KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

XR Laboratory Prototype for Chemical Engineering Education

Jessica Lizeth Domínguez Alfaro

Supervisor:
Prof. dr. ir. Peter Van Puyvelde

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Chemical Engineering

August 2023

XR Laboratory Prototype for Chemical Engineering Education

Jessica Lizeth DOMÍNGUEZ ALFARO

Examination committee:
em. Prof. dr. ir. Jean-Pierre Celis, chair
Prof. dr. ir. Peter Van Puyvelde, supervisor
Prof. dr. ir. Ilse Smets
Prof. dr. ir. Kristel Bernaerts
Prof. dr. ir. Tom Van Gerven
Prof. dr. rer. nat. Thies Pfeiffer
(University of Applied Sciences Emden/Leer)
Assoc. Prof. dr. Daniel Cermak-Sassenrath
(IT University of Copenhagen)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Chemical Engineering

August 2023

© 2023 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Jessica Lizeth Domínguez Alfaro, Celestijnenlaan 200J box 2424, B-3001 Leuven (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface

"*Freedom is nothing but a chance to be better*"

– Albert Camus

This acknowledgement is to all those who, over the last few years and throughout my life, have given me freedom and allowed me to become a better person every day.

First, I would like to thank my supervisor. Peter, I thank you for giving me the opportunity to carry out this project and trusting me every step of the way. Also, I would like to thank all the members of the examination committee for taking the time to read my thesis, suggesting improvements, and above all, for guiding me throughout the whole PhD as part of the CHARMING supervisors. Thank you very much!

Next, I would like to thank the SMaRT group. During this unusual time, you gave me a place (even if I was the black sheep of rheology) and made me part of the group. Thank you all for the great company. In the same way, I thank all the ESRs of CHARMING for being my allies on this unpredictable path. Without you, I would certainly not be here now writing these lines.

Also from CHARMING, I would also like to thank Rabab Nasser for her dedication in managing this project, Hans Muster and Sarah Pascal for welcoming me at ACTA, Matthias Jung, Andrea Raddatz and Sofía García Fracaro for welcoming me at Merck. Thank you all for your support, energy and interest in the project.

I am deeply grateful to Stefanie and Wout, my two master's students, for being instrumental in my personal and professional growth. Thank you both for the great work that we achieved together.

This journey would have been very difficult without the people who came to stay. Alina, I thank you for becoming my new sister, for our long-deep-hilarious talks,

and for always being there, in the fancies meals, but also in the storms. Sou, I thank you for bringing smiles to brighten my day and your company. Yanshen, I thank you for your generosity and your thousand meals. Iurii (chuvak), I thank you for bringing calm to my days and always considering me even when my days were looking very busy. Sanne, thank you for those catch-up Onan coffees and deep discussions of life. Serkan, thank you for listening to my broken French and being nerds together. Also, to my adopted brothers, Michael, Philippe, and Pedro, thank you for letting me be myself and sharing with me who you are.

I would also like to thank my two closest collaborators and now my dearest friends. Michi, I thank you for everything really. For our weekly meetings that brought me back down to earth, for letting me mumble jumble with ideas, for teaching me what you know and for being my accomplice in every step. To Jonas, because definitely, if I had not met you, this thesis would probably have a different shape; I thank you for teaching me, supporting me, rocking with me, welcoming me, and always being available. Thank you both so much for allowing me to grow together with you.

One does not get better overnight; I would like to thank all my friends in life who have grown up with me over the years and are part of my life. In alphabetical order, Bjerke, Bre, Chris, Dany, Lu, Miguel, Mickey, Richie, Sofi, Turi, Vane, Yaz and to all the members of "el tren del ..." Armando, Edgar, Luismi and Mario. I love you all, and I will always be grateful for having you in different steps of my life.

