Immersion is more about feeling high presence and being involved sensory and psychologically in a virtual or simulated environment (Cummings & Bailenson, 2016). Thus, immersive environments are interactive, but not all interactive environments are immersive. Another comparative study could further investigate the role of interactivity by manipulating the level of interactivity in both game-based AR and 2D environments. By systematically varying the degree of interactivity and keeping other factors as constant as possible, researchers could determine whether the observed interest differences are primarily driven by immersion or interactivity. This would provide a more comprehensive understanding of the factors influencing user engagement and help guide future design considerations for game-based (AR) applications.

Conclusion and implications

As was already found by Garzón and Acevedo (2019), AR is an excellent way of improving the knowledge for vocational education and university students. This study shows that game-based AR is also beneficial for students in late primary school and early secondary school. Moreover, game-based AR can also be used to convey declarative knowledge more effectively than a 2D version with the same content. As for interest, there was no significant difference between the game-based AR group and the interactive 2D group when pre-interest was used as a covariate. Although technical and methodological limitations should be considered when drawing this conclusion, interest scores were generally high. This implies that children enjoyed playing both versions of the game.

While this study helps understanding game-based AR and challenges current conceptions about game-based AR and immersive learning in general, such as the CAMIL model, the implications of these findings extend beyond academic research. First, these findings can inspire developers of educational material and game developers to create more and better game-based AR applications to improve learning in late primary school and early secondary school. Second, teachers who have the availability of tablets or

smartphones in their classrooms can use these to play AR games in the classroom to effectively improve the knowledge of their students.

Literature:

Akçayır, M. & Akçayır, G. (2017). Advantages and challenges associated with augmented reality for education: A systematic review of the literature. *Educational Research Review*, 20, 1–11. https://doi.org/10.1016/j.edurev.2016.11.002

Alper, A., Oztaş, E. S., Atun, H., Cinar, D., & Moyenga, M. (2021). A systematic literature review towards the research of game-based learning with augmented reality. *International Journal of Technology in Education and Science, 5*(2), 224-244. https://doi.org/10.46328/ijtes.176

Anderson, L. W., Krathwohl, D. R., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., & Wittrock, M. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy.* Longman Publishing.

Arztmann, M., Domínguez Alfaro, J. L., Blattgerste, J., Jeuring, J., & van Puyvelde, P. (2022a). Maries' ChemLab: A mobile augmented reality game to teach basic chemistry to children. 16th European Conference on Game-Based Learning, Lisbon. https://lirias.kuleuven.be/retrieve/679819

Arztmann, M., Hornstra, L., Jeuring, J., & Kester, L. (2022b). Effects of games in STEM education: A meta-analysis on the moderating role of student background characteristics. *Studies in Science Education*, 1–37. https://doi.org/10.1080/03057267.2022.2057732

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford University Press.

Blancha, J. M., Alarcón, R., Arnau, J., Bono, R., & Bendayan, R. (2017). Non-normal data: Is ANOVA still a valid option? *Psicothema, 29*(4), 552–557. https://doi.org/10.7334/psicothema2016.383

Bos, J. (2020). Research Ethics for Students in the Social Sciences. Springer Cham. https://doi.org/10.1007/978-3-030-48415-6

Cai, S., Wang, X., & Ciang, F. (2014). A case study of Augmented Reality simulation system application in a chemistry course. *Computers in Human Behavior, 37*, 31–40. https://doi.org/10.1016/j.chb.2014.04.018

Cai, S., Chiang, F. K., Sun, Y., Lin, C., & Lee, J. J. (2017). Applications of augmented reality-based natural interactive learning in magnetic field instruction. *Interactive Learning Environments, 25*(6), 778–791. https://doi.org/10.1080/10494820.2016.1181094.

