



ARETE – DELIVERABLE

WP4 – D4.5 ARETE Use Scenarios: Analysis of Educational Technologies Usage in Situ – UPDATE

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 856533

Deliverable number:	D4.5 (update for D4.1)
Due date:	30 April 2022
Nature ¹ :	R
Dissemination Level:	PU
Work Package:	4
Lead Beneficiary:	UDUR
Beneficiaries:	NUID UCD, CLB, WWL, SVU, EUN, CNR, UNW, VICOM, OU

1

Nature:

R = Report, P = Prototype, D = Demonstrator, O = Other

Dissemination level

PU = Public

PP = Restricted to other programme participants (including the Commission Services)

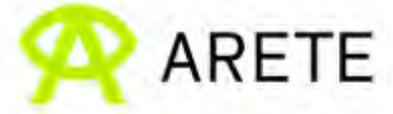
RE = Restricted to a group specified by the consortium (including the Commission Services)

CO = Confidential, only for members of the consortium (including the Commission Services)

Restraint UE = Classified with the classification level "Restraint UE" according to Commission Decision 2001/844 and amendments

Confidential UE = Classified with the mention of the classification level "Confidential UE" according to Commission Decision 2001/844 and amendments

Secret UE = Classified with the mention of the classification level "Secret UE" according to Commission Decision 2001/844 and amendments



Document History

Version	Date	Description
1	03.04.2022	Initial complete draft
2	05.04.2022	Release version for comment
3	15.04.2022	Comments addressed

Disclaimer

The contents of this document are the copyright of the ARETE consortium and shall not be copied in whole, in part, or otherwise reproduced (whether by photographic, reprographic or any other method), and the contents thereof shall not be divulged to any other person or organisation without prior written permission. Only members of the ARETE Consortium, entered the ARETE Consortium Agreement, dated 24.04.2019, and the European Commission can use and disseminate this information.

Content provided and information within this report is the sole responsibility of the ARETE Consortium and does not necessarily represent the views expressed by the European Commission or its services. Whilst this information contained in the documents and webpages of the project is believed to be accurate, the authors and/or any other participant of the ARETE consortium makes no warranty of any kind with regard to this material.



Table of Contents

Executive Summary	3
1. Introduction	5
2. Human-centred Design (HcD) Methods	5
3. Pilot 1: Interaction Design of WordsWorthLearning (WWL) Application	6
3.1 Heuristic Evaluation of the WordsWorthLearning ARETE Read & Spell app	6
3.1.1 Procedure	6
3.1.2 Findings	7
3.2 Focus Groups with Teacher Coordinators	12
3.2.1 Before the Intervention	12
3.2.2 After the Intervention	14
4. Pilot 2: Interaction Design of CleverBooks (CLB) Applications	16
4.1 Heuristic Evaluation	16
4.1.1 Procedure	16
4.1.2 Findings – Geography Workbook and Geography Application	16
4.1.3 Findings – Geography map	23
4.1.4 Findings – Geometry workbook and geometry application	23
4.2 Focus Groups with Teacher Coordinators	26
4.2.1 Before the intervention	26
4.2.2 After the intervention	28
5. Pilot 3: Interaction Design of PBIS Prototypes	30
5.1 Proposal for the PBIS User Interfaces Design	30
5.1.1 General Interface improvement suggestions	30
5.1.2 Menu bar suggested improvements	30
5.1.3 Routine selection improvements	31
5.1.4 Leaderboard design suggestions	32
5.1.5 AR mode text consistency issues	32
5.2 Usability Design Requirements for AR-PBIS Applications	33
6. Pilot 4: MirageXR	33
6.1 Heuristic Evaluation of MirageXR	33
6.1.1 Findings	34
6.1.2 Improvements after heuristic evaluation	41
6.2 User-based Usability Tests of MirageXR	42
6.2.1 Basic features for an example XR lesson	42
6.2.2 Procedure	43
6.2.3 Results	43
6.3 Discussion	46
7. Innovative Approaches to Evaluating eXtended Reality (XR)	51
8. Conclusion	52



Executive Summary

This deliverable D4.5 reports the evaluation methods and results that WP4 has implemented and obtained since the submission of D4.2 in M18. In close collaborations with WP3, WP5 and WP6, the HCI team of WP4 endeavoured to provide timely feedback to the respective teams for improving the interaction quality of the design artefacts generated for the four Pilots. Utilising the established human-centred design methods, we have identified usability problems (UPs). For each UP, we assigned an importance level (low, medium, high) for it to be fixed and proposed recommended modifications for the development team to consider.

The table below presents an overview of the evaluation studies undertaken. The design artefacts for each Pilot were at different development stages when under evaluation. The number of usability problems (UPs) varied with the complexity of the software package involved. Specifically, the Pilot 2 package comprised five components - two interactive apps, two workbooks and one map – accounting for a higher number of UPs, albeit the majority was of low importance. Encouragingly, the revised versions of the design artefacts evaluated were found to be more usable and pleasurable to use.

Pilot: Design Artefacts	Methods	Main Results
1: Read & Spell (beta release for pilots)	<ul style="list-style-type: none"> ▪ Heuristic Evaluation ▪ Focus Groups with teacher coordinators and pilot teachers 	<ul style="list-style-type: none"> ▪ 28%-low, 22%-medium, 50%-high; all 18 UPs fixed. ▪ Overall teachers were contented; the app was engaging for students; the device freezing issue occurred randomly; some tasks with stringent criteria were challenging; requested improving device compatibility and child-friendliness
2: Geography app with workbook and map + Geometry app with workbook (beta release for pilots)	<ul style="list-style-type: none"> ▪ Heuristic Evaluation ▪ Focus Groups with teacher coordinators and pilot teachers 	<p>1.43%-low, 20%-medium, 37%-high; majority of 46 UPs fixed.</p> <p>2. Overall teachers were satisfied; the app was enjoyable for students; experienced translation and usability issues; requested content customisability and simplification of login process.</p>
3: AR-PBIS application (medium-fidelity prototype)	Heuristic Evaluation	Five areas: general UI; menu bar; routine selection, leaderboard; AR text mode (under re-development)
4. MirageXR v.1.5 – v.1.8 (high-fidelity prototype)	<ul style="list-style-type: none"> ▪ Heuristic Evaluation ▪ User-based usability tests with 10 proxy participants 	<ul style="list-style-type: none"> ▪ 6%-low, 42%-medium, 42%-high; majority of 26 UPs fixed ▪ Strong potential of MirageXR as educational tool recognised; Unresponsiveness as most serious issue

Furthermore, WP4 explored the innovative approaches to evaluating eXtended Reality (XR) applications with the wider research community. A workshop was held where emergent XR



devices (e.g., smart glasses) and innovative methods (e.g., multisensory fusion) were explored.



1. Introduction

This is the third deliverable of WP4, preceded by D4.1 (M9) and D4.2 (M18). Significant changes and progresses have been made since the release of D4.2 from the scientific as well as societal perspective. Particularly relevant is the recent removal of the restrictions imposed by the pandemic, allowing face-to-face empirical studies to take place. For instance, in Feb 2022 the WP4 HCI team were able to conduct user-based usability tests in our lab, which are documented in detail in Section 6.2.

Furthermore, the three Pilots and additional Pilot 4 involved a range of design artefacts and methodological protocols. The HCI team has collaborated closely with the respective teams in WP3, WP5 and WP6 to evaluate these project outputs on an ongoing basis, identifying what, why and how to improve their quality. Specifically, we applied different human-centred methods (Section 2) to analyse systematically the WordsWorthLearning's Read & Spell application for Pilot 1 (Section 3), the Cleverbook's Geography workbook, application and map as well as Geometry workbook and application (Section 4), the PBIS applications (Section 5) and the MirageXR application (Section 6). Usability problems and recommended modifications of different importance levels have been identified and fed back to the development team for consideration. Depending on constraints (e.g., time, effort), some recommendations have been implemented and proved effective (e.g., Table 2 in Section 3.1.2), albeit at different points of time, whereas some have been archived for potential development in the future (e.g., the potential use of artificial intelligence to create teachers' avatars that behave and appear human-like, discussed further in Sections 6.2.3-4), if resources are available.

2. Human-centred Design (HcD) Methods

In this section, we summarise the key Human-centred Design (HcD) methods which the HCI team have applied to evaluate the design artefacts of the three Pilots (1, 2 and 3) and to prepare Pilot 4. The key concepts underpinning these methods, including Usability, User Experience, Formative Evaluation and Summative Evaluation, are documented in D4.1 (M9).

- **Survey:** Survey methods, including *questionnaire* and *interview*, are widely used in the field of HCI to gather participants' subjective data (Lazar et al. 2017). In WP4, we have employed these two methods in the context of requirements analysis, usability tests and reflective workshops. Specifically, standardized questionnaires such as SUS (System Usability Scale; Brooke, 1996) and HARUS (Handheld Augmented Reality Usability Scale; Santos et al., 2015) have been administered. Semi-structured interviews are conducted to gain additional and deeper insights.
- **Focus Group:** Focus Groups are semi-structured group interviews facilitating the discussion of topics that are of interest. Typically, a focus group is moderated by two researchers, with one presenting questions to the group and managing the group dynamics while the other one is observing and taking notes. Participants are encouraged to share their feelings and thoughts by prompts. For WP4, focus groups are performed with teachers to collect their input and feedback on scenarios and functionality options or interface design alternatives as well as collecting their



impressions and feedback after using the different ARETE software artefacts and materials.

- **Heuristic Evaluation (HE):** Heuristic Evaluation is a usability inspection method where feedback on design artefacts is generated by HCI specialists without involving end-users. Specifically, prototypes are inspected for compliance with or violation of usability heuristics (cf. the ten widely used ones proposed by Jakob Nielsen, 1994 and proved applicable for today's technologies²). This method can be applied throughout the software development lifecycle, including low-, medium- and high-fidelity prototypes as well as alpha/beta releases. The main result of HE is a list of usability problems (UPs) with their importance being classified as high, medium, or low (Section 3.1). To support developers, this list typically includes recommended modifications that can be implemented to address and resolve the usability problems identified.
- **Think-aloud (TA):** TA is a common HCI method for usability testing. There are two major types: concurrent (CTA) and retrospective (RTA). In case of CTA participants are asked to verbalize their thoughts when interacting with a system to complete given tasks whereas in case of RTA participants are asked to first perform the tasks in silence and then make a verbal report on the interaction, typically immediately but reporting can also be delayed by hours or days. The strength of CTA is that real-time verbalizations reflect truly cognitive processes underlying the actual interaction with the system under evaluation. RTA is recognized for providing deeper insights into the reasons behind behavioural and emotional responses to the interaction. We employed CTA for our usability tests (Section 6).

3. Pilot 1: Interaction Design of WordsWorthLearning (WWL) Application

This section presents the collaborations that WP4 did with WWL, specifically the evaluation of the digital artefact ARETE Read & Spell application that was used for Pilot 1. We first present the heuristic evaluation results of the ARETE Read & Spell app that discusses usability issues we identified in the development process and improvements that were implemented. Then, we summarise the discussion and feedback from the teacher coordinators in the focus group that we conducted before and after interventions.

3.1 Heuristic Evaluation of the WordsWorthLearning ARETE Read & Spell app

3.1.1 Procedure

A team of three HCI specialists went through the process of walkthrough analysis for the ARETE Read & Spell application to check for any usability issues or software bugs. Several tasks and sequences were tried out to ensure the correct performance of the ARETE Read & Spell app under different circumstances.

The main evaluation session lasted 2.5 hours during which the specialists assumed the roles of teacher and student while being aware of the fact these end-users typically possess a wide

² <https://www.nngroup.com/articles/ten-usability-heuristics/>



range of computer literacy. Some key usability and user experience concepts, including the aesthetic and affective factors, were taken into account for this evaluation.

The device for running the application was an Android phone (Oneplus8T Android 11.0.12.12 Model KB2003 12GB RAM) with a screen resolution of 2400x1800px.

After the heuristic evaluation session, the usability observations made were circulated around the team so that each member could independently assign an importance level (H - high / M - medium / L - low) for fixing each issue:

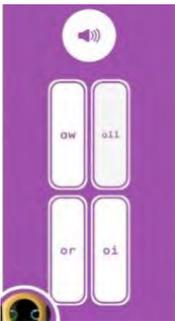
- Low importance (L) rating is given for issues, which would be noticed by end-users and might affect their overall sense of the quality of the interface, but such issues would not hinder them significantly in achieving their objectives.
- Medium importance (M) rating is given for issues, which would be noticed by end-users and may confuse, delay or distract them briefly and temporarily.
- High importance (H) rating is given for issues, which would be an obstacle for end-users, either preventing them from achieving their goals or causing significant delay, disruption, confusion or annoyance.

Finally, discrepancies in importance scores were discussed and a consensus was reached for each usability observation.

3.1.2 Findings

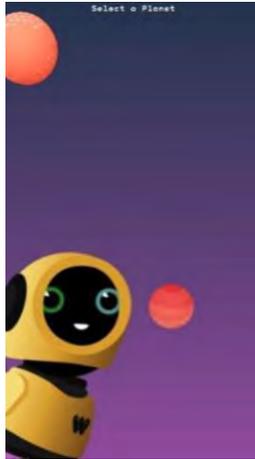
Recommended modifications: Based on our observations and discussions regarding the usability of the ARETE Read & Spell application, we proposed a list of modifications (Table 1).