Without a doubt, I would not be the person I am without the freedom that my family has given me throughout my life. Thanks to my parents because they have never, despite the difficulties, stopped my dreams; they have accompanied me and encouraged me to be better every day. I love you both with all my heart. To my dear sister Diana and her new extensions, thank you for taking care of me since I was a "pollito" and for supporting me in everything; I love you, shadow. To the Domínguez-Alfaro family, cousins, aunts, uncles and grandmas and grandpas, thank you for always supporting me and giving me a piece of you. To my adoptive family Trine, Idunn and Kjell, you know I will always be thankful for changing my life. And to my new family Pérez-Rey, for accepting and supporting "La Jessy" as if she was one of the club of the stripes.

Finally, I would like to thank my best friend. Luis, thank you for being "mi país", my adopted supervisor, my therapist, and my partner, and for not only letting me grow as a person but for showing me that with you, I will always be free.

Thank you to every one of you from the bottom of my heart.
Jessie

Abstract

The knowledge acquired during practical laboratory courses is essential for the comprehensive training of any experimental scientist. Unfortunately, not all students have access to this infrastructure due to logistical reasons, lack of resources, or simply unpredictable situations such as a pandemic. Without adequate practical experience, students may face difficulties in understanding abstract concepts and performing future practical endeavours. As such, it is essential to diversify the tools available to support students and strengthen their learning and practical skills.

Augmented Reality (AR) and its use in science education are gaining attention as technology advances. It has been shown that the current affordances of this technology enable learners to interact with knowledge in previously not possible ways, such as visualizing invisible phenomena, training with unsafe situations, or improving spatial skills. Previous studies have extensively explored the benefits of AR for learners, but little is known about how students interact with technology.

This dissertation investigates how augmented reality can be used to support chemical engineering education. To this end, the MAR Lab app, a mobile augmented reality prototype, was designed and developed.

First, an exploratory study was performed on publicly available educational AR applications to help define the basic requirements for the prototype. The findings suggested that most available applications are marker-based, lack content and interactivity, and users experience technical difficulties that hinder their use.

Subsequently, formative evaluations were performed to design, develop and improve MAR Lab. Overall, the application was designed to leverage the main affordances of AR, such as real-time interaction and the blend of the real and digital environment, to provide students with an accessible way to perform experiments without physically being in the laboratory using a handheld device. A titration experiment was used as a case study.

The evaluations specifically investigated how students with different levels of expertise interacted with the application. Furthermore, improvements to the app were proposed by analyzing the log data of their actions using process mining. The results showed that the prototype's usability improved over the iterations and that users could use MAR Lab to learn about titration. However, specific problematic interactions in the AR environment were also identified. Specifically, participants reported problems with tracking the environment and interacting with specific objects. On the other hand, participants of the different groups recognize that they could benefit from the learning experience, for instance, by remembering previous experiences or being introduced to new experiments.

Finally, a summative evaluation of the application was conducted with the target group, namely, novice chemical engineering students. This intervention assessed the relationship between cognitive load, in-app performance and knowledge acquisition. Although limited by its small sample size, the results suggest that cognitive load likely increases with extended use of the application, but knowledge acquisition also increases with more playing time. The cognitive load imposed is moderate, and the participants could improve their performance with additional levels played. Overall, students demonstrated learning from the application with a moderate effect size.

The final use of the MAR Lab app is envisioned as a tool for preparing students during their theoretical courses, before real-life experimental sessions and providing an alternative when experiments must be conducted remotely. By providing students with a way to gain practical experience with experiments, the application could potentially improve the comprehensive training of experimental scientists and help students overcome the challenges of accessing practical knowledge.

Beknopte samenvatting

De kennis die tijdens praktische laboratoriumsessies wordt opgedaan, is essentieel voor de opleiding van elke experimentele wetenschapper. Helaas hebben niet alle studenten toegang tot deze infrastructuur omwille van bijvoorbeeld logistieke redenen, gebrek aan middelen of gewoonweg onvoorspelbare situaties zoals een pandemie. Zonder voldoende praktijkervaring kunnen studenten moeilijkheden ondervinden bij het uitvoeren van praktische zaken en het begrijpen van abstracte concepten. Daarom is het nuttig de hulpmiddelen die studenten kunnen gebruiken om praktische vaardigheden om te trainen, te diversifiëren.