Chang, L. Y. H. (2012). Group processes and EFL learners' Motivation: A Study of group dynamics in EFL classrooms. *Tesol Quarterly, 44*(1), 129-154. https://doi.org/10.5054/tq.2010.213780

Cheng, M. T., Lin, Y. W., & She, H. C. (2015). Learning through playing virtual age: Exploring the interactions among student concept learning, gaming performance, in-game behaviors, and the use of in-game characters. *Computers & Education, 86*, 18–29. https://doi.org/10.1016/j.compedu.2015.03.007.

Chen, C., ping, & Wang, C. H. (2015). Employing augmented-reality-embedded instruction to disperse the imparities of individual differences in earth science learning. *Journal of Science Education and Technology, 24*(6), 835–847. https://doi.org/10.1007/s10956-015-9567-3.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin, 122*(1), 155-159. https://www2.psych.ubc.ca/~schaller/528Readings/Cohen1992.pdf

Collins, J. (2006). Education techniques for lifelong learning: Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radio Graphics, 26*(2), 543-551. https://doi.org/10.1148/rg.262055145

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology, 102*(4), 874–888. https://doi.org/10.1037/a0027403

Cummings, J. J., & Bailenson, J. N. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media Psychology, 19*(2), 272–309. https://doi.org/10.1080/ 15213269.2015.1015740.

Deci, E. L. (1992). The relation of interest to the motivation of behavior: A self-determination theory perspective. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 43 - 69). Hillsdale, NJ: Lawrence Erlbaum. Associates. https://www.researchgate.net/profile/Edward-Deci/publication/232512697_The_relation_of_interest_to_the_motivation_of_behavior_A_self-determination_theory_perspective/links/56a642d408aeca0fddcb4c6e/The-relation-of-interest-to-the-motivation-of-behavior-A-self-determination-theory-perspective.pdf?

Dehnad, A., Nasser, H., & Hosseini, A. F. (2014). *A Comparison between Three-and Four-Option Multiple Choice Questions. 98*, 398-403. https://doi.org/10.1016/j.sbspro.2014.03.432

Di Serio, Á., Blancha Ibáñez, M., & Delgado Kloos, C. (2013). Impact of an augmented reality system on students' motivation for a visual art course. *Computers & Education, 68*, 586-596. https://doi.org//10.1016/j.compedu.2012.03.002

Dunlevy, M., Dede, C., & Mitchell, R. (2009). Affordances and limitations of immersive participatory augmented reality simulations for teaching and learning. *Journal of Science Education and Technology*, 18, 7–22. https://doi.org/10.1007/s10956-008-9119-1

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics (5th edition)*. SAGE Publications.

Garzón, J., Pavón, J., & Baldiris, S. (2019). Systematic review and meta-analysis of augmented reality in educational settings. *Virtual Reality*, 23, 447–459. https://doi.org/10.1007/s10055-019-00379-9

Garzón, J. & Acevedo, J. (2019). Meta-analysis of the impact of augmented reality on students' learning gains. *Educational Research Review*, 27, 244–260. https://doi.org/10.1016/j.edurev.2019.04.001

Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischbach, A., & Preckel, F. (2014). "My Questionnaire is Too Long!" The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology, 39*(3), 188–205. https://doi.org/10.1016/j.cedpsych.2014.04.002

Gros, B. (2007). Digital games in education: The design of games-based learning environments. *Journal of Research on Technology in Education, 40*(1), 23–38. https://doi.org/10.1080/15391523.2007.10782494

Deci, E. L & Ryan, R. M. (2000). Self-determination theory and the facilitation of Intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78. https://doi.org/10.1037//0003-066X.55.1.68

Harackiewicz, J. M., Smith, J. L., & Priniski, S. J. (2016). Interest matters: The importance of promoting interest in education. *Policy Insights From the Behavioral and Brain Sciences, 3*(2), 220–227. https://doi.org/10. 1177/2372732216655542.

Henrysson, S. (1963). Correction of item–total correlations in item analysis. *Psychometrika, 28*(2), 211–218. http://dx.doi.org/10.1007/BF02289618

Herpich, F., Ribeiro Jardim, R., Becker Nunes, F., Bierhalz Voss, G., Manzoni Fontoura, L., & Duarte Medina, R. (2014). *XVI Virtual Lab: An immersive tool to assist in the teaching of software engineering*. 118–126. https://doi.org/doi:10.1109/SVR.2014.36.