Table 1: Findings of the heuristic evaluation of ARETE Read & Spell application

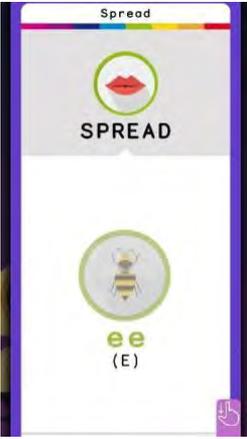
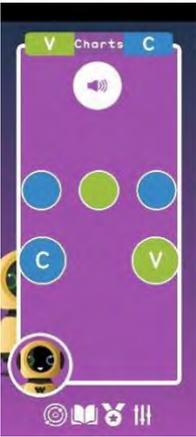
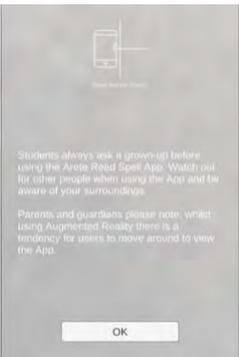
ID	Usability Observation	Recommended Modification	I*	Developer's Responses
WWL _1	Some questions do not have a correct answer, which prevents students from progressing entirely. Therefore, Students cannot progress further and lost interest 	Check all the quizzes to guarantee that they have a correct answer to progress. Also, consider adding a skip button.	H	Fixed.
WWL _2	The planet selection interface should show whether the planet is locked or not. Students could try to press the lock planet and cause frustration.	Add a locked icon for the locked planets.	H	Set planets as Monochrome.



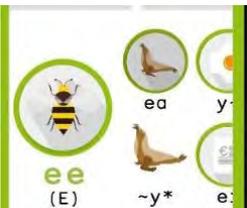
ID	Usability Observation	Recommended Modification	I*	Developer's Responses
WWL _3	The planet selection interface also indicated the “finished” planet. Students may forget which planet is already learnt.	Add a clear icon for the cleared planets.	H	Never happened in WWL 7 Levels.
WWL _4	Tiptop should stay on the current planet, not resetting position to the start planet after the lesson. Students may forget which planet is already learnt.	Tiptop position on the current planet.	H	Never happened in WWL 7 Levels.
WWL _5	The selected option is not clear. At a glance, students can mistake the selected option from unselected.	Make the selected option more visible.	H	Change each selection button to colour grey.
WWL _6	In some quizzes, Tiptop should be more explicit that the students can select multiple choices. Students could be frustrated by the quiz without explanation.	Add explicit explanation that multiple selections are expected.	H	This would be too much info on screen & the TT audio lesson explains the task.
WWL _7	Learning vowels take a long time because it requires AR calibration for every vowel. Students might get annoyed by the AR calibration process and lost interest	Provide alternative options to AR or provide a better way for AR calibration.	H	See below for the position of this Info "Camera needs background details to calibrate a position for the AR object"



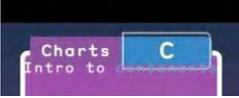


ID	Usability Observation	Recommended Modification	I*	Developer's Responses
WWL_8	<p>Cannot retry vowels after finished learning, need to start an entire lesson to retry. Students can accidentally close AR mode and cannot retry, there is a chance that the device cannot locate a flat surface and retry is needed.</p> 	<p>Allow the students to retry the learnt vowels.</p>	H	Fixed
WWL_9	<p>The automatic alignment of drag and drop in this quiz make it arduous to place the third element, the user has to place the third element behind the second element in a tiny grey area gap. Students cannot place the third elements, which can cause frustration.</p> 	<p>Remove automatic alignment or provide a larger placeable gap.</p>	H	Reduced size of counters, works OK for Tablets.
WWL_10	<p>AR mode warning should display at the start of the program, not every time AR mode start. Students may be annoyed by the AR warning and lost interest.</p> 	<p>Provide a warning once when the application starts.</p>	M	<p>This is a stipulation that is necessary for Google Apps. Added calibration info: "Camera needs background details to calibrate a position for the AR object"</p>



ID	Usability Observation	Recommended Modification	I*	Developer's Responses
WWL_11	<p>Consider adding a planet name for each planet. Students may not remember planet name, which prevents them from discussing with each other</p> 	<p>Add planet name in the planet selection screen.</p>	M	<p>Add planet names on top of each planet.</p>
WWL_12	<p>The reward screen's objective is not clear. Students may not understand the purpose of the reward screen and lost interest.</p> 	<p>Add reward screen explanation.</p>	M	<p>Added Wording for start prize Reveals a msg box & press OK "Well Done! You've earned a piece for your galactic jigsaw"</p>
WWL_13	<p>The replay button should be removed after the quiz start. Students may press replay instruction without app response.</p> 	<p>Remove the replay instruction button, or allow the instruction to be replayed during the quiz.</p>	M	Fixed
WWL_14	<p>Tiptop should face the user when placed in the AR mode. Students cannot see the Tiptop face without repositioning.</p> 	<p>Make sure that Tiptop faces the devices when placed.</p>	L	Fixed
WWL_15	<p>Video mode in vowel library shows a couple of frames of the last played video. Students may be confused by</p> 	<p>Remove a start of the video that show incorrect information.</p>	L	Fixed



ID	Usability Observation	Recommended Modification	I*	Developer's Responses
	incorrect information displayed.			
WWL_16	<p>The replay button after each lesson did not reset the “highlight” effect. The effect on “C” keep playing after replaying the lesson. Students may be confused when replaying content.</p> 	Make sure the interfaces are properly reset.	L	Fixed
WWL_17	<p>The replay button after each lesson misaligns the contents. Students may be confused when replaying content.</p> 	Make sure the interfaces are properly reset.	L	Fixed
WWL_18	<p>Students can only go to the next lesson or replay after the congratulation screen. Therefore, students cannot pause to see their progress</p> 	After the congratulation screen considers adding a button for going back to the module selection screen to help students get a better sense of progress and an option to pause learning.	L	Fixed

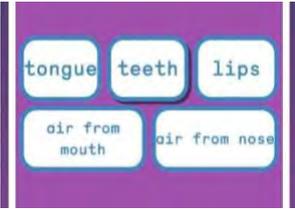
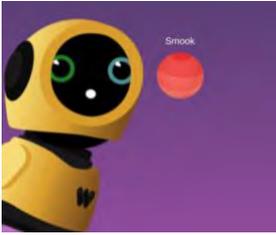
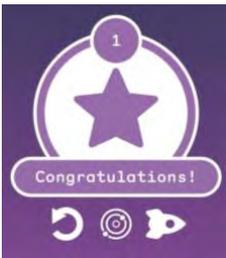
Positive observations: Tiptop is a colourful and child-friendly virtual companion, which would definitely attract children’s attention and motivate them to go through the extensive lesson within ARETE Read & Spell application. The voice acting is also fun, clear, pleasant to listen to.



The application design enables the children to progress through lessons step-by-step, which is essential for developing correct pronunciation. Furthermore, the rewarding system is designed to provide students with a sense of achievement after finishing each lesson. Finally, the amount of lessons in this application is comprehensive, and we envision that students could use the ARETE Read & Spell application as a reference even after the students completed all the lessons in the application.

Improvements after heuristic evaluation: During the heuristic evaluation, 5 low-, 4 medium-, and 9 high-importance issues were found while using the ARETE Read & Spell application. Most of these issues had already been addressed by the WWL developers, but not released yet. After the update, the application was significantly improved in terms of usability, consistency, and stability. Some of the visible improvements are shown in following Table 2.

Table 2. Examples of the usability improvement of ARETE Read & Spell application.

Before usability improvement	After usability improvement	Improvement description
		The selected choices in the exercise interface are more perceptible.
		Added planet name in the planet selection screen to help students reference their lesson
		Added a (middle) button for going back to the module selection screen to help students get a better sense of progress and an option to pause learning.

3.2 Focus Groups with Teacher Coordinators

3.2.1 Before the Intervention

For the Pilot 1 Teacher Coordinators workshop, we used a combination of a focus group and an online feedback platform to facilitate the discussion. Padlet (<https://padlet.com>), a password-protected online feedback platform, was used to enable the teacher coordinators to note down their observations and feedback via notes, audio clips, and photos, during the workshop. Instruction documents to Padlet were sent to the participating teachers prior to the workshop and introduction slides were used to remind the teacher coordinators of the Padlet during the workshop. Screen capture of the Padlet containing teacher coordinators'



feedback for Pilot 1 is shown in Figure 1. The teacher coordinators' feedback on the Padlet was used to facilitate discussion in a focus group on the second day of the workshop.

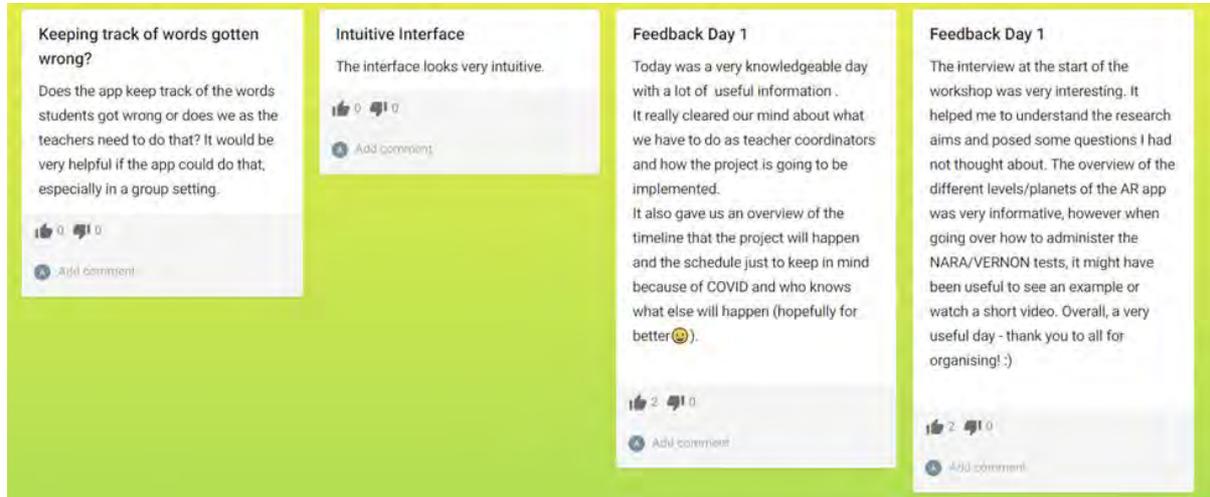


Figure 1. Screenshot of teacher coordinator's feedback on Padlet for Pilot 1 (pre-intervention).

During the focus group, we asked general questions to avoid bias, such as, “What are the application requirements?”, “What do you like most?”, and “What do you like least?”. The focus group ran for 1 hour with the two teacher coordinators whose responses were generally positive. Both of them praised that the user interfaces looked intuitive, accessible, and engaging. One of the teacher coordinators said that the application would be “picked up very quickly by kids”. This view was echoed by the other teacher coordinator, who complimented the application space theme: “most kids like space and this (design) certainly would attract their attention”. Both teachers remarked that the virtual assistant “Tiptop” looked child-friendly, and one of them even suggested that Tiptop could be incorporated into creative writing exercises for other learning topics. Concerning performance, one teacher coordinator commented “the application loads very quickly, especially for the application of this size”. Furthermore, both teachers praised the design of the lesson. They liked the step-by-step learning progression from one planet to the next, because they thought that the students could see their own progress and become motivated. While one teacher liked the gamification elements, they suggested that those could be improved by incorporating more AR elements into the game.

These findings may be somewhat limited as both the teachers noted that they had not spent enough time on the application. Nevertheless, some concerns were expressed about how students might go through the application, especially the exercise. One teacher commented that the exercise passing grade, which required a perfect score, might be too restrictive and could cause frustration in students. Both suggested more flexible alternatives such as an option for teachers to manually move students to the next level, reduce a passing grade after each attempt, or include hints from Tiptop to encourage students to finish the exercise or lesson.

Aside from their concerns, teacher coordinators also requested additional features to help their teaching. Both teacher coordinators requested a dashboard to monitor students'



progression and performance. The dashboard should be simple and accessible, while also displaying essential information such as individual student progress, last login time, learning duration, and record of student’s mistakes, as teacher coordinators would like to use this information to detect possible issues in their class. One teacher coordinator also requested manuals, supporting materials, and teaching templates for teachers.