Augmented Reality (AR) en het gebruik ervan in het onderwijs krijgen steeds meer aandacht naarmate de technologie voortschrijdt. Er is aangetoond dat de huidige mogelijkheden van deze technologie lerenden in staat stellen om op voorheen onmogelijke manieren met kennis om te gaan, zoals het visualiseren van onzichtbare verschijnselen, het trainen met onveilige situaties, of het verbeteren van ruimtelijke vaardigheden. Eerdere studies hebben de voordelen van AR voor leerlingen uitgebreid onderzocht, maar er is weinig bekend over hoe leerlingen met de technologie omgaan.

Dit proefschrift beoogt te onderzoeken hoe augmented reality kan worden gebruikt ter ondersteuning van het chemisch ingenieursonderwijs. Hiertoe is de MAR Lab app, een mobiel augmented reality prototype, ontworpen en ontwikkeld.

Eerst werd een verkennend onderzoek uitgevoerd naar openbaar beschikbare educatieve AR-toepassingen om de basisvereisten voor het prototype te helpen definiëren. Uit de bevindingen bleek dat de meeste beschikbare toepassingen op markers zijn gebaseerd, inhoud en interactiviteit missen, en dat gebruikers technische problemen ondervinden die het gebruik ervan belemmeren.

Vervolgens werden formatieve evaluaties uitgevoerd om MAR Lab, ontwikkelen en verbeteren. In het algemeen werd de applicatie ontworpen om de belangrijkste voordelen van AR, zoals real-time interactie en de mix van de echte en digitale

omgeving, te benutten om studenten een toegankelijke manier te bieden om experimenten uit te voeren zonder fysiek in het laboratorium te zijn met behulp van een handheld apparaat. Een titratie-experiment werd gebruikt als casestudy.

In de evaluaties werd specifiek onderzocht hoe studenten met verschillende expertiseniveaus met de applicatie omgingen. Verder werden verbeteringen aan de app voorgesteld door de loggegevens van hun acties te analyseren met behulp van process mining. De resultaten toonden aan dat de bruikbaarheid van het prototype tijdens de iteraties verbeterde en dat gebruikers MAR Lab konden gebruiken om te leren over titratie. Er werden echter ook specifieke problematische interacties in de AR-omgeving geïdentificeerd. Deelnemers meldden problemen met het volgen van de omgeving en de interactie met specifieke objecten. Anderzijds erkenden de deelnemers van de verschillende groepen dat zij voordeel konden halen uit de leerervaring, bijvoorbeeld door zich eerdere ervaringen te herinneren of kennis te maken met nieuwe experimenten.

Tot slot werd een summatieve evaluatie van de toepassing uitgevoerd met de doelgroep, namelijk beginnende studenten chemische technologie. Deze interventie beoordeelde de relatie tussen cognitieve belasting, in-app prestaties en kennisverwerving. Hoewel beperkt door de kleine steekproef, suggereren de resultaten dat de cognitieve belasting waarschijnlijk toeneemt bij langdurig gebruik van de toepassing, maar dat de kennisverwerving ook toeneemt bij langere speelduur. De opgelegde cognitieve belasting is matig, en de deelnemers konden hun prestaties verbeteren met meer gespeelde niveaus. In het algemeen lieten de studenten zien dat ze leerden van de toepassing met een matige effectgrootte.

Het uiteindelijke gebruik van de MAR Lab app is bedoeld als een instrument om studenten voor te bereiden tijdens hun theoretische cursussen, voorafgaand aan echte experimentele sessies, en als een alternatief wanneer experimenten op afstand moeten worden uitgevoerd. Door studenten een manier te bieden om praktische ervaring op te doen met experimenten, kan de toepassing mogelijk de uitgebreide opleiding van experimentele wetenschappers verbeteren en studenten helpen de uitdagingen van toegang tot praktische kennis te overwinnen.