Hessen, D. & Van Erp, S. (2020). *Reader: Assessment en evaluatie 2020-2021.* Universiteit Utrecht.

HHU. (2020). *G*Power*. Retrieved May 17, 2022, from https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

Hidi, S., & Anderson, V. (1992). Situational interest and its impact on reading and expository writing. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 215–238). Lawrence Erlbaum Associates, Inc.

Höft, L., Bernholt, S., & Blankenburg, J. S. (2017). Knowing more about things you care less sectional analysis of the opposing trend and interplay between conceptual understanding and interest in secondary school chemistry. *Journal of Research in Science Teaching, 56*(2). https://doi.org/10.1002/tea.21475

Huang, W. (2020). *Investigating the novelty effect in virtual reality on STEM learning* [Dissertation, Arizona State University]. https://keep.lib.asu.edu/_flysystem/fedora/c7/224783/huang_asu_0010E_20075.pdf

Huang, T. C., Chen, C. C., & Chou, Y. W. (2016). Animating eco-education: To see, feel, and discover in an augmented reality-based experiential learning environment. *Computers & Education, 96*, 72–82. https://doi.org/10.1016/j.compedu.2016.02.008.

Hung, Y. H., Chen, C. H., & Huang, S. W. (2016). Applying augmented reality to enhance learning: A study of different teaching materials. *Journal of Computer Assisted Learning*, 1–15. https://doi.org/10.1111/jcal.12173.

JASP Team. (2022). *JASP (Version 0.16.3)*. Department of Psychological Methods: University of Amsterdam. https://jasp-stats.org/

Kamarainen, A. M., Metcalf, S., Grotzer, T., Browne, A., Mazzuca, D., Tutwiler, M. S., Dede, C. (2013). EcoMOBILE: Integrating augmented reality and probeware with environmental education field trips. *Computers & Education, 68*, 545–556. https://doi.org/10.1016/j.compedu.2013.02.018.

Karagozlu, D. (2018). Determination of the impact of augmented reality application on the success and problem-solving skills of students. *Quality and Quantity, 52*(5), 2393–2402. https://doi.org/10.1007/s11135-017-0674-5.

Ke, F., & Grabowski, B. (2007). Gameplaying for maths learning: Cooperative or not? British Journal of *Educational Technology, 38,* 249– 259. doi:10.1111/j.1467-8535.2006.00593.x

Koch, M., Luck, K. Von, Schwarzer, J., & Draheim, S. (2018). The novelty effect in large display deployments – Experiences and lessons-Learned for evaluating prototypes. *IEEE Access, 6*(99), 3140–3148. https://doi.org/10.18420/ecscw2018

Liu, T. Y., & Chu, Y. L. (2010). Using ubiquitous games in an English listening and speaking course: Impact on learning outcomes and motivation. *Computers & Education, 55*(2), 630–643. https://doi.org/10.1016/j.compedu.2010.02.023.

Mahle, M. (2011). Effect of interactivity on student's achievement and motivation in distance education. *The Quarterly Review of Distance Education, 12*(3), 207-512. https://web.p.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=9ddc74f2-c38e-456b-812d-e78e0f2ddb16%40redis

Makransky, G., Andreasen, N. K., Baceviciute, S., & Mayer, R. E. (2020). Immersive virtual reality increases liking but not learning with a science simulation and generative learning strategies promote learning in immersive virtual reality. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000473.

Makransky, G., Borre-Gude, S., & Mayer, R. E. (2019a). Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. *Journal of Computer Assisted Learning., 35*(6), 691–707.

Makransky, G., & Lilleholt, L. (2018). A structural equation modeling investigation of the emotional value of immersive virtual reality in education. *Educational Technology Research and Development, 5*, 1–24. https://doi.org/10.1007/s11423-018-9581-2.