Table 3: Pilot1 After Intervention HARUS Questionnaire Results (T1-4: teachers’ responses)

Handheld Augmented Reality Usability Scale (1: Strongly Disagree – 7: Strongly Agree)							
#	Statements	T1	T2	T3	T4	Mean	SD
Q1	I think that interacting with this application requires a lot of mental effort.	5	7	5	4	5.25	1.09
Q2	I thought the amount of information displayed on screen was appropriate.	5	4	5	3	4.25	0.83
Q3	I thought that the information displayed on screen was difficult to read.	2	2	4	4	3	1.00
Q4	I felt that the information display was responding fast enough.	2	1	3	4	2.5	1.12
Q5	I thought that the information displayed on screen was confusing.	2	6	2	3	3.25	1.64
Q6	I thought the words and symbols on screen were easy to read.	6	7	7	5	6.25	0.83
Q7	I felt that the display was flickering too much.	1	2	2	3	2	0.71
Q8	I thought that the information displayed on screen was consistent.	5	4	6	4	4.75	0.83
Comprehension Score (%)		66.67	47.92	66.67	54.17	58.85	8.12
Q9	I think that interacting with this application requires a lot of body muscle effort.	1	1	1	2	1.25	0.43
Q10	I felt that using the application was comfortable for my arms and hands.	4	4	7	6	5.25	1.30
Q11	I found the device difficult to hold while operating the application.	2	2	7	3	3.5	2.06
Q12	I found it easy to input information through the application.	5	6	6	4	5.25	0.83
Q13	I felt that my arm or hand became tired after using the application.	2	2	1	2	1.75	0.43
Q14	I think the application is easy to control.	3	6	6	4	4.75	1.30
Q15	I felt that I was losing grip and dropping the device at some point.	2	1	1	2	1.5	0.50
Q16	I think the operation of this application is simple and uncomplicated.	6	6	5	4	5.25	0.83
Manipulation Score (%)		72.92	83.33	79.17	68.75	76.04	5.61

3.2.2 After the Intervention

Similar to the workshop we had before the intervention, we held a one-hour focus group, which was attended by four teachers; two of them had also attended the workshop prior to the intervention. The focus group was structured as three parts. First, the teachers were requested to share their experience with the training materials they had been given before the intervention. The opener question to stimulate the discussion was: "Was the training material adequate to prepare you for the intervention?" Second, they were asked to complete the questionnaire HARUS (Section 2) to gauge the perceived usability of the application with the two scores: comprehension and manipulation. The former represents how well the user understands the information offered by the AR system (Table 3: Q1–8), whereas the latter



indicates the ease of handling the AR device as the user performs the task (Table 3: Q9–16). Third, the teachers were asked to discuss "*What do you like most/least?*" and "*Which issue would you like to see improved the most?*".

Regarding the training materials before the intervention, all the teachers agreed that the training materials were helpful as a starting point but not adequate for using the application in the classroom. As one teacher stated, "*The training guide only gives you an overview of the app and it would be beneficial if there was some more support, such as handbooks or answer keys, to support the teachers because the games (inside the app) were very difficult, even I found it challenging.*" Another teacher said that the lessons inside the application were highly specialised, alluding that the lessons are "borderline speech therapy" and suggesting that most teachers would not continue with the application without adequate support. The teachers agreed that the training materials should include the expected learning outcomes of the game and more detailed information about the solutions of the exercises and how the application works.

Regarding the application usability, Table 3 shows the above-average HARUS scores for both comprehensibility ($M = 58.85$, $SD = 8.12$) and manipulability ($M = 76.04$, $SD = 16.97$). A closer look at the HARUS results indicated that the teachers gave notably low ratings to Q1 and Q4, which were connected to mental effort and response time. When asked to elaborate about the mental effort (Q1), all the teachers indicated that the problems were in the application's exercise. As one teacher stated, "*the interface is not necessarily complicated but requires the students to store a lot of information in their heads during the exercise.*" Another teacher concurred: "*there are parts that are more intensive than others; the first planet exercise is quite challenging, but it gets easier after the second planet.*" The teachers agreed that the perfect passing grade was the source of the issue, as they said: "*ten questions exercises are quite hard, students get 9 out of 10, but they have to go back and do the test again*". Then, they added: "*sometimes they redo the exercise, and they get a lower score than the first time, because they also don't know which questions they got wrong.*" One of the teachers said that the exercises were rather demoralising for the students since they focused on the wrong answer rather than the correct answer, and it turned into a memory game as they tried to recall the questions that they got wrong. She also provided an example "*students are confused between the sounds that are represented by lowercase and uppercase (th and TH), and they do not have any hint at all that they sound different and they have to rely on their memory during the exercise.*" To solve this issue, teachers suggested that the passing grade should be relaxed and the application should give constructive feedback rather than just a score notification since the students would quickly lose interest. Furthermore, teachers pointed out that the application relied on students to check their answers with the knowledge base inside the app, which 8–9-year-old students would not do.

Regarding the response time (Q4), the teachers attributed the low score to the application's performance and compatibility issues. One teacher commented: "*I am reluctant to use the AR button because it keeps freezing and there is no way to get out of the AR without restarting the application from the beginning, which is quite frustrating and takes a lot of time to get through.*" Another teacher also pointed out that the freezing problems were not restricted to AR mode. The teachers also found compatibility issues across different devices; as one



teacher said, *"we have ten iPads, and only three are compatible, and it seems to work on certain software updates and models."* However, another teacher claimed that even devices with the same model and software version might freeze at random, and that some students had the problem while others did not. Aside from the freezing problems, one teacher suggested a change in the application's language usage. She said that the symbol "&" confused students and proposed that the application should use the word "and" to make the application more child-friendly. Another teacher found that some students could gain access to all of the materials for the teacher's account; nevertheless, the teacher noted that this problem had been resolved since.

Despite the technical difficulties, teachers reported getting positive feedback from students, for example, students said that *"I like the games and there are a lot of levels I can learn in the game"* and *"The app is cool and marvelous."* One teacher claimed that the app provided the students with confidence during their spelling work, and that the students felt better about themselves after the session since they felt like they had accomplished something special. Another teacher agreed, saying, *"When it's running well, it's pretty straight forward"*. As a result, when we asked the teachers for suggestions for improvement, they unanimously agreed that resolving the technical challenges (freezing), and improvements in training materials and in-app exercises (as discussed above) should be the top priority. Furthermore, it was suggested that students should have the opportunity to consolidate their learning and revisit these teachings in different contexts, rather than going through these exercises and then never looking at it again. Additionally, another teacher requested that AR activities include more interactivity, stating that the existing AR activities are "nice novelty" but rather limited.

4. Pilot 2: Interaction Design of CleverBooks (CLB) Applications

4.1 Heuristic Evaluation

4.1.1 Procedure

Two HCI specialists of the ARETE team performed the heuristic evaluation study, which was aimed to evaluate the usability of the CLB artefacts used for Pilot 2: The ARETE Geography app, together with the CLB Geography workbook and the world map, as well as the ARETE Geometry app, together with the CLB Geometry workbook. Several tasks and sequences were tried out to ensure the correct performance of the ARETE Geometry and ARETE Geography applications under different circumstances. The evaluation session lasted about 2.5 hours during which the procedure similar to Section 3.1.1 was carried out.

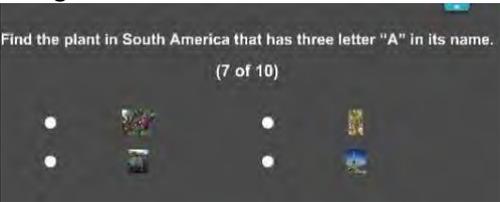
The devices for running the applications were two Android phones (Oneplus8T Android 11.0.12.12 Model KB2003 12GB RAM and Huawei P20 with 4GB RAM running Android 10). The screen resolution was 2400x1800 and 2244x1080, respectively.

4.1.2 Findings – Geography Workbook and Geography Application

Based on our observations and discussions regarding the usability of the ARETE Geography application <https://play.google.com/store/apps/details?id=eu.cleverbooks.arete.geography>, we recommended the following modifications (Table 4 where GG = GeoGraphy; I* = Importance; see Section 3.1.1)



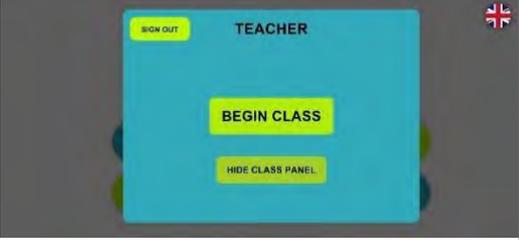
Table 4. Findings of heuristic evaluation of CLB Geography workbook and application

ID	Usability Observation	Recommended Modification	I*
CLB_GG01	<p>Any user can “Sign in as teacher”, thus students could misuse or be confused by this option.</p> 	<p>Create a separate teacher app, so that students don't get the option to sign in as a teacher.</p> <p>Or based on the password used to access the app, determine the user to be a student or teacher, e.g. all teacher passwords could have a “t_” at the start of their actual password.</p>	H
CLB_GG02	<p>There are some grammar and layout issues in the workbook, for example on the bottom of page 2: “You can this app download free from your mobile apps store (Android or Apple)”</p>	<p>Especially for primary school students correct grammar and spelling is very important, thus those issues need to be fixed.</p>	H
CLB_GG03	<p>If the teacher “assigns a quiz” to the students, on the student screen the top of the content is cut off.</p> 	<p>Make sure all content is always visible regardless of aspect ratio.</p>	H
CLB_GG04	<p>There are typos in the quiz.</p> 	<p>Especially for primary school students correct grammar and spelling is very important, thus the app should be checked and errors fixed.</p>	H
CLB_GG05	<p>The images of the plants in the “Plants” quiz are tiny, so it is very hard to recognise them.</p> 	<p>Make the images bigger.</p>	H

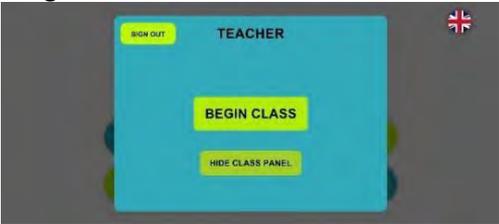


ID	Usability Observation	Recommended Modification	I*
CLB_ GG06	<p>If the teacher already started a class and a student tries to join, then a pop-up with a Russian error message appears. Even though the students use the English language setting for the app.</p> 	<p>Fix this bug and ensure that the error message matches the chosen language.</p>	H
CLB_ GG07	<p>Multiple 3D AR models are not easily recognisable and can therefore not be matched easily to the workbook content (e.g., the Sidney Opera House or plant models).</p>	<p>When selecting a model on the AR continent, show a larger, rotatable model superimposed on the screen.</p>	H
CLB_ GG08	<p>The “Antarctica” marker is not recognised.</p>	<p>Add the Antarctica marker to the app.</p>	H
CLB_ GG09	<p>The animals on page 11 do not match the animals in the app.</p>	<p>Add more animals to the app.</p>	H
CLB_ GG10	<p>At the end of the flags game for Australia, a Russian message appears.</p> 	<p>Translate this message or ensure that the message matches the chosen language.</p>	H
CLB_ GG11	<p>There are no animals in the “Animal” game.</p> 	<p>Fix this possible bug.</p>	H

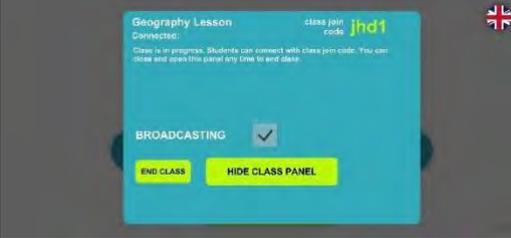
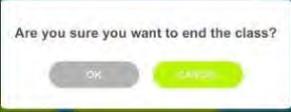


ID	Usability Observation	Recommended Modification	I*
CLB_ GG12	If a student accidentally clicks the home button while in a game they just see a blank screen and there seems to be no way back to the game for them.	Offer the student a way to return to the game.	H
CLB_ GG13	<p>There are no heritages in the “Heritage” game.</p> 	Fix this possible bug.	H
CLB_ GG14	<p>It is unclear what the “class panel” is, as the pop-up is labelled “TEACHER”.</p> 	Change label to something like “Class Panel (Teacher)”.	M
CLB_ GG15	On the student waiting screen it is unclear what is currently happening.	Add a message “waiting for your teacher to start the broadcast” and an “alive” indicator (e.g., spinning circle).	M
CLB_ GG16	The functionality of the “hide class panel” button in the student waiting screen is unclear. It should only appear once the teacher has started the class and the panel can actually be hidden.	Remove the “hide class panel” button from the student waiting screen.	M
CLB_ GG17	After selecting a category under “continents” (e.g. Animals or Water Animals) the teacher cannot go back, they have to use the home button and then select “continents” again to get back to the selection screen, which is an unnecessary step.	Add a back button.	M



ID	Usability Observation	Recommended Modification	I*
CLB_GG18	<p>The labels in the “Start Multiuser Activity” screen are confusing, “Assign to class” needs to be selected to start the activity, whereas “Start” goes “back” to the previous screen.</p> 	Label the Start button with “back”.	M
CLB_GG19	It is hard to focus on the “Europe” part of the “Eurasia” marker.	Separate the two markers.	M
CLB_GG20	<p>The button to show the class panel is a bit tiny.</p> 	Make it a bit bigger.	L
CLB_GG21	The “Cancel” button in the “Start Multiuser Activity” screen seems to do the same as the “Start” button.	Make sure they are both needed and do different things or remove one of them.	L
CLB_GG22	<p>It is unexpected that the “hide class panel” is a button on the same level as “begin class”.</p> 	Put “hide class panel” at the top right-hand corner and maybe change to an “X”.	L



ID	Usability Observation	Recommended Modification	I*
CLB_ GG23	<p>There is a lot of white space in the “Join Class” screen for students.</p> 	<p>Put the interface elements in the centre of the pop-up.</p>	L
CLB_ GG24	<p>Teachers cannot untick the “broadcasting” tickbox.</p> 	<p>If this is supposed to be an indicator that the teacher is broadcasting, not an interface element to change the broadcasting status, it should be visualised differently (e.g., without the box, just a tickmark).</p>	L
CLB_ GG25	<p>Class panel pop-up has a lot of white space, which could be used better.</p> 	<p>Class join code should be displayed bigger and more prominently in the centre of the screen. The number of connected students could be shown more prominently.</p>	L
CLB_ GG26	<p>Help text is confusing “You can close and open this panel any time to end class.”</p> 	<p>Should read something like “You can close and open this panel any time to access the information about this class and the button to end class.”</p>	L
CLB_ GG27	<p>Labels in the “end class confirmation pop-up” could be misleading.</p> 	<p>Instead of “OK” and “Cancel” the buttons should be labelled “Yes” and “No” for better clarity.</p>	L



ID	Usability Observation	Recommended Modification	I*
CLB_ GG28	<p>To avoid accidentally or unintentionally clicking the “end class” button it should have a different colour.</p> 	<p>Change the colour of the “end class” button.</p>	L
CLB_ GG29	<p>The language list seems to be in a random order, which can make it hard to find the language you are looking for.</p>	<p>Order the languages in alphabetical order.</p>	L
CLB_ GG30	<p>Students might not speak English, the flags help to identify the language to select, but might not be enough.</p> 	<p>The language should not be written in English in the language list, but in the native language for each entry (e.g. “español” instead of “spanish”).</p>	L
CLB_ GG31	<p>The language used in the workbook might not be entirely suitable for primary school students, e.g., “peculiarities”.</p>	<p>Make sure that the workbooks only use age-appropriate words.</p>	L
CLB_ GG32	<p>The “Choose continent” screen for the “Flags” game is inconsistent (grey) compared to the screen for the other two games (blue and white).</p> 	<p>Make sure the app UI is consistent.</p>	L



4.1.3 Findings – Geography map

Based on our observations and discussions regarding the usability of the ARETE Geography application (<https://play.google.com/store/apps/details?id=eu.cleverbooks.arete.geography>) and Geography map, we recommended the following modifications (Table 5).