List of Abbreviations

Acronym	Description
APK	Android Application Package
AR	Augmented Reality
CLES	Common Language Effect Size
CLT	Cognitive Load Theory
CTML	Cognitive Theory of Multimedia Learning
DIF	Design Implementation Framework
ELT	Experiential Learning Theory
HMDs	Head-Mounted Displays
ISCED	International Standard Classification of Education
MAR	Mobile Augmented Reality
MCQs	Multiple Choice Questions
NHST	Null Hypothesis Significant Testing
PM	Process Mining
R-TLX	Raw Task Load Index
RP	Rapid Prototyping
SDKs	Software Development Toolkits
SLAM	Simultaneous Localization and Mapping
STEM	Science, Technology, Engineering and Mathematics
SUS	System Usability Scale
TLX	Task Load Index
UX	User Experience
VR	Virtual Reality

Contents

Abstract iii

Beknopte samenvatting v

List of Abbreviations vii

Contents ix

List of Figures xiii

List of Tables xvii

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Problem Statement . . . 2
- 1.3 Objectives and Structure . . . 3
- 1.4 Research Methodology . . . 4

2 State-of-the-art 7
- 2.1 Immersive Technologies . . . 7
- 2.2 Augmented Reality . . . 10
  - 2.2.1 AR types . . . 13
  - 2.2.2 MAR in Education . . . 14
  - 2.2.3 Development of AR applications . . . 16
- 2.3 Theoretical Foundations . . . 17
  - 2.3.1 Learning Theories . . . 17
  - 2.3.2 Experiential Learning . . . 17
  - 2.3.3 Meaningful Learning in Science . . . 19
  - 2.3.4 Cognitive Load Theory & Multimedia Learning . . . 21
  - 2.3.5 Complex Learning & 4C/ID model . . . 22
  - 2.3.6 Inquired based learning . . . 28

2.3.7 Embodied Cognition 28
2.4 Common Perspectives of Learning Theories 29
2.5 Conclusion 30

3 Mobile Augmented Reality Apps in Education 33
3.1 Introduction 33
3.1.1 Background and Related Work 34
3.1.2 User Experience of MAR Applications 35
3.1.3 User Reviews and App Data as a Source of Information 35
3.2 Methodology 36
3.2.1 App Extraction 36
3.2.2 Review Extraction 37
3.2.3 Review Classification 38
3.2.4 App Characteristics Extraction 39
3.2.5 Classification under Fields and Levels of Education . . . 39
3.3 Results and Discussion 40
3.3.1 What are the main problems, feature demands, and opinions expressed by the users of educational MAR apps published in the Google Play Store? 40
3.3.2 What are the main characteristics of price, size, rating score, installs, release dates, and AR types of educational MAR apps? 47
3.3.3 Which are the levels and fields of education that educational MAR apps published in the Google Play Store target? 51
3.4 Limitations 52
3.5 Conclusions 54

4 Design and Development 57
4.1 MAR Lab Design and Development 57
4.1.1 Research Methodology 57
4.2 Low Fidelity Prototype 58
4.2.1 General Concept 59
4.2.2 Analysis Needs and Educational Content 59
4.2.3 Defining Learning Objectives 62
4.2.4 Learning Experience Design 62
4.2.5 Level Structure 64
4.2.6 User Evaluation 65
4.2.7 Results 65
4.3 Digital Prototype 69
4.3.1 Technical Implementation 69
4.3.2 Level Flow 75
4.3.3 Methodology 77

4.3.4 Results . . . 79
4.4 Conclusions . . . 83
4.5 Acknowledgements . . . 83

5 Second Digital Prototype 85
5.1 Prototype Improvements . . . 86
5.1.1 Redesign of MAR Lab . . . 86
5.2 Technical Implementation . . . 90
5.3 Process Mining . . . 90
5.4 Formative Evaluations . . . 90
5.5 Methodology . . . 91
5.5.1 Instruments . . . 92
5.5.2 Data Analysis . . . 93
5.6 Results . . . 95
5.6.1 ACTA Results . . . 95
5.6.2 MERCK Results . . . 108
5.6.3 KU Leuven Results . . . 120
5.7 Discussion . . . 130
5.8 Limitations . . . 133
5.9 Conclusions . . . 134
5.10 Acknowledgements . . . 134