Makransky, G. & Petersen, G. B. (2021). The Cognitive affective model of immersive learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review, 33,* 937–958. https://doi.org/10.1007/s10648-020-09586-2

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019b). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction, 60*, 225–236. https://doi.org/10.1016/j.learninstruc.2017.12.007.

Mark, M. M., & Lanz-Watson, A. L. (2011). Experiments and quasi-experiments in field settings. In A. T. Panter & S. K. Sterba (Eds.), Handbook of ethics in quantitative evaluation (pp. 185–209). Routledge. https://www.researchgate.net/profile/Nguyen-Trung-Hiep/post/What-textbook-could-you-suggest-for-quantitative-research-for-graduate-level/attachment/59d62ca279197b807798af40/AS%3A347540928122888%401459871622356/download/Handbook+of+Ethics+in+Quantitative+Methodology.pdf#page=206

Moore, J. W., & Fletcher, P. C. (2012). Sense of agency in health and disease: A review of cue integration approaches. *Consciousness and Cognition, 21*(1), 59–68. https://doi.org/10.1016/j.concog.2011.08.010.

Norman D. (2013). *The design of everyday things: Revised and expanded edition.* Basic Books. https://ebookcentral.proquest.com/lib/uunl/reader.action?docID=1167019

Nuutila, K., Tapola, A., Tuominen, A., Kupiainen, S., Pásztor, A, & Nimivirta, M. (2020). Reciprocal Predictions Between Interest, Self-Efficacy, and Performance During a Task. *Frontiers in Education*, 5. https://doi.org/10.3389/feduc.2020.00036

Qualtrics. (2023). *Qualtrics Provo* (March 2023). https://survey.uu.nl/

Radu, I. (2014). Augmented reality in education: A meta-review and cross-media analysis. *Personal and Ubiquitous Computing, 18*, 1533–1543. https://doi.org/10.1007/s00779-013-0747-y

Razali, N. M. & Was, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J*ournal of Statistical Modeling and Analytics, 2*(1), 21-33. https://www.researchgate.net/publication/267205556_Power_Comparisons_of_Shapiro-Wilk_Kolmogorov-Smirnov_Lilliefors_and_Anderson-Darling_Tests#fullTextFileContent

Rigby, S., & Ryan, R. M. (2011). *Glued to games: how video games draw us in and hold us spellbound.* Praeger. Https;//doi.org/10.5860/CHOICE.49-0099

Rosa, N., Van Bommel, J., Hurst, W., Nijboer, T., Veltkamp, R., & Werkhoven, P. (2019). *Embodying an extra virtual body in augmented reality*. 1138–1139. http://www.cs.uu.nl/groups/MG/multimedia/publications/art/IEEE%20VR2019.pdf

Sheridan, T. B. (1992). Musings on telepresence and virtual presence. *Presence: Teleoperators and Virtual Environments, 1*(1), 120–126. https://doi.org/10.1162/pres.1992.1.1.120.

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology, 64(*2), 489–528. https://doi.org/10.1111/j.1744-6570.2011.01190.x

Tabachnick, B. G., & Fidell, L. S. (2013) *Using multivariate statistics (6th ed.)*. Pearson.

Taber, K. S. (2018). The Use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*, 1273-1296. https://doi.org/10.1007/s11165-016-9602-2

Uttl, B. (2005). Measurement of individual differences: Lessons from memory assessment in research and clinical practice. *Psychological Science, 16*(6), 460–467. https://doi.org/10.1111/j.0956-7976.2005.01557.x

Vandercruysse, S., Vandewaetere, M., Cornillie, F., & Clarebout, G. (2013). Competition and students' perceptions in a game-based language learning environment. Educational Technology Research and Development, 61, 927-950. https://doi.org/10.1007/s11423-013-9314-5

Vansteenkiste, M., Soenens, B., Sierens, E., Luyckx, K., & Lens, W. (2009). Motivational profiles from a self-determination perspective: The Quality of Motivation Matters. *Journal of Educational Psychology, 101*(3), 671–688. https://doi.org/10.1037/a0015083