Table 5. Findings of heuristic evaluation of CLB geography map

ID	Usability Observation	Recommended Modification	I*
CLB_GG33	<p>If you are in the “Earth Layer” view and lose the map marker from focus/view and put it back into view, the slider interface element is gone. If you click on the “Earth Layer” icon again, to get it back, the earth stays in its open state, but the slider is to the left (in the “Earth closed” position), so it is not possible to close the earth using the slider. Moving it has unexpected consequences.</p> 	Reset the earth to “closed” when the slider is being reset.	H
CLB_GG34	When in “Solar System mode” the speaker becomes yellow when you press it, it is unclear what that means.	If clicking the speaker mutes it, use the same muted speaker symbol as on other screens.	L

4.1.4 Findings – Geometry workbook and geometry application

Based on our observations and discussions regarding the usability of the ARETE Geometry application (<https://play.google.com/store/apps/details?id=eu.cleverbooks.arete.geometry>), we recommended the following modifications (Table 6 where GM = GeoMetry; I* = Importance level)

Table 6. Findings of heuristic evaluation of CLB geometry application

ID	Usability Observation	Recommended Modification	I*
CLB_GM01	The workbook has several typos and grammatical errors, e.g on page 14 “Circle the shape that doesn’t below to the group.”	Especially for primary school students correct grammar and spelling is very important, thus the app should be checked and errors fixed.	H



ID	Usability Observation	Recommended Modification	I*
CLB_GM02	<p>The Spanish version is half-Spanish and half-German (on a phone with German OS) other parts are in English (e.g. end of quiz message).</p> 	<p>Make sure all strings are translated to the selected language.</p>	H
CLB_GM03	<p>If the teacher selects the math game, students are shown the buttons to select the kind of math game (e.g., addition) and the game does not progress automatically from there.</p>	<p>Fix this possible bug.</p>	H
CLB_GM04	<p>When switching away from the student waiting screen, the app briefly shows the buttons to select “Shapes”, ... to the student.</p>	<p>Instead, switch to the correct activity straight away.</p>	M
CLB_GM05	<p>It looks like the “Shape Game” has several levels, but it is unclear how to access anything other than Level 1.</p>	<p>Either make other levels accessible or remove the label “Level 1”.</p>	M
CLB_GM06	<p>In the “Shapes and Maths game” there is a black shadow plane that appears when moving back from the marker.</p> 	<p>Remove this plane to avoid confusion.</p>	M
CLB_GM07	<p>At the end of the shape game, if you lost all your lives there is a reset button that does nothing.</p>	<p>Remove the reset button or make it interactive.</p>	L
CLB_GM08	<p>At the end of the shape game if you made it through all the questions there is a “next” arrow that does nothing.</p>	<p>Remove it or make it interactive.</p>	L
CLB_GM09	<p>Same menu issues found in Geography application (i.e., CLB_GG01, 14, 22-30)</p>	<p>Same recommended modifications.</p>	L



ID	Usability Observation	Recommended Modification	I*
CLB_GM10	<p>The app starts in German (if the operating system is German), although this is not available in the list of languages.</p> 	<p>Make sure the app starts in one of the languages on the language list.</p>	L
CLB_GM11	<p>The label on top of the camera focus area says “card”.</p>	<p>Should say “workbook” instead.</p>	L
CLB_GM12	<p>The shape game cannot be played without sound, as only the voiceover announces, which shapes are to be found (at least for some of the questions, for example the first one).</p>	<p>Add the name of the shape to be found as text as well.</p>	L

Positive observations

- The application has multi-language support and works cross-platform.
- The application has sizable lessons and quizzes to explore.
- The application and learning materials are fun, colourful, and interactive, which should engage young students.
- The sound inside the application is a nice addition, which adds immersion to the application.
- The collaboration feature, which allows the teachers to monitor the students, shows that the application designer considered real user scenarios in the design process.
- Having a game “Catch the owl” to play while waiting is very nice for the students on the student waiting screen.

Improvements after heuristic evaluation

20 low-, 9 medium-, and 17 high-importance issues were identified during the heuristic evaluation of the ARETE Geography and Geometry Applications. The Cleverbooks developers have already fixed most of these issues. The application's usability, consistency, and stability were considerably enhanced after the update. Table 7 shows some of the most noticeable improvements.



Table 7. Examples of the usability improvement and bugs fix of ARETE Geography and Geometry app

Before usability improvement	After usability improvement	Improvement description
		<p>The “Shapes and Maths game” no longer has a black shadow plane that appears when moving the marker.</p>
		<p>Choice of animals appear correctly in the geography application’s animal game</p>
		<p>Choice of cultural heritage appear correctly in the geography’s application heritage game</p>

4.2 Focus Groups with Teacher Coordinators

4.2.1 Before the intervention

Similar to Pilot 1, we used the Padlet platform and a focus group to gather feedback and requirements from the teacher coordinators in the Pilot 2 teacher Coordinators workshop (cf. Section 3.2.1). We received more feedback in Padlet as compared to Pilot 1, presumably because of the higher number of participants. Screen capture of teacher coordinators’ feedback for Pilot 2 is shown in Figure 2. The focus group was conducted with 9 teacher coordinators, and it ran for approximately 1 hour.

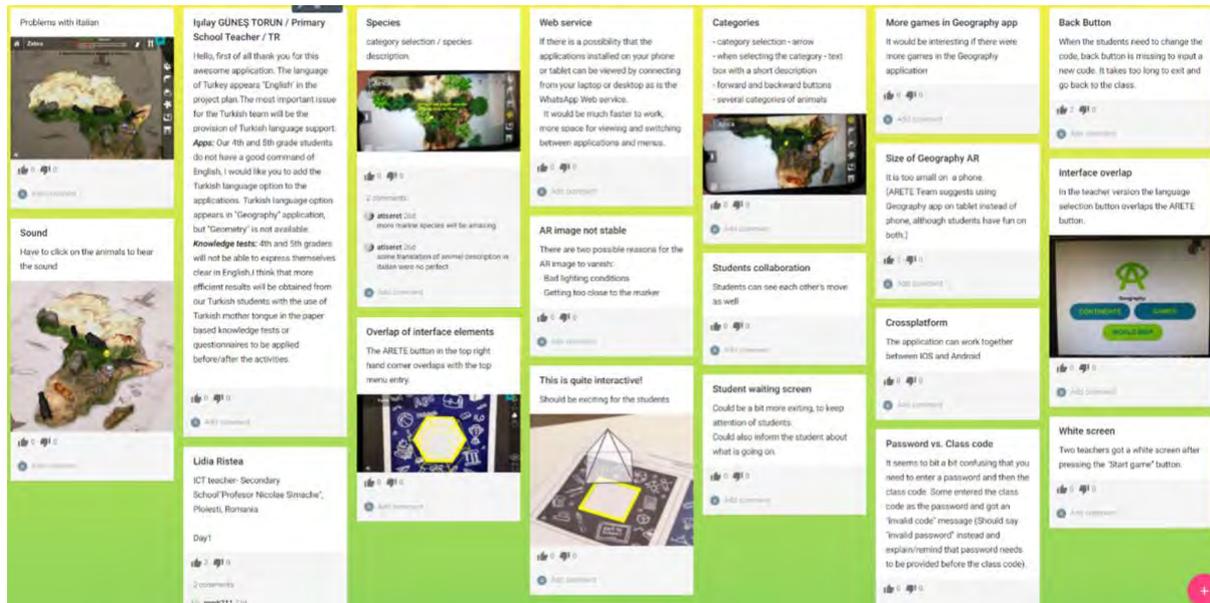


Figure 2. Screenshot of teacher coordinator's feedback on Padlet for Pilot 2 (pre-intervention).

On the first day of the workshop, teacher coordinators had hands-on activities to familiarize themselves with the ARETE Geography and Geometry AR applications from both student and teacher perspectives. Thus, during the focus group, we asked the teacher coordinators some general questions such as “What did you like most?”, “What did you like least?”, and “Do you have any improvement suggestions?” for both student and teacher perspectives.

From the student perspective, teachers praised that the application has multi-language support, as one teacher put it: “I think that it is great that the students will use the application in our language”. Another teacher even alluded to the notion that the application could be used for learning a foreign language. However, there were a few teachers who reported that the translation in their language was not correct and required corrections. The majority of teachers complimented the AR feature and learning materials (e.g., workbooks and maps). For example, one teacher felt that augmenting digital 3D objects on the printed material would show students that the technology can be used as a learning tool. This view was echoed by other teachers who liked that the application was connected to the workbook, as one teacher commented that “it is something that we can use from time to time in our curriculum. Using Cleverbook in combination with the app could be very good for teaching”. Many teachers also praised the application's other features such as “I like that the students can collaborate and work together” and “The sound of the animal is also very nice and engaging”.

From the teacher perspective, the majority of teachers responded positively about the application. They said that the application: “is simple to use”, “have small requirements”, and “can work in different devices and software (IOS and Android)”. Several teachers also commented about the monitoring feature. One individual stated: “I like that I can see the progress of the students, (which is a) really good feature of the app” and another commented: “I like the possibility of control, I can give students some tasks and choose what they will do, can also see how students involve in the task. I can guide students through the experience”. However, opinions differed as to whether the application was practical. Some teachers felt



that the application was flexible and could be used with different levels of class as they could choose materials for their students and thought that some students could explore more than the available curriculum on their own. While other teachers argued that the application would need more customizable options, as one teacher pointed out that they could not see anything that they could change and asked whether they could change 3D objects or materials in the application. There were also some suggestions to separate games by difficult levels (e.g., easy, normal, hard).

While the teacher coordinators generally provided positive feedback about the application, a number of issues were identified in the focus group. In one case, a teacher encountered an error when quitting the lesson. Another teacher also commented that the process to join/re-join the lesson was rather complicated, as they could not join the lesson correctly. Other issues that teachers commented on included: *"cannot see the menu on some devices"*, *"need more explanation regarding the animal"*, and *"cannot see the app on the (Google) store"*. A few teachers also suggested improvements for the user interfaces and user experience. For instance, one teacher pointed out that the application had a button that students did not need to use and it should be removed. Another teacher also expressed frustration that the students had to wait for the teacher to end the lesson and suggested that the application should allow the students to explore freely after finishing the lesson since students were likely to finish the lesson at different times.

4.2.2 After the intervention

We conducted the focus group with the same structure as described in Section 3.2.2. Fourteen teachers attended the one-hour focus group, but only 13 responded to the HARUS questionnaire.

When asked about the training materials distributed before the intervention, we received mixed reactions. According to one teacher, the training materials were beneficial in preparing them to use the application. However, some other teachers identified two shortcomings. First, teachers said that they got the materials after school had already started (October 20, 2021), leaving them with little time to prepare. The second issue was that teachers felt the training materials were insufficient, stating, *"the materials at the beginning were quite limited (just Africa), so it's not enough for me to prepare for the whole year."* Another teacher added, *"I feel like I was left alone to incorporate the materials into the class."* When asked how we might improve the training materials further, the teacher suggested a contact point with a person who understands how the application works or a webinar for discussion. Some teachers also mentioned the two-month gap between the training workshops and the start of the intervention, which they found too long. The teachers also expressed doubts about the accompanying materials (geography map, geography workbook, and geometry workbook) for practical use in the classroom, stating that the materials were too easy for target students and that *"the workbook is entertaining for children, but I am not sure how informative the workbook actually was."* Another teacher suggested that instead of only exercises, learning materials should be included in the book. Additionally, several teachers anticipated more content, as they commented that *"cultural heritage is very poor," "there is no heritage in Greece,"* and *"European plants are very limited."*



Regarding the application usability, Table 8 shows the above-average HARUS scores for both comprehensibility (M = 66.67, SD = 13.10) and manipulability (M = 71.15, SD = 18.90). Nonetheless, teachers expressed concerns about the translation, noting that it was incorrect and did not match the context in numerous languages (Croatia, Serbian, etc.). Additionally, teachers said that there had been cases in which another language was shown when selecting a different language. The teachers also stated that since the device must be held near the workbook or map, the interfaces should be simpler to read from a distance. Moreover, the teachers argued that using the code to join the class took too long, and if there were issues with the teachers' devices, the students would be unable to attend the lesson. One teacher even encountered a circumstance in which students gained access to the teacher's account and distributed a lesson code, making the lesson inaccessible to the teacher and other students. Thus, they suggested that we simplify the lesson joining process and introduce another way for students to access the lesson. Furthermore, other issues were reported in the focus group, including bugs (e.g., "*there are no cards in the animal game,*" "*some games did not even work; only the flags game was usable,*" "*cannot see all the students connected to the game*"), freezing (e.g., "*students' applications froze*"), content mistakes (e.g., "*incorrect answers in the quiz*"), and incompatibility (e.g., "*multiple devices are not compatible with the app*").