6 Final Prototype Evaluation 135
6.1 AR & Cognitive Load . . . 136
6.2 Learning & Performance . . . 137
6.3 The Proposed Model . . . 138
6.4 Methodology . . . 139
6.4.1 Procedure . . . 139
6.4.2 Instruments . . . 140
6.4.3 Data Analysis . . . 141
6.5 Results . . . 143
6.6 Discussion . . . 156
6.7 General Recommendations for Practitioners . . . 160
6.8 Limitations . . . 161
6.9 Conclusion . . . 161

7 Conclusion and Outlook 163
7.1 Mobile Augmented Reality Apps in Education . . . 163
7.2 First Prototype of MAR Lab . . . 164
7.3 Formative Evaluations . . . 164
7.4 Summative Evaluation . . . 165
7.5 Outlook of MAR Lab . . . 166
7.6 General Insights for Developing AR Learning Tools . . . 168

7.7 Future of AR in Education . . . 169
7.8 Conclusion . . . 170

A Appendix of Chapter 3 **171**
A.1 Content of metadata extracted for each app . . . 171
A.2 ARdoc tool Architecture . . . 172
A.3 ARdoc Categories . . . 173
A.4 Levels of education according to the International Standard Classification of Education (ISCED) . . . 174

B Appendix of Chapter 4 **175**
B.1 Online Survey . . . 175
B.2 Chemistry Test . . . 179
B.3 Demographic Questions . . . 182
B.4 Statistical Analysis . . . 183

C Appendix of Chapter 5 **185**
C.1 Titration Code . . . 185
C.2 Questionnaires from evaluations . . . 193
C.2.1 ACTA Test . . . 193
C.2.2 Merck Test . . . 197
C.2.3 KU Leuven Test . . . 200
C.3 Interview Questions . . . 204
C.3.1 Details of the app . . . 204
C.3.2 Experience with AR and future use . . . 205
C.4 Themes and Codes . . . 206
C.5 Statistical Analysis . . . 206

D Appendix of Chapter 6 **213**
D.1 Final Questionnaire . . . 213
D.1.1 R-TLX . . . 213
D.2 Statistical Analysis . . . 214

Bibliography **219**

Curriculum vitae **237**

List of publications **239**

List of Figures

1.1 The DIF mapped into thesis structure 5

2.1 Reality-Virtuality Continuum 8
2.2 XReality (XR) Framework 9
2.3 Google Maps Live View 10
2.4 The Housecraft app 11
2.5 National Geographic Explore VR 11
2.6 Ivan Sutherland's HMD 12
2.7 Tom Caudell's AR prototype 12
2.8 Marker-based AR example 13
2.9 Location-based AR example 14
2.10 Cognitive theory of multimedia learning 22
2.11 4C/ID Model 23

3.1 The General methodology for data processing of app reviews . 37
3.2 Problem discovery broad topics 42
3.3 Information giving broad topics 43
3.4 Feature request broad topics 45
3.5 Weighted average score per install range 49
3.6 Educational MAR apps released since 2013 50

4.1 Rapid Prototyping Model adapted to MAR Lab 58
4.2 Titration set-up 60
4.3 Indicator colour change 60
4.4 Titration Curve from weak acid- strong base 61
4.5 Design of the MAR Lab learning experience based on the 4C/ID model 63
4.6 Inquiry-based cycle phases mapped into MAR Lab's stages . . 65
4.7 Low-fidelity prototype 66
4.8 Low-fidelity prototype proposed modifications 67

4.9 Overview of the software used for MAR Lab's development . . 69
4.10 MAR Lab's Unity Hierarchy 71
4.11 Custom Event Code 73
4.12 TrainAR Interact Action 74
4.13 TrainAR Grab Action 74
4.14 TrainAR Combine Action 74
4.15 MAR app level flow 76
4.16 First User Study Design 78
4.17 SUS score of first prototype 80
4.18 Learning Test: Boxplot First study 81