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research, 34*, 229–243. doi:10.2190/FLHV-K4WA-WPVQ-H0YM

Webster, R. (2016). Declarative knowledge acquisition in immersive virtual learning environments. *Interactive Learning Environments, 24*(6), 1319–1333. https://doi.org/10.1080/10494820.2014.994533

Wouters, P., Van Nimwegen, C., Van Oostendorp, H., & Van der Spek, E. D. (2013). A Meta-Analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology, 105*(2), 249–265. https://doi.org/10.1037/a0031311

Xerte Project. (2022). *Xerte* (3.11). Apereo Community. https://xerte.org.uk/index.php/en/

Appendix A: Interactive 2D condition

Link to interactive 2D condition:

https://xerte.uu.nl/play.php?template_id=2825

Appendix B: Learning questions

Pretest questions:

Q1: How do you measure acidity?

A: With pH-strips

B: With the Fe-schale

C: With UV-indexes

D: With the colorimetry-scale

Q2: Which product contains the most acid?

A: Watermelon

B: Lemon

C: Apple

D: Banana

Q3: What happens when you squeeze lemon juice in ripe blackberry juice

A: Nothing

B: Color changes

C: Becomes warmer

D: Starts smoking

Q4 (removed after pilot): What is the opposite of acidity?

A: Salt

B: Base

C: Sweet

Q5: Which product contains the least acid?

A: Vinegar

B: Detergent

C: Milk

D: Coffee

Q6 (removed after experiment): Blackberries don't become more acidic once they ripen

A: True

B: False

C: Don't know

Q7: Which product contains the most acid?

A: Vinegar

B: Water

C: Detergent

D: Coffee

Q8: What color does a ripe blackberry have (dutch translation not a problem, blackberry = braam; maybe ceiling effect)

A: Black

B: Red

C: Green

D: Orange

Table 4

Reliabilty and score per question in the pretest and posttest

Question	Item-rest correlation		Percentage correct	
	pretest	posttest	pretest	posttest
Q1	.28	.28	41%	84%
Q2	n.a.	.17	100%	88%
Q3	.05	.26	56%	66%
Q5	.12	.01	12%	25%
Q7	.00	.02	85%	72%
Q8	-.02	.02	73%	42%

Table 5

Pilot: Reliabilty and score per question in the pretest and posttest

Question	Item-rest correlation		Percentage correct	
	pretest	posttest	pretest	posttest
Pilot Q1	.11	n.a.	71%	100%
Pilot Q2	.26	-.04	92%	96%
Pilot Q3	.32	.38	79%	81%
Pilot Q4	.31	.11	84%	81%
Pilot Q5	.17	.36	21%	15%
Pilot Q6	.05	.14	84%	69%
Pilot Q7	.37	.41	84%	85%
Pilot Q8	.44	.14	34%	42%

Note: Guttman's λ2 pretest = .58, Guttman's λ2 posttest = .51

Appendix C: Ethical considerations

Informed consent

We have passive consent to collect data. This consent was given by the ethical board of Utrecht University. All participants were minors in this study. Therefore, we also needed to inform parents or legal representatives (Bos, 2020). The schools emailed all of the parents. Another important aspect of working with children is that we needed to inform them in a way that they understand. Thus, we kept the explanation relatively simple.

Effort and voluntary participation

The children participated for a maximum of 30 minutes. They got time for this during class, so they did not have to be compensated for participating in their free time. A concern was be that the children do not choose to participate but are selected by the teacher and the researchers. However, their teacher decided that testing Marie's ChemLab is a meaningful additive to the lessons. Therefore, testing the application was part of their lessons.