Table 8: Pilot2 After Intervention HARUS Questionnaire Results

#	Statements	Mean	SD
Q1	I think that interacting with this application requires a lot of mental effort.	3.15	1.61
Q2	I thought the amount of information displayed on screen was appropriate.	4.38	1.39
Q3	I thought that the information displayed on screen was difficult to read.	3.23	2.01
Q4	I felt that the information display was responding fast enough.	4.62	0.84
Q5	I thought that the information displayed on screen was confusing.	4.08	1.27
Q6	I thought the words and symbols on screen were easy to read.	4.23	1.19
Q7	I felt that the display was flickering too much.	4.15	1.61
Q8	I thought that the information displayed on screen was consistent.	4.15	1.10
Comprehension Score (%)		66.67	13.10
Q9	I think that interacting with this application requires a lot of body muscle effort.	4.69	1.54
Q10	I felt that using the application was comfortable for my arms and hands.	4.77	1.48
Q11	I found the device difficult to hold while operating the application.	3.54	2.06
Q12	I found it easy to input information through the application.	4.00	1.36
Q13	I felt that my arm or hand became tired after using the application.	4.08	1.94
Q14	I think the application is easy to control.	4.15	1.35
Q15	I felt that I was losing grip and dropping the device at some point.	4.15	1.87
Q16	I think the operation of this application is simple and uncomplicated.	4.77	0.80
Manipulation Score (%)		71.15	18.90

Nonetheless, teachers praised the application for its accessibility and expressed optimism for future iterations, stating that "*the application is very simple to use*" and "*the application by itself is good, and I am optimistic that the translation will be solved in the future.*" Furthermore, the teachers happily shared that their students were passionate about the application and physically engaged in using it. For instance, one teacher said that "*our students liked feeding the animals, being active, and playing games.*" Additionally, the



teachers also lauded the application's child-friendly user interfaces and collaborative capabilities, noting that *"the colour and the sound of the animal were very good; there was also a wide range of content"*, *"students had freedom to explore and were not afraid of using the application,"* and *"the students have a chance to collaborate and help each other."*

Aside from fixing the language and technical issues, the teachers requested a customising option. They pointed out that the in-app quiz was currently static and not adaptable to different local curricula. Hence, they proposed that teachers should be allowed to change the quiz or select the questions that would appear on the quiz. Another feature request was for result history; teachers claimed it was nice to see the students' scores, but they could not see the history of the results and specific questions students answered incorrectly, so they could not perform the evaluation afterwards. One teacher also suggested adding more AR interaction to the geometry app, saying that the AR interaction in the geometry app was limited in comparison to the geography app.

5. Pilot 3: Interaction Design of PBIS Prototypes

In this section, we detail the collaboration with WP5 which contributes to the design of AR-PBIS prototypes. In Section 5.1, we discuss the proposal and design considerations for the user interface of the AR-PBIS prototype using heuristic evaluation. As the AR-PBIS application is still under development at the time of writing this deliverable, WP4 compiled usability requirements as a guideline to ensure that the AR-PBIS application would be suitable for young students. The usability requirements are listed in Section 5.2.

5.1 Proposal for the PBIS User Interfaces Design

We have evaluated several UI features and proposed improvement suggestions, which are presented in the following.

5.1.1 General Interface improvement suggestions

The "Leaderboard" icon should be removed and this functionality should be accessed by clicking on the "Token" icon in the top left-hand corner. This helps to declutter the interface and makes use of the token icon, which is used for displaying information only.



5.1.2 Menu bar suggested improvements

Problem: the current icons are difficult to interpret. The students may misunderstand the icons or misuse the associated applications, ending up in the wrong mode and thus not being able to achieve their goals. To solve this problem, we suggest:



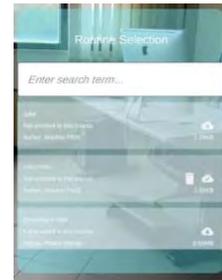
- Provide labels in the user interfaces to name each of the icons.
- Provide a tutorial on the first startup of the application (maybe including the Alien pointing at the different buttons, explaining what they do or could use more traditional visualisations, like labels with an arrow or pointy bit towards the button or overlay with descriptions to explain each icon's functionality.



- c. Combine “Play mode” (left button) and “Explore/discovery mode” (middle button) into one, by enabling access to Mirage XR scenarios also via a QR tag. We think that for the users it is unclear why there is a separation between the two, if possible, the separation should be removed from the interface. The chat button can then go to where the “Leaderboard” button is now (top right-hand corner of the screen).

5.1.3 Routine selection improvements

Problems: Students could be overwhelmed by the list of available routines in the Moodle, which are displayed in the UI. In addition, students may have difficulty searching and selecting the correct routine. Or accidentally or deliberately selecting the wrong list entry, disrupting their learning. To mitigate this issue, we suggest two possible solutions:



Solution 1: Inspired by a discussion with the WP5 partner, we propose an animation illustrated in the storyboard in Figure 3a. First, the Alien flies down in a UFO saucer, jumps out of the saucer, and asks “Which routine do you like to learn?”. Then the UFO saucer becomes a rotatable wheel as a UI object for the students to choose from the available options (Figure 3b). This kind of interactive visual can be fun for the students.

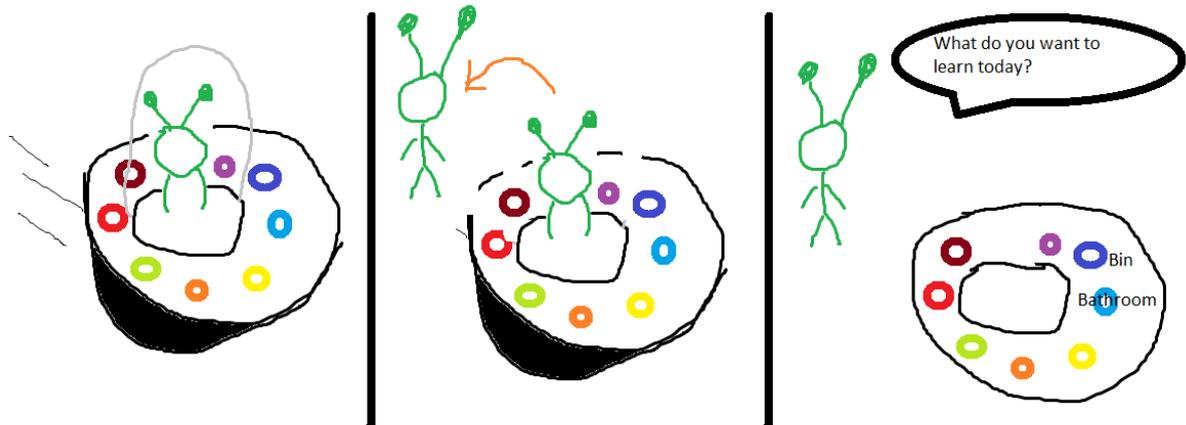


Figure 3a: Alien flies down in a saucer (left); Alien departs the saucer (middle); Alien initiates the dialogue bubble and the saucer tilts to full view showing the routines options as buttons (right)

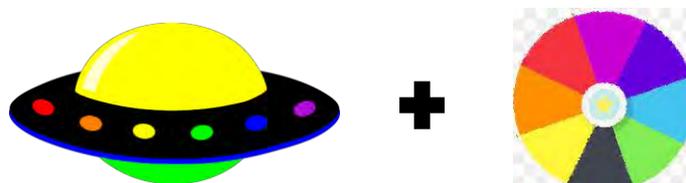


Figure 3b: Visuals of an example saucer and a rotatable wheel

Solution 2: Provide a feature to generate a QR code for selected routines in MirageXR for the instructor to print and put on appropriate locations. When the students scan the QR code the PBIS app can take the students directly to the associated routine. This method allows the

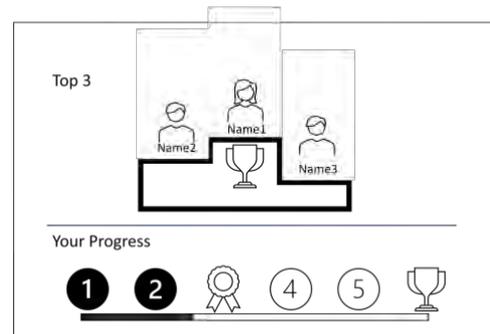


students to access the MirageXR routine (“Play mode” left button in the menu) in the same way the students access PBIS AR scenarios (“Explore/discovery mode” middle button in the menu). This solution would make finding a solution for selecting the topic and scenario for today’s lecture obsolete, as this information would be contained in the QR code that has been generated for this scenario. In addition, the menu user interface can be combined or removed, referring to Section 5.1.2 point c.

5.1.4 *Leaderboard design suggestions*

The leaderboard could be a great motivation tool for students; nevertheless, careful considerations are needed to avoid discouraging the students who are at the bottom of the list. Discussed design considerations are listed below:

- Show only the top 3 students on a podium like in the Olympics games to recognise student achievement and motivate them;
- Display a progress bar to show a student's personal progress.
- Enable the students to collect badges and tokens, which can be exchanged for real-life benefits.
- The ranking might not be based on the number of earned badges, since the students may earn all the badges sometimes. However, if ranking is based on the completion time, there would be a danger that students would rush to a solution without carefully thinking about the content; thus, we recommend not to use the time to rank students. We suggest adding some multiple-choice questions with more than two options with each option carrying different points, which can lead to more variation in the points earned; therefore, the leaderboard would become more meaningful. For instance, after asking a binary (true/false) question “How do you evaluate the behaviour?”, we can ask follow-up questions such as “what is wrong with the behaviour?” or “what is correct with the behaviour?” Nonetheless, we can imagine that it needs more work of the CNR or VU team to develop such multiple-choice questions.



5.1.5 *AR mode text consistency issues*

Some text in the user interfaces do not have a background which makes them difficult to read over the camera stream. A semi-transparent colour background, such as the prompt in the below figure, can be used to make the text more prominent. In addition, it would improve the user interfaces design consistency across the applications.





5.2 Usability Design Requirements for AR-PBIS Applications

In collaboration with WP5, we have formulated usability design requirements targeting primary school students for AR-PBIS application. Usability is a particularly important design consideration for students to learn behavioural routines during PBIS lessons with high efficiency, effectiveness and pleasure. Based on the ARETE systematic literature survey (D4.1), analysis of user requirements (D4.2), students questionnaire results (D5.1), and general usability design guidelines (LaViola et al. 2017), six usability design requirements have been identified and summarised as follows (details are presented in D5.2):

1. *Learnability*: the system must enable students who are inexperienced with the app to quickly comprehend the tasks inside the app and determine the necessary steps to complete the tasks.
2. *Simplicity*: the system must provide clear and clean user interfaces.
3. *Engagement*: the system must be engaging and fun to use,
4. *Help*: The system must provide help options when the students get stuck or lost.
5. *Suitability*: The app and lesson design should consider real-world classroom scenarios and limitations.
6. *Feedback and Error Handling*: The system should provide feedback for every student's action, especially when students interact with the AR elements.

6. Pilot 4: MirageXR

The following section presents the collaboration between WP4 and WP3 in preparation for Pilot 4, which aims to evaluate the MirageXR authoring toolkits from the teacher perspective. Heuristic Evaluation of MirageXR was conducted iteratively during the development process to ensure that the teachers can use MirageXR to create lessons with high efficiency, effectiveness, and satisfaction. Furthermore, user-based usability tests with ten participants were conducted to further evaluate the usability of MirageXR with users unfamiliar with Augmented Reality. The MirageXR application can be run in HoloLens, tablets and phones.

6.1 Heuristic Evaluation of MirageXR

The main evaluation for MirageXR version 1.5 took place in July 2021 when two HCI specialists conducted the heuristic evaluation study for about 3 hours (cf. Section 2). Follow-up heuristic evaluations of MirageXR subsequent releases (version 1.7 and 1.8) took place periodically following an iterative design process. The equipment involved include: Microsoft HoloLens 2, an Android phone (Oneplus8T Android 11.0.12.12 Model KB2003 12GB RAM).



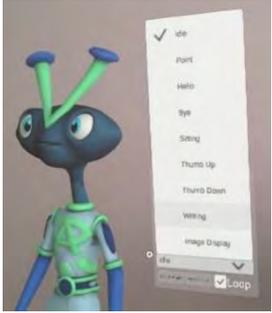
6.1.1 Findings

Based on our observations and discussions regarding the usability of the MirageXR version 1.5, we recommended a list of modifications shown in Table 9 (MXR = MirageXR; I* = importance; Section 3.1.1). For the subsequent releases of MirageXR, we list our follow-up observations and discussions separately in Table 10.