5.1 Introductory Screens of MAR Lab 87
5.2 MAR Lab modified flow 88
5.3 Interaction with the logbook 89
5.4 Marie Curie's cartoon 89
5.5 User Study Design of the Second Prototype 91
5.6 Data Sources and Tools 93
5.7 ACTA Test 96
5.8 SUS Scores for ACTA test 97
5.9 Normalized values of each SUS item: ACTA test 98
5.10 Learning Test: Boxplot ACTA operators 102
5.11 Learning Test: Boxplot ACTA students 102
5.12 Number of hints given to ACTA participants 103
5.13 ACTA operators' Process Map 106
5.14 ACTA students' Process Map 107
5.15 Merck Test 110
5.16 SUS Scores for Merck test 110
5.17 Normalized values of each SUS item: Merck test 111
5.18 Learning Test: Boxplot Merck group 114
5.19 Number of hints given to Merck participants 115
5.20 Merck Process Map Level 1 116
5.21 Merck Process Map Level 2 117
5.22 Merck Process Map Level 3 118
5.23 SUS Scores for KU Leuven test 121
5.24 Normalized values of each SUS item: KU Leuven test 122
5.25 Learning Test: Boxplot KU Leuven 124
5.26 Number of hints given to KU Leuven participants 125
5.27 KU Leuven Process Map Level 1 127
5.28 KU Leuven Process Map Level 2 128
5.29 KU Leuven Process Map Level 3 129
5.30 SUS Scores of all tests 131

6.1 Model relating in-app performance, knowledge acquisition, and cognitive load . . . 138
6.2 Final User Study Design . . . 140
6.3 Data Sources and Tools . . . 142
6.4 MAR Lab Final Evaluation . . . 144
6.5 SUS Scores for KU Leuven final test . . . 144
6.6 Normalized values of each SUS item: final KU Leuven test . . . 145
6.7 Number of hints given to KU Leuven students . . . 147
6.8 Final Evaluation: Level 1 Process Map . . . 149
6.9 Final Evaluation: Level 2 Process Map . . . 150
6.10 Final Evaluation: Level 3 Process Map . . . 151
6.11 Learning Test: Boxplot Final Evaluation . . . 152
6.12 Boxplot of Total Time per Level . . . 153
6.13 Boxplot of Number of Assembly Errors per Level . . . 154
6.14 Boxplot of R-TLX average score of KU Leuven Students . . . 155

A.1 ARdoc tool architecture . . . 172

C.1 Titration Curve from strong acid- strong base . . . 186

List of Tables

2.1 Popular AR Platforms 16

3.1 Classification of informative reviews 40
3.2 Apps per type of AR 51
3.3 Apps per level of education 51
3.4 Classification of the apps according to ISCED 53

4.1 Most common issues on low-fidelity prototype 67
4.2 Results of log data recorded 80

5.1 Custom events registered by Unity Analytics 95
5.2 Problems reported by 26 ACTA participants 99
5.3 Problems reported by seven Merck participants and three trainers 112
5.4 Problems reported by 29 KU Leuven participants 123

6.1 Average R-TLX scores and cumulative data distributions of scores per task reported by Grier [1] 155

A.1 ARdoc Categories 173
A.2 Levels of education according to ISCED 174

B.1 Paired Samples T-Test 184
B.2 Test of Normality (Shapiro-Wilk) 184
B.3 Descriptives 184

C.1 Thematic Analysis Codes 207
C.2 ACTA operators: Wilcoxon signed-rank test 208
C.3 ACTA operators: Test of Normality (Shapiro-Wilk) 208
C.4 ACTA operators: Descriptives 208
C.5 ACTA students: Paired Samples T-Test 209
C.6 ACTA students: Test of Normality (Shapiro-Wilk) 209