Another concern was that children in the AR condition got to play a game, while the children in the textbook condition played a presumably less engaging game. Although they were not tested in the same rooms and the time on task was lower for the textbook condition, I could see why children could be disappointed afterwards. Bos (2020) calls this asymmetrical treatment. The problem of asymmetrical treatment is rooted in the principles of egalitarian justice, which suggests that individuals should have equal access to the benefits, rather than simply being protected from harm (Bos, 2020). Mark and Lenz-Watson (2011) therefore suggested that once the results indicate that the experimental condition had significant benefits, the control condition has the right to get access to it. Therefore, after finishing the digital textbook task, children ware able to play the game if they wanted.

Sensitivity

There was no sensitivity issue regarding the instruments. The questions were not provocative since they were just about acids and bases. Also, learning and motivation results would not be published per individual student. This meant that pressure or fear of failing are no hindering factors for validity.

Storage

The data is not sensitive, because participant's identity and gender are made anonymous and, information about their ethnicity, political opinions, medical history, sexual orientation, religious background, and philosophical beliefs were not tested (Bos, 2020). However, this does not mean that data could be stored everywhere or shared with everyone.

Ensuring the safekeeping of research data is a critical aspect of research ethics, particularly in the current age of hacking and data breaches (Bos, 2020). It is also subject to increasing regulation worldwide. The primary goals of secure storage are to allow for verification of the data and to enable its reuse for secondary analysis (Bos, 2020). Data is saved on the U-drive because it is the safest way to store data. The servers are local, 2-factor authentication is needed to access the data and the servers are managed by an IT-team.

Appendix D: Data cleaning steps and JASP-syntax

Data cleaning:

nr.	Action
1	Scoring the values that are not scored yet. For example the score 2 for the knowledge questions. This was a result of adding answer options based on the pilot, but not coding them. Every 2 is coded as a 0.
2	Combine all different datasets in one excel file for the pretest knowledge questions. Filter for ID - lowest to highest. Repeat for posttest
3	Add the posttest results to the new dataset. This is done by copy-pasting them with their ID's still in the file. This way, you have two ID rows, and you can match these. In cases where there is no posttest or no pretest, leave the row blank. In cases where there is a double ID, choose the first entry. There was one occasion where I saw ID416 two times in the pretest, one ID416 in the posttest, but also an ID415 in the posttest. There is no ID215 in the pretest, so it was impossible to match them. Therefore, I had to exclude both ID416's from the experiment. Moreover, ID557 and ID559 were double, but I was able to trace them back based on timestamps. They are now ID701 and ID702.
4	Add the pre- and post interest questions from Qualtrics for the 2D condition.
5	Add the pre- and post interest questions from a separate Excel file for the 3D condition.

Analysis:

nr.	Action	Code or steps
1	Descriptives age	jaspDescriptives::Descriptives(version = "0.17.1", formula = ~ Age)
2	Distribution age	jaspTTests::TTestIndependentSamples(version = "0.17.1", formula = ~ Age, group = "Condition")
3	Distribution gender	jaspFrequencies::ContingencyTables(version = "0.17.1", formula = Gender ~ Condition_2)

Analysis:

nr.	Action	Code or steps
4	Education level distribution	jaspFrequencies::ContingencyTables(version = "0.17.1", formula = Level ~ Condition, countsExpected = TRUE)
5	Adjusted standardized residuals for education level	Adjusted standardized residuals formula in Excel: (observed - expected) / (ROOT(expected * (1-(row total/population total)) * (1-(column total/population total))))
6	Pretest reliability	[Reliabilty] -> [Unidimensional reliability] -> Add all pretest scores to [Variables] -> Select [Guttman's *λ2*] -> Select [Item-rest correlation].
7	Posttest reliability	[Reliabilty] -> [Unidimensional reliability] -> Add all posttest scores to [Variables] -> Select [Guttman's *λ2*] -> Select [Item-rest correlation].
8	Ceiling check pretest	jaspDescriptives::Descriptives(version = "0.17.1", formula = ~ PRE_LO_1 + PRE_LO_2 + PRE_LO_3 + PRE_LO_8 + PRE_LO_7 + PRE_LO_6 + PRE_LO_5, frequencyTables = TRUE)
9	Ceiling check posttest	jaspDescriptives::Descriptives(version = "0.17.1", formula = ~ POST_LO_1 + POST_LO_2 + POST_LO_3 + POST_LO_5 + POST_LO_6 + POST_LO_7 + POST_LO_8, frequencyTables = TRUE)
10	Make sum score pretest	[Compute column] -> [Define column through R-code] -> [PRE_LO_1 + PRE_LO_2 + PRE_LO_3 + PRE_LO_8 + PRE_LO_7 + PRE_LO_5]
11	Make sum score posttest	[Compute column] -> [Define column through R-code] -> [PRE_LO_1 + PRE_LO_2 + PRE_LO_3 + PRE_LO_8 + PRE_LO_7 + PRE_LO_5]
12	Normality check and descriptive statistics	jaspDescriptives::Descriptives(version = "0.17.1", formula = ~ Int_pre + Int_post + PRETEST + POSTTEST, kurtosis = TRUE, shapiroWilkTest = TRUE, skewness = TRUE, splitBy = "Condition_2")
13	Random distribution for learning	jaspTTests::TTestIndependentSamples(version = "0.17.1", formula = ~ PRETEST, descriptives = TRUE, equalityOfVariancesTest = TRUE, equalityOfVariancesTestType = "levene", group = "Condition_2", mannWhitneyU = TRUE)
14	Random distribution for interest	jaspTTests::TTestIndependentSamples(version = "0.17.1", formula = ~ Int_pre, descriptives = TRUE, equalityOfVariancesTest = TRUE, equalityOfVariancesTestType = "levene", group = "Condition_2", mannWhitneyU = TRUE)

Analysis:

nr.	Action	Code or steps
15	Pre-interest per education level	jaspAnova::Anova(version = "0.17.1", formula = Int_pre ~ Level, contrasts = list(list(contrast = "none", variable = "Level")), descriptives = TRUE, effectSizeEstimates = TRUE, effectSizeEtaSquared = FALSE, effectSizePartialEtaSquared = TRUE, homogeneityTests = TRUE, postHocCorrectionBonferroni = TRUE, postHocCorrectionTukey = FALSE, postHocSignificanceFlag = TRUE, postHocTerms = ~ Level)
16	Quality check	jaspRegression::Correlation(version = "0.17.1", assumptionCheckMultivariateShapiro = TRUE, pairwiseDisplay = TRUE, pearson = FALSE, significanceFlagged = TRUE, spearman = TRUE, variables = list("Int_pre", "Int_post", "PRETEST", "POSTTEST"))
17	RQ1 - learning	jaspAnova::AnovaRepeatedMeasures(version = "0.17.1", betweenModelTerms = ~ Condition, betweenSubjectFactors = "Condition", contrasts = list(list(contrast = "none", variable = "Knowledge"), list(contrast = "none", variable = "Condition"), list(contrast = "none", variable = list("Condition", "Knowledge"))), descriptivePlotErrorBar = TRUE, descriptivePlotErrorBarType = "se", descriptivePlotHorizontalAxis = "Knowledge", descriptivePlotSeparateLines = "Condition", descriptives = TRUE, effectSizeEstimates = TRUE, effectSizeEtaSquared = FALSE, effectSizePartialEtaSquared = TRUE, homogeneityTests = TRUE, marginalMeanTerms = list("Knowledge", "Condition"), repeatedMeasuresCells = list("PRETEST", "POSTTEST"), repeatedMeasuresFactors = list(list(levels = list("Pretest", "Posttest"), name = "Knowledge")), withinModelTerms = list("Knowledge"))
18	RQ2 - interest	jaspAnova::Ancova(version = "0.17.1", formula = Int_post ~ Condition + Int_pre, covariates = "Int_pre", contrasts = list(list(contrast = "none", variable = "Condition")), descriptives = TRUE, effectSizeEstimates = TRUE, effectSizeEtaSquared = FALSE, effectSizePartialEtaSquared = TRUE, homogeneityTests = TRUE)