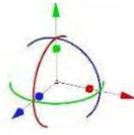
Table 9. Findings of heuristic evaluation of MirageXR v. 1.5

ID	Usability Observation	Recommended Modifications	I*	Developer Responses
MXR _1	<p>3D objects placed out of physical spaces on creation</p>  <p>(In this figure, “task station”, represented by an orange diamond, is placed behind a physical wall and selecting ray is blocked, and users cannot interact with UI elements). 3D objects appear beyond available physical space, which makes the objects difficult to manipulate</p>	Placed 3D objects models within available physical space.	H	Developers have now introduced a SpawnPoint, which is visually indicated with - currently - a red sphere. This still requires some testing, I believe, whether it is handled now by all augmentations consistently.
MXR _2	Selecting rays blocked by physical objects (also shown in issues1 figure). Users cannot select panels or buttons behind the physical objects because selecting rays are blocked by physical objects.	Reconfigure ray-cast filter to filter out physical objects or placed 3D objects models within available space	H	For EditMode only - in PlayMode, it must be used to handle occlusion. For Hololens, this seems to be the case already - but should be checked for consistency. On mobile devices, I am not sure, whether recent changes have picked this up already. It is important to be able to select all virtual elements in EditMode unobstructed. But in PlayMode, this must be a feature.
MXR _3	The activity step list does not follow the “task-station”. Users	The activity step list should follow the	H	We have now refactored the voice commands (to



ID	Usability Observation	Recommended Modifications	I*	Developer Responses
	<p>cannot find the activity step panels when users move from one space to another.</p> 	<p>“task station”. In addition, additional space is needed on the activity list title bar for selecting</p>		<p>call action list panel) and since it anyways only exists on HL1 and HL2 (and no longer on mobiles), I close this ticket.</p>
MXR _4	<p>Unavailable animations should be greyed out. Users might find it confusing when selecting an animation and the setting reset to “Idle”.</p> 	<p>Grey out unavailable animations</p>	H	<p>Removed unavailable animations from animation list</p>
MXR _5	<p>Transparent menu panels</p>  <p>(In this figure, the transparent annotation list blocked the users from interacting with the activity list). Users may try to interact with objects behind transparent panels and be blocked by them, which can cause confusion.</p>	<p>Avoid using transparent panels.</p>	H	<p>Disable "Raycast target" for empty rows (so selecting ray can go through)</p>



ID	Usability Observation	Recommended Modifications	I*	Developer Responses
MXR _6	Cannot replay 2D/3d audio in authoring mode. Users cannot debug the record sound.	Add a replay button during authoring mode	H	Actually you can by pressing the annotation button. When you press it the audio editor will be opened again and you can play or even record and replace the audio again.
MXR _7	Lack of scale function. Users cannot change the scale of 3D objects to grab learners' attention, this is important for 3D objects such as pictures, labels, and visual effects (VFX).	Provide handle widgets for scale manipulation	H	Some models can be resized. We need to investigate which augmentations do not have this functionality and add it to them. The scale should be stored in ToggleObject.scale
MXR _8	Lack of an in-application explanation. Multiple features lack in-application explanation, users might not know about each feature. 	Showing help text when a cursor hovers a UI object.	H	Add Hover tutorial text for character setting
MRX _9	Lack of 3D objects manipulation widget. Users can only manipulate 3D objects through freehand manipulation, which is difficult for new users or mobile users. This makes the 3D objects difficult to place in the desired location.	 Provide handle widgets for 3D manipulations.	M	We had the handle widget above for the character models at some point, and they were too fiddly to use. The bounding boxes we had in the past on Sketchfab – 3D objects handle widgets were added for character and some models.
MXR _10	Small 3D UI objects for AR, especially “+” and scroll bar.	Increase the UI objects size for both	M	In v1.6 we have a brand new UI which is



ID	Usability Observation	Recommended Modifications	I*	Developer Responses
	<p>Multiple 3D UIs objects are too small for AR, and users may have a difficult time interacting with 3D UI objects, especially on mobile devices.</p> 	<p>HoloLens and mobile versions (For mobile, integrated 3D UIs into a flat menu as planned)</p>		<p>developed only for mobile devices. For HoloLens we will work on the related ticket (18)</p>
MXR_11	<p>Duplicate “+” sign buttons. Users might be confused by different button functionality, especially in the mobile version when both appear at the same time.</p> 	<p>Use two different icons</p>	M	<p>This is resolved in 1.6 - only the plus button in the screen space UI remains.</p>
MXR_12	<p>Selecting-ray passes through a character in some parts of the model. Users might find it difficult to select or manipulate characters.</p> 	<p>Have colliders that match the character model to improve selecting-ray collision detection</p>	M	<p>Using teleporting ring design in VR user interface, so the user can use the ring to move the character and use the arrow to see the character facing direction.</p>
MXR_13	<p>The modification panel often overlaps with its character model. Users might be frustrated when interacting with the modification panel.</p>	<p>Move the modification panel to the side to avoid overlapping</p>	M	<p>Fixed in v1.6</p>



ID	Usability Observation	Recommended Modifications	I*	Developer Responses
				
MXR _14	<p>3D objects are not snapped to surfaces. Users cannot judge whether 3D objects are placed on the surface or floating, which can make objects such as waypoints or characters appear unnatural.</p> 	Snap 3D objects to the surface, wall, or floor when nearby.	M	This was now refactored along with issue 12, so that the marker ring around the character is tacked to the floor plane.
MXR _15	<p>3D objects placed out of view (e.g., behind user) on creation. Difficult finding 3D objects, objects can be placed in an unknown location, and users have to search in a 3D space to find the objects.</p>	Provide a temporary directional indicator to guide the users when clicking on the object names on the annotation list.	M	Currently, all augmentations will be spawned on top of the annotation menu. Maybe we can just specify it with a colour. Like a red sphere
MXR _16	<p>Panels rotation. Users have to view the panels at an angle, multiple panels do not face the users directly.</p> 	<p>Change from “billboard” behaviour to curved UI around the users based on task station, similar to Oculus’s curved UI.</p> 	M	Resolved (on mobile this is replaced with screen space UI, on HoloLens we have implemented the curved UI).
MXR _17	<p>Characters do not face the user on creation.</p>	<p>Make sure the character is always facing the users when initialised.</p>	M	Fixed in v1.6



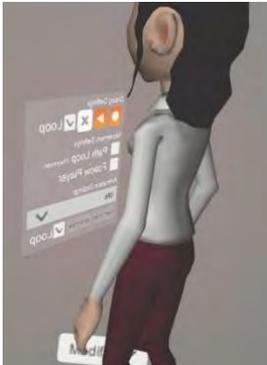
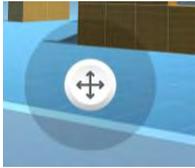
ID	Usability Observation	Recommended Modifications	I*	Developer Responses
	 <p>The panels next to the character are hard to manipulate when the characters are facing away from the users.</p>			
MXR_18	<p>The augmentation menu presents as a list. Users have to scroll and select annotations from a 2D menu, which is difficult to select.</p> 	<p>Consider using 3D objects with icons to replace buttons, eliminate scroll bars if possible. As an example, the figure below shows 3D objects interfaces by spatial.io.</p> 	L	Replaced with larger 2D square icons design based on HoloLens UI menus.
MXR_19	<p>A short dropdown list requires multiple steps. In a dropdown list, users have to select a small dropdown button, then select another small entry.</p> 	<p>Replace short dropdown list with dedicated buttons</p>	L	Simplify dropdown to radio button

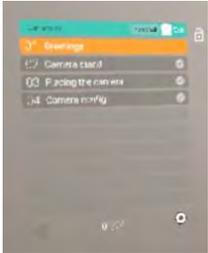
Table 10. Follow-up recommended modifications for the release of MirageXR v.1.6-1.8

ID	Usability Observation	Recommended modification	I*
MXR_20	3D objects still go through the floor and make them difficult to interact with. Selecting ray still	Ensure that selecting ray is not blocked by the physical objects.	H



ID	Usability Observation	Recommended modification	I*
	stuck on the floor or obstacles, make editing 3D objects below the floor or behind obstacles very difficult.		
MXR _21	Character models detached from the ring after creating a walking path, it seems like there are issues with the character collision. 	Ensure that the character models are not detached from the ring	H
MXR _22	(HoloLens2 Only) Inconsistent hand interaction. Currently, direct hand interactions are enabled only in some 3D models and 2D UIs, and the users might be confused by the inconsistent modes of interaction.	Enable direct hand interactions for all 3D models and 2D UIs for HoloLens	H
MXR _23	Added 3D models often appear dark, which may be caused by the lack of lighting or only use fixed lighting in the 3D environment. This prevents the user from observing the 3D objects from different viewpoints. 	Allow users to place other types of lighting, such as point light and spot light, in the scene.	M
MXR _24	(Mobile only) Lack of screen space control option for mobile UIs. With the limited screen space on the mobile platform, the user can have difficulty interacting with 3D widgets.	Adding 2D Joystick on the screen space for more precise control of 3D objects manipulation in mobile user interfaces. 	M



ID	Usability Observation	Recommended modification	I*
MXR_25	Lack of “Undo” option; thus the users cannot fix their mistake easily. Fixing the mistake in a 3D environment can be arduous, which could discourage the users from using the application.	Add an Undo button, to undo 3D actions such as, move, rotate, and scale.	L
MXR_26	Confusing step navigation UIs on the step list. Currently, the users can move to the next step by pressing the left-right arrows under the step list. But, users will try to select the step directly on the step list, which is currently not interactable, and become frustrated. 	Enable the user to select the next step or previous step directly on the step list, and "grey out" not interactable steps. For example, if the current step is 2, steps 1 and 3 should be selectable and "grey out" step 4 forward.	L

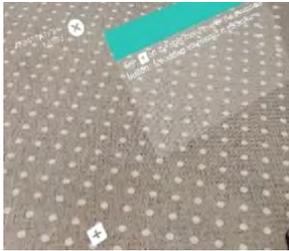
6.1.2 Improvements after heuristic evaluation

During the heuristic evaluation, 2 low-importance, 9 medium-importance, and 8 high-importance issues were found while using MirageXR. Follow-up heuristic evaluations found additionally 2 low-importance, 2 medium-importance, and 3 high-importance issues. After the update, we found that MirageXR significantly improved in terms of usability, consistency, and stability. However, several issues are still currently addressed in the development process, including multiple fixes, optimization, redesign of the user interface for mobile platform, which will be reported in the future deliverable. Some of the improvements from the heuristic evaluation are shown in Table 11.

Table 11. Examples of the usability improvement of MirageXR after the developers’ updates

Before usability improvement	After usability improvement	Improvement description
		The improvements of the character marker that is not overlapped with the character model and also show the direction that the character is facing



Before usability improvement	After usability improvement	Improvement description
		<p>Previously, the user had to view the panels at an angle and multiple panels do not face the users directly. After the changes, panels are angled to face the user for easier interaction.</p>
		<p>Previously, users had to select augmentation from a list menu, which is difficult to select in AR due to the icon's thin size. The update replaces the icon with larger 2D square icons design based on HoloLens UI menus, which make them easier to select.</p>

6.2 User-based Usability Tests of MirageXR

In the section, we present the material, procedure and results of the user-based usability tests of MirageXR.

6.2.1 Basic features for an example XR lesson

There are five basic features (F1-F5) of XR lesson authoring toolkits of which MirageXR is an instance. An example XR lesson making use of these MirageXR features has been developed. Participants were given this example prior to creating their own XR lesson, enabling them to apply these features and provide detailed feedback on them.

- *F1: Temporal tool:* teachers should be able to plan their lessons and students' learning sequence
- *F2: Teachers' representation:* teachers' presence in the classroom could help students become more immersed and engaged.
- *F3: Viewer guidance:* teachers should be able to communicate with students about the best vantage point from which to see the instruction.
- *F4: 3D objects and visual cues:* teachers should have a multitude of 3D objects and visual cues on hand to use in their XR lesson.
- *F5: 3D user interfaces:* teachers should have access to user interfaces that are straightforward and clean, allowing novice teachers to quickly grasp.

Participants were asked to assume the role of photography teachers, teaching the basics of tripods and camera setup. This topic was chosen because it would be easy to demonstrate and allow various types of interactions. To begin creating the XR lesson, participants had to divide the lesson into three steps (F1). In the first step, the participants had to create 3D characters (F2) that introduce the students to the teaching subject, i.e., a physical tripod. The participants were asked to use the available 3D user interfaces (F5) and task stations (F3) to



configure the 3D character to face and point at the tripod and record their introduction dialogue. Then, the participants should also place labels (F4) on specific parts of the tripod as well. In the second step, the participants were instructed to use the ghost features (F2) to record themselves demonstrating how to properly ensemble the tripod. The participants also had to view their ghost recordings to ensure that the demonstration was done properly. In the final steps, the participants had to find a camera model (F4) from Sketchfab and place it on the tripod and use *act* visual cues (F4) to provide students with instruction on how the camera should be placed on the tripods. Finally, a picture was taken using MirageXR features (F4) and the image was placed near the virtual camera model as an example for their students.