C.7 ACTA students: Descriptives 209
C.8 Merck test: Wilcoxon signed-rank test 210
C.9 Merck test: Test of Normality (Shapiro-Wilk) 210
C.10 Merck test: Descriptives 210
C.11 KU Leuven test: Wilcoxon signed-rank test 211
C.12 KU Leuven test: Test of Normality (Shapiro-Wilk) 211
C.13 KU Leuven test: Descriptives 211

D.1 Final Evaluation Learning Test: Paired Samples T-test 215
D.2 Final Evaluation Learning Test: Test of Normality (Shapiro-Wilk) 215
D.3 Final Evaluation Learning Test: Descriptives 215
D.4 Total time per level: Descriptives 216
D.5 Total time per level: Friedman Test 216
D.6 Total time per level: Conover's Post Hoc Comparisons - RM Factor 1 216
D.7 Assembly errors per level: Descriptives 217
D.8 Assembly errors per level: Friedman Test 217
D.9 Assembly errors per level: Conover's Post Hoc Comparisons - RM Factor 1 217
D.10 Bayesian Informative Hypothesis Testing: Descriptive statistics 218
D.11 Bayesian Informative Hypothesis Testing: Covariance matrix . 218

Chapter 1

Introduction

1.1 Motivation

When someone goes to the supermarket to buy cooking vinegar for sauerkraut or gherkins for burgers, how can one choose the best product for this purpose? In principle, one would have to know the ideal acid concentration needed; otherwise, the pickles would simply not be as appetising. While it is true that high school or undergraduate students studying chemistry have been introduced to the basic concepts of acids and bases, they can, in theory, make an educated choice when cooking. It is important to note that understanding these principles may differ among individuals. At this level of education, students are expected to be introduced to the fundamental concepts of acids and bases. However, it is not uncommon for some students to struggle with grasping these concepts.

Traditionally, students are introduced to the subject of acids and bases through theoretical lectures and lab work, which can provide a helpful foundation for understanding these concepts. However, these approaches may not always be sufficient for fully comprehending and applying these principles, especially if they are not equally available to all students. Many factors influence a student's learning journey, such as their resources, the resources of their institution, their teachers, their personal characteristics, and motivation, making the teaching-learning process complex. According to Novak's theory of human constructivism and meaningful learning [2], the role of educators is to facilitate knowledge construction by providing learners with diverse experiences that enable them to self-manage it. This is especially important in science due to the complexity of the subject matter being taught. Therefore, it is crucial for educators to

provide students with a variety of learning experiences to facilitate meaningful and long-lasting learning.

1.2 Problem Statement

According to Education 4.0, experience-based learning is an essential part of the new paradigms that will enable future professionals to solve tomorrow's challenges [3]. Hands-on activities can help students develop a deep understanding of abstract concepts by allowing them to map the concepts into concrete experiences [4] and improve their performance [5]. This approach can also help them develop valuable skills such as problem-solving, critical thinking, collaboration, and communication. Additionally, experience-based activities can increase student motivation and engagement by enabling active participation in their own learning [6].

In the context of Science, Technology, Engineering and Math (STEM) education, laboratory practice is a key form of experience-based learning. Science and engineering degrees typically include laboratory activities that provide students with hands-on experiences that support the concepts they learn in theoretical courses. For instance, engaging laboratory experiences have shown to enhance students' conceptual understanding as a result of the conceptual reflections on experiments, improve their performance, data analysis and confidence while performing experimental work, and enhance their ability to explain and correct their own misconceptions [7].

Nonetheless, new challenges, such as reduced time and frequency of laboratory sessions, have created the need for alternative ways to provide laboratory practice to STEM students in order to foster valuable learning experiences. Moreover, conventional laboratory experiences may sometimes fail to support student learning and limit their scope to following "cookbook" recipes. As a result, students may not fully exploit the opportunities that experiments can provide [8]. Thus, novel approaches that complement laboratory practice and provide students with engaging and effective learning experiences are crucial.