6.2.2 Procedure

Ten volunteering participants (three females and seven males, aged 20–50) were recruited. Their disciplinary backgrounds ranged from art undergraduate students to engineer post-graduate students. No one in the study had any prior knowledge of the MirageXR application. In terms of previous XR experience, five had none, while five considered themselves to be beginners. Four of the experienced participants use XR yearly and one uses it monthly. After completing the consent form and pre-test questionnaire, we showed the participants a 12-minute introductory video that demonstrated how to use MirageXR. Then, the participants were given a five-minute training session to help them get acquainted with the Microsoft HoloLens2 interface. Following the practice session, participants were given an example of a photography XR lesson (see above) and asked to create their own XR lesson based on the example provided. Participants were told to follow a think-aloud protocol to describe their thinking process and actions throughout the evaluation. Following the XR session, participants completed a post-experiment questionnaire for subjective evaluations and a semi-structured interview for feedback. The entire user study took around 50 minutes.

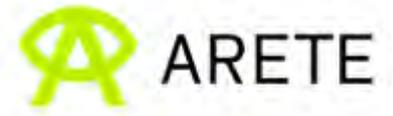
To get a better understanding of how participants used the MirageXR, we captured their point of view using the Microsoft HoloLens2 capabilities, which resulted in approximately 251 minutes of video footage. We also collected subjective ratings of a task's difficulty, enjoyment, focus (Sauro et al., 2009), and mental effort (Zijlstra et al., 1985) in the post-test questionnaires. Also, we used a modified version of HARUS: Handheld Augmented Reality Usability Scale (Santos et al., 2015) to measure MirageXR's usability based on comprehension and manipulation score. We also used Simulation Sickness Questionnaire to quantify simulator sickness (SSQ).

6.2.3 Results

In the following text, we first present the quantitative results based on the analysis of the post-test questionnaires, followed by the qualitative results derived from the video analysis.

Quantitative findings

Seven participants finished the XR lesson in about 25 minutes, and two finished it with more than 30 minutes. One participant skipped some steps and could not complete it. Descriptive statistics of the post-test questionnaires are shown in Table 12. On the 7-point Likert scale, participants assessed the task's difficulty as kind of challenging (Q1: $M = 3$, $SD = 0.74$), which corresponded to participants' mental effort evaluation of "rather hard to do" to "pretty hard



to do" (Q4: $M = 63$, $SD = 25.24$). However, the majority of participants (Q2: $M = 5.1$, $SD = 1.56$) enjoyed the experience and were able to concentrate on the task (Q3: $M = 5.2$, $SD = 1.03$).

Table 12. Usability post-test questionnaire results

Handheld Augmented Reality Usability Scale (1: Strongly Disagree – 7: Strongly Agree)							
#	Statements	T1	T2	T3	T4	Mean	SD
Q1	I think that interacting with this application requires a lot of mental effort.	5	7	5	4	5.25	1.09
Q2	I thought the amount of information displayed on screen was appropriate.	5	4	5	3	4.25	0.83
Q3	I thought that the information displayed on screen was difficult to read.	2	2	4	4	3	1.00
Q4	I felt that the information display was responding fast enough.	2	1	3	4	2.5	1.12
Q5	I thought that the information displayed on screen was confusing.	2	6	2	3	3.25	1.64
Q6	I thought the words and symbols on screen were easy to read.	6	7	7	5	6.25	0.83
Q7	I felt that the display was flickering too much.	1	2	2	3	2	0.71
Q8	I thought that the information displayed on screen was consistent.	5	4	6	4	4.75	0.83
Comprehension Score (%)		66.67	47.92	66.67	54.17	58.85	8.12
Q9	I think that interacting with this application requires a lot of body muscle effort.	1	1	1	2	1.25	0.43
Q10	I felt that using the application was comfortable for my arms and hands.	4	4	7	6	5.25	1.30
Q11	I found the device difficult to hold while operating the application.	2	2	7	3	3.5	2.06
Q12	I found it easy to input information through the application.	5	6	6	4	5.25	0.83
Q13	I felt that my arm or hand became tired after using the application.	2	2	1	2	1.75	0.43
Q14	I think the application is easy to control.	3	6	6	4	4.75	1.30
Q15	I felt that I was losing grip and dropping the device at some point.	2	1	1	2	1.5	0.50
Q16	I think the operation of this application is simple and uncomplicated.	6	6	5	4	5.25	0.83
Manipulation Score (%)		72.92	83.33	79.17	68.75	76.04	5.61

The HARUS score indicated further issues, with lower-than-average scores for both comprehensibility (Table 12: Q5-12, $M = 46.04$, $SD = 16.68$) and manipulability (Table 12: Q13-20, $M = 43.33$, $SD = 16.97$). Further inspection of the HARUS scores revealed that the participants gave particularly low ratings to questions 5, 8, 13, 16, and 18, which correlated to mental effort, responding time, bodily exertion, inputs, and controls (Table 12). In terms of simulation sickness, the analysed Virtual Reality Sickness Questionnaire (Kim et al., 2018) and SSQ scores yielded no significant difference between before and after the test, indicating that MirageXR was unlikely to induce simulation sickness.

Qualitative results

To determine the root causes of the usability issues, we analysed video footage from the participants' point of view and counted the number of times they struggled to perform a



particular action throughout the usability test. The issues were classified according to the categories shown in Table 13.

Table 13: Video analysis result divided by categories

#	Category description	Total	SD
Comprehension issues			
1	Participants misunderstood the UI, interacted with incorrect UIs, or tried to interact with non-UI elements.	83	2.97
2	Participants cannot find the menu panels/objects in 3D environments.	31	1.14
3	The target interfaces were obstructed by other interfaces or face away from participants.	19	1.30
4	Participants cannot find the correct option in the 2D menu.	17	1.35
5	Participants misunderstood that the function was already activated, e.g., they thought that the ghost/voice was already recorded, and vice versa.	15	1.36
6	Participants had problems recover from error/mistakes. (e.g., participants select “follow player” by mistake then cannot manipulate the character).	5	0.67
7	Participants misunderstood the features’ function, e.g., trying to talk to an AI 3D character while selecting pre-recorded.	4	0.66
8	Participants have problems dividing steps, skipping step creation, or creating steps in the wrong order.	3	0.46
Manipulation issues			
9	Participants’ gestures were not registered or missed the target UI.	133	4.58
10	Participants used the wrong gesture to interact with the UI (scroll instead of air-tap, direct interaction instead of air-tap).	57	3.00
11	Participants had problems with 3D object manipulation, moving, rotating, and scaling (scale when intending to rotate).	24	1.28
12	Participants accidentally selected an unintended target.	4	0.66
Issues by steps			
	Step 1: Create Step (F1), Character (F2), Label (F4). Then, configure the character to face a physical object (F3, F5).	247	5.46
	Step 2: Make a step (F1), record Ghost (F2), and view Ghost (F5).	53	2.49
	Step 3: Create a step (F1), download and position a 3D model (F4), and position act visual cues (F4) near a physical object (F3).	91	4.04

Comprehension issues. The major comprehension issues appeared to be participants’ misinterpreting the purpose of the user interfaces and being unable to locate the correct user interface in 3D environments. The bulk of the issues arose during the first step of the tasks, which required participants to configure 3D character models as their representation. The issue is that participants often want to edit the character model directly, as if they were adjusting a mannequin; their expectations differ from those of MirageXR user interfaces, which employ additional menu panels and bounding boxes to adjust the character model. Furthermore, due to the limited field-of-view of HoloLens2, the character context menu panels and bounding box were often presented outside of the participants’ field-of-view, making them difficult to detect. Character models also obscured and hampered access to



other user interfaces, adding to the participants' frustration. Additionally, the character configuration contains the most complicated features in the applications, which confused the inexperienced participant and many times caused them to pick the erroneous choice.

Aside from the problems encountered in creating character augmentation, we found other frequent user interface issues, such as unnecessary and non-functional buttons. These buttons should be eliminated to prevent additional misunderstanding. These results were consistent with the remarks of participants during the semi-structured interview. The majority of participants requested that 3D menus be streamlined in order to reduce occlusion in 3D environments. To assist users in mastering advanced functions, one user recommended a tutorial video clip within the application to explain how various character augmentation features work.

Manipulation issues. According to our observations, all participants struggled to perform the "air-tap" gesture, which was used to select the bulk of MirageXR's user-interface components, resulting in several mistakes during the evaluation. These issues, however, seemed to be caused by the inexperience of the participants, as seen by a decrease in the issue count as the evaluation progressed. Nonetheless, all participants instinctively attempted to touch the 3D user interfaces many times and complained when the UI was not responsive when touched, which may be caused by the lack of system feedback. These findings suggest that the MirageXR should support direct touch control and provide haptic or sound feedback as much as possible, and that 3D UI should be designed around direct touch control rather than gesture control. This advice was consistent with the opinions expressed by interviewees, who suggested that we replace the air-tap gesture.

Other feedback. Despite the difficulties with the user interface, the majority of participants appreciated the MirageXR for its extensive features. "I like how the users can record themselves and add multiple objects into the scene" one participant said. Another participant praised the MirageXR's available 3D objects such as *visual effects* and *act*, saying, "it's quite lively, and should catch the students' interest." Most participants had a positive perspective and believed that MirageXR would be highly beneficial for the teacher if the user interfaces were upgraded. Regarding the teachers' representation, several individuals said that the *ghost* recording appearance was rather unsettling and that it would be preferable if it seemed more human. They felt that the feature would be better if the *ghost* resembled the MirageXR's 3D character model. Some participants, however, disagreed with this recommendation, arguing that all teachers' representations should be enhanced further since they did not appear human.

6.3 Discussion

Based on the results of the usability evaluations, analytical (heuristic evaluation, Section 6.1) as well as empirical (user-based; Section 6.2), we identified a number of issues with the current prototype of MirageXR user interfaces on Microsoft HoloLens2. We suggest the following improvements.

The usability findings show that the WIMP principle does not translate well to 3D settings because users have an extra cognitive demand when identifying items in the 3D environment, and certain users may not execute gestural control appropriately. Thus, the design of the



menu panels is the key priority for overcoming these issues. The menu panels should be positioned within easy reach of users and should follow them in 3D space, promoting direct touch input, eliminating occlusion, and reducing the need of the "air-tap" gesture control. Another observation is that users interact with virtual things as if they were real; hence, the application should offer interaction that is as close to 3D object affordance as possible. As seen by how participants engage with 3D character models, participants attempt to control the character model in the same way that they would adjust a mannequin; hence, this control technique may be a viable approach for reducing the complexity of customizing the character model. Another alternative for changing 3D character models is to use the ghost recording technique, which seems to be intuitive for participants as seen by reduced issues in step 2 (Table 13), enabling them to "act out" the movement they want for the model to follow.

The system's unresponsiveness was also identified as a serious concern in this usability study. This problem most likely happened because adding 3D objects to the scene typically takes a long time, and MirageXR failed to notify the user of the current system state. Another factor influencing perceived responsiveness is a lack of feedback; numerous times during the evaluation, participants misinterpreted that they interacted with the UI when they did not. MirageXR should keep users aware of the system status by including a loading screen and providing haptic, audio, or visual feedback as they interact with the UI to increase the sense of responsiveness. Furthermore, the MirageXR's UI should be optimised by removing unnecessary buttons to minimise misunderstandings that occurred several times throughout the evaluation. Finally, teachers are unlikely to utilise avatars to represent themselves in XR lessons if they find the avatar to be unsettling and disagreeable. Several participants claimed that the ghost recording appeared unnatural, while certain 3D characters in the collection appeared to slip into the uncanny valley (unease caused by the avatar that resembles a human but not quite); thus, MirageXR should improve the appearance of ghost recording and consider improving character appearance further to avoid the uncanny valley (Moore, 2012).

Based on the results of the usability study and video analysis of MirageXR version 1.8, we suggest the following changes, which are shown in Table 14 (MXR = MirageXR; I* = importance).

Table 14: Improvement suggestions based on the usability evaluation of MirageXR 1.8.

ID	Usability Observation	Suggest solutions	I*
MXR_27	Participants used the wrong gestures to interact with the UI (scroll instead of air-tap, direct interaction instead of air-tap).	Ensure that every UI supports direct touch interaction.	H
MXR_28	Participants cannot find the menu panels or objects in 3D environments. The target interfaces can also be obstructed by other interfaces or face away from participants.	The menu panels should be positioned within easy reach of users and should follow them in 3D space, promoting direct touch input, eliminating occlusion, and reducing the need for "air-tap" gesture control.	H
MXR_29	Participants cannot find the correct option in the 2D menu and have problems configuring the menu, especially the character menu.	The character sub-menu should be rearranged based on the complexity of the features. Frequently used features such as audio recording and animation	H

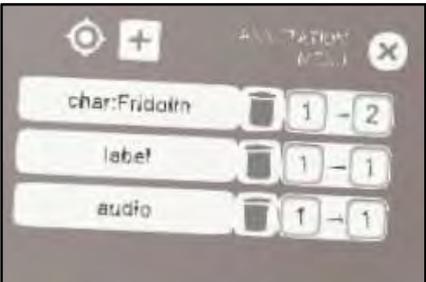


ID	Usability Observation	Suggest solutions	I*
		<p>settings should appear first. Some advanced features should be grouped and hidden away. Buttons such as “play” or “x” buttons should not appear when there is no audio to play.</p>	
MXR_30	<p>Participants were confused between the duplicate “+” button on the activity list and the augmentation list.</p>	<p>Change button to be more explicit: “add activity” and “add augmentation”.</p>	H
MXR_31	<p>Participants select the “+” button when the augmentation menu is already activated.</p> 	<p>Hide the “+” button when the augmentation menu is active, add the “back” button instead to go back to the activity list.</p>	H
MXR_32	<p>Participants could not find the character bounding box control and often selected the “scale” handle when trying to rotate the character.</p> 	<p>Increase the bounding box size to character waist height to make it easier to spot and have enough room to interact with rotate handles.</p>	H
MXR_33	<p>Bounding box UIs for <i>act</i> and <i>VFX</i> lack handles to rotate; only scale handles are available.</p> 	<p>Ensure that rotation handles are available for all bounding box</p>	H



ID	Usability Observation	Suggest solutions	I*
MXR _34	<p>Participants were unsure whether the ghost had already been recorded since the visual indication (a small ghost) was not very clear. Multiple participants thought they have to control the <i>ghost</i> via a small ghost.</p> 	<p>Add voice to notify that the ghost is recording.</p>	H
MXR _35	<p>Participants were unsure whether the activity had already been saved due to the lack of visual or sound feedback</p>	<p>Add a message that the activity is already saved.</p>	H
MXR _36	<p>Multiple participants chose the “Accept” button before performing the recording during the evaluation.</p> 	<p>Hide the “Accept” button before recording or taking a photo for ghost, audio, and image augmentations. Similarly, the play button should be hidden when there is no audio record to play.</p>	H
MXR _37	<p>Multiple participants were confused when the application is unresponsive when adding augmentation to the scene.</p>	<p>Ensure that “loading” and progress bars are used when loading large objects into the scene.</p>	H
MXR _38	<p>Multiple participants commented that the system was not responsive because they were not sure whether they had already selected the button due to the lack of feedback.</p>	<p>Add audio feedback when participants click the buttons.</p>	H
MXR _39	<p>Participants have to create augmentations from the beginning when they want to change text in the augmentation.</p>	<p>Allow users to edit text in the augmentation.</p>	M
MXR _40	<p>Participants struggle to navigate back and forth between the augmentation menu and the existing augmentation list. When participants select the “+” button, the augmentation list disappears, and they cannot go back unless they select one of the augmentations. In the same manner, participants had problems finding the annotation menu when it was replaced by the augmentation list.</p>	<p>Display the augmentation menu and augmentation list side by side instead of switching back and forth.</p>	M



ID	Usability Observation	Suggest solutions	I*
			
MXR_41	<p>Most features lack description, for instance, this menu should have “choose the gender of your ghost recording”.</p> 	Add a brief feature description for most of the submenu	H
MXR_42	Participants commented that the ghost recording was disturbing.	Improve the appearance of the ghost. Some participants suggest using the 3D character model instead of the ghost model.	H
MXR_43	Participants rarely read the tutorial text. This issue might be because the tutorial text is too long, or the tutorial text is outside of the participants’ view.	Add an in-app short video tutorial to demonstrate the feature instead.	L
MXR_44	<p>Participants try to interact with the augmentation list instead of the actual augmentation, which causes some confusion.</p> 	When participants interact with the augmentation list, there should be visual cues to guide participants to the actual augmentation.	L
MXR_45	When participants select “label” augmentation, the keyboard should be available right away. There is no need for participants to select textbox before the keyboard appears.	Automatically select the textbox to bring the keyboard up as soon as participants select “label” augmentations.	L
MXR_46	Participants expected that the ghost recording will also record their finger movements.	Record fingers movement as well	L



ID	Usability Observation	Suggest solutions	I*
MXR _47	Participants expected that taking a photo in the MirageXR will also capture the augmentation as well	Include Hologram in the photo	L

7. Innovative Approaches to Evaluating eXtended Reality (XR)

For evaluating the design artefacts of the ARETE Pilots, the HCI team applied the established human-centred design methods (Section 2), which proved viable as well as valuable. Nonetheless, AR/VR/MR/XR are evolving rapidly with advanced features that entail new evaluation methods beyond the traditional ones (e.g., questionnaires). The HCI team were motivated to explore innovative approaches with the wider research community.

On 19th July 2021, about 25 participants from different lines of Mixed Reality (MR) research got together to present and discuss their work and to get inspired by the work of others. The occasion was the one-day workshop “Beyond Questionnaires: Innovative Approaches to Evaluating Mixed Reality”, which was part of the 34th *British Human-Computer Interaction (HCI) conference*. The goal of this workshop was to explore which evaluation methods are available for MR applications in particular to determine the usability and user experience, and how those methods could be extended and expanded beyond the omnipresent questionnaires. Shortcomings of questionnaires, which motivated this goal, were discussed; for example, the limitations in expressiveness, flexibility, and modality of feedback that questionnaires support, and the indirect influence of end-user's comments, which are typically interpreted by researchers.

The workshop programme consists of two invited talks and six paper presentations, covering a wide variety of MR applications and their evaluation, including a learning tool that brings digital skeletons into the homes of veterinary students (Xu et al. 2021), a design tool for improving collaborating with robots (Branco et al., 2021), and an evaluating tool for trust in holographic artificial intelligence (Huang & Wild, 2021). A tool to support end-users in designing MR interfaces and content was also presented (Heintz & Law, 2021), as well as a mapping space to identify questionnaires or other evaluation instruments for MR applications (Saeghe 2021), and trace plots that allow teachers to analyse student activities in virtual environments and improve their teaching accordingly (Thanyadit et al., 2021).

In addition, two interactive discussion sessions were conducted. For the first one the workshop participants were divided into four groups to discuss the aspects of the future of MR based on the following questions:

- Are MR Glasses next-generation smartphones? Why?
- How should MR technology be enhanced to make it become the future mainstream digital communication device?
- What are possible use scenarios?
- What innovative methods and tools need to be developed to evaluate such MR technology?

After the lively group discussion, a delegate of each group summarised their discussion points to all workshop participants. While each group thought differently in terms of MR enhancements, the consensus seemed to be that the MR Glasses are the devices of the



future; nevertheless, they are unlikely to replace smartphones any time soon since MR technology requires multiple improvements, ranging from technological advancements (e.g., miniaturization, improved interaction) to social solutions (e.g., privacy, trust, acceptance) as well as cheaper solutions to be suitable for the mass market.

In the second interactive discussion section, the participants were asked to identify grand challenges of MR evaluation methodology, based on their experience and research agendas. The discussion was then compiled into the following diagram (Figure 4). For instance, one theme portrayed in the upper left-hand corner is about psychophysiological measurement; existing tools and sensors (e.g., EEG, GSR, EMG, heart rate) require further improvement in terms of accuracy and precision, and innovative solutions are needed to remove noise and interpret the data correctly.

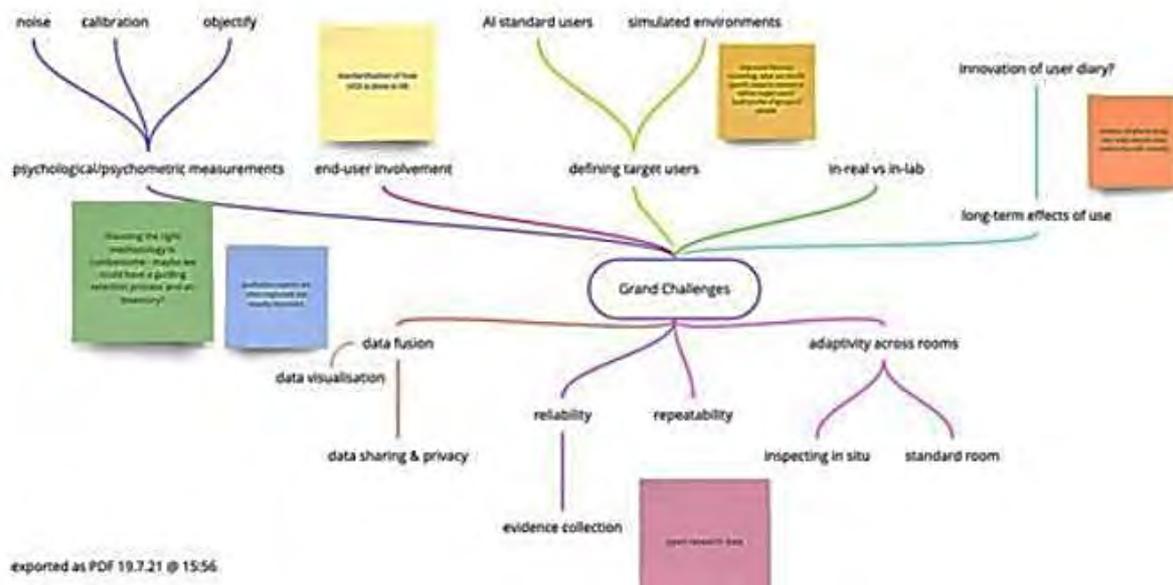


Figure 4. Concept map of the themes arisen from the BHCI workshop group discussion

Overall, the workshop was considered a success with exchanges of ideas, stimulating future work on this burgeoning topic.

8. Conclusion

In the last 12 months the ARETE project has made visible progress in the implementation of Pilot 1 and Pilot 2, the preparation for Pilot 3 and the planning for the newly proposed Pilot 4. WP4 has been actively involved in the design and evaluation of the prototypes of the respective applications, including Read & Spell, Geography/Geometry apps and workbooks, PBIS app and MirageXR. While we have still relied on the use of Heuristic Evaluation, which proved effective for identifying usability problems and recommended modifications, we have employed user-based evaluations through online focus-groups and, more recently, lab-based in-person usability tests, thanks to the relaxation of the pandemic restrictions.

For each of the prototypes evaluated, we identified different strengths and weaknesses as detailed in the preceding sections. Through regular communications with the development



teams of the respective applications, many of the usability issues detected have been resolved satisfactorily. Inevitably, some issues emerged only when the applications were used in real-life contexts (e.g., the device freezing problem in classroom). Comments and suggestions from the teachers provided valuable insights for the future development of ARETE and similar applications. Overall, the work reported in this deliverable lends further evidence to the significance of the Human-centred Design approaches (Section 2), which supports the collection of the evaluation feedback from stakeholders (or their proxies), analysis, and addressing on an ongoing basis the quality enhancement of design and software artefacts and thus user acceptance.

In the coming months, WP4 will continue its role in supporting the development of the PBIS and MirageXR applications in the context of Pilot 3 and Pilot 4, respectively. Furthermore, the HCI team will contribute to the standardisation effort ISO/IEC JTC 1/SC 24/WG 11 "Health, safety, security and usability of Augmented & Virtual Reality (AR/VR)". Based on this work and to further expand it, the team will sustain the effort in exploring innovative HCI methods for designing and evaluating XR-based educational applications.

References

- Branco, D. Silva, P.A., Almeida, J., Menezes, P., Bermúdez I Badia, S., Pilacinski, A. (2021). Virtual Reality, a tool for safe testing of user experience in collaborative robotics. In Proceedings of 34th British HCI Workshop and Doctoral Consortium (HCI2021-WDC). DOI: 10.14236/ewic/HCI2021-W2.3
- Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 189–194). London, UK: Taylor & Francis.
- Heintz, M., & Law, E. L-C. (2021). Beyond Paper: PDart - Participatory Design Augmented Reality tool. In Proceedings of 34th British HCI Workshop and Doctoral Consortium (HCI2021-WDC). DOI: 10.14236/ewic/HCI2021-W2.5
- Huang, X., & Wild, F. (2021). A new metric scale for measuring trust towards holographic intelligent agent. In Proceedings of 34th British HCI Workshop and Doctoral Consortium (HCI2021-WDC). DOI: 10.14236/ewic/HCI2021-W2.4
- Kim, H.K., Park, J., Choi, Y. and Choe, M., 2018. Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Applied ergonomics*, 69, pp.66-73.
- LaViola Jr, J.J., Kruijff, E., McMahan, R.P., Bowman, D. & Poupyrev, I.P., (2017). 3D user interfaces: theory and practice. Addison-Wesley Professional.
- Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research methods in human-computer interaction*. Morgan Kaufmann.
- Moore, R.K., 2012. A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Scientific reports*, 2(1), pp.1-5.
- Saeghe, P. (2021). How To Identify Questionnaires For Mixed Reality Applications. In Proceedings of 34th British HCI Workshop and Doctoral Consortium (HCI2021-WDC). DOI: 10.14236/ewic/HCI2021-W2.2
- Santos, M.E.C., Polvi, J., Taketomi, T., Yamamoto, G., Sandor, C. and Kato, H., 2015. Toward standard usability questionnaires for handheld augmented reality. *IEEE computer graphics and applications*, 35(5), pp.66-75.
- Sauro, J. and Dumas, J.S., 2009, April. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1599-1608).
- Thanyadit, S., Heintz, M., & Law, E. L-C. (2021). Data Visualization for Asynchronous VR Classroom. In Proceedings of 34th British HCI Workshop and Doctoral Consortium (HCI2021-WDC). DOI: 10.14236/ewic/HCI2021-W2.6
- Xu, X., Pan, X., Mangina, E., Kilroy, D., Kumar, A. & Campbell, A.G. (2021). Augmented Reality for Veterinary self-learning during the pandemic: a holistic study protocol for a remote, randomised, cross-over study. In



Proceedings of 34th British HCI Workshop and Doctoral Consortium (HCI2021-WDC). DOI: 10.14236/ewic/HCI2021-W2.1

- Zijlstra, F.R.H. and Van Doorn, L., 1985. The construction of a scale to measure subjective effort. *Delft, Netherlands*, 43(1985), pp.124-139.