Emerging technologies, such as Virtual Reality (VR) and Augmented Reality (AR), have provided alternative hands-on experience in recent years. Both AR and VR have been used in chemical education to support diverse learning activities [9, 10, 11, 12, 13, 14, 15].

In particular, VR has been used to develop scenarios for training laboratory skills, and studies have suggested that virtual environments can enhance learners' confidence when performing chemical experiments, as well as serving as pre-

training for real laboratory practice [16, 17]. Similarly, AR has been shown to be effective for training students in laboratory skills, with studies showing that students can gain confidence when using AR tools to perform experiments [18].

When comparing VR and AR, AR is generally considered to be less intrusive to the user [19] and can be deployed on handheld devices, making it an accessible tool for students [20].

Data from several studies suggest that AR can provide a range of learning benefits, such as improving spatial abilities, increasing memory retention, decreasing cognitive overload, and boosting learners' motivation [19, 21, 22, 23]. Additionally, AR has been shown to have a positive effect on students' academic performance and achievement [24, 25, 26, 27, 28].

On the other hand, using AR in educational settings also presents some challenges. For example, from a pedagogical perspective, it may not always be evident how to integrate educational content and AR technology [19, 29]. Moreover, ensuring that all students can access the learning environments is one of the priorities of schools and institutions [30].

However, using technology in education presents a range of challenges. For example, many schools lack the necessary infrastructure and support systems, and the cost and maintenance of certain technologies can be high, which prevents teachers from using them in their practice. As a result, inequalities among students can arise. To ensure technology, including AR and VR, is used effectively and equitably to support learning and education, it is essential to find sustainable and scalable solutions. To achieve this, researchers, developers, and practitioners in the field of educational technology must prioritize finding solutions to these challenges.

1.3 Objectives and Structure

As part of the CHARMING project[1], this work explores how augmented reality (AR) can support chemical engineering education. Mobile AR addresses limitations in resources and offers unique learning affordances. Therefore, a mobile AR prototype called **MAR Lab** is proposed as a supportive interactive learning environment. MAR Lab incorporates instructional design principles to enhance the user experience and learning outcomes, allowing users to experiment with titration experiments. Additionally, the TrainAR framework [31] is leveraged to address issues of scalability and accessibility. Two specific objectives are proposed: 1) to develop a tool that supplements the learning experience of

[1] https://charming-etn.eu/

students without being physically present in a chemical laboratory, and 2) to investigate the system's usability among undergraduate chemical engineering students and explore their perceptions of usefulness and requirements. Overall, this work aims to answer the following research question:

How can mobile augmented reality be used to support learning chemical practical skills?

1.4 Research Methodology

The Design Implementation Framework (DIF) guides the research described in this thesis. This framework was chosen because, unlike others, the DIF recognises usability and user experience as part of the learning experience and implements feedback and evaluation at various points in the iterative cycle [32].

The DIF comprises five phases: problem definition and assessment, ideation, design and user experience, experimental evaluations, and feedback and implementation. The framework is also based on different user-centred design methodologies and is adapted to educational technologies by emphasising three key principles: participatory design, iterative design and usability. Figure 1.1 shows the correspondence between each chapter and each framework phase.

The initial phase consists of defining and assessing the problem related to educational technology. Chapter 1 presents the research problem statement, and Chapter 2 provides an overview of related research areas such as immersive technologies, mobile augmented reality (MAR), and pedagogical and psychological theories.

Next, the ideation phase involves researchers and developers collaborating to develop possible solutions. Chapter 3 presents the exploratory study of mobile augmented applications to guide the design, focusing on usability and user experience. The next chapter also complements this phase.

The third phase is called design and user experience. In this phase, designs are refined using usability and user experience methods. In addition, the design can be refined using methods such as rapid prototyping until no apparent problems remain. Chapter 4 presents the design and development of the first digital prototype of MAR Lab using the TrainAR framework [31] and the rapid prototyping methodology [33, 34, 35]. Moreover, Chapter 5 describes the formative evaluations and iterations of the prototype using a mixed methodology and process mining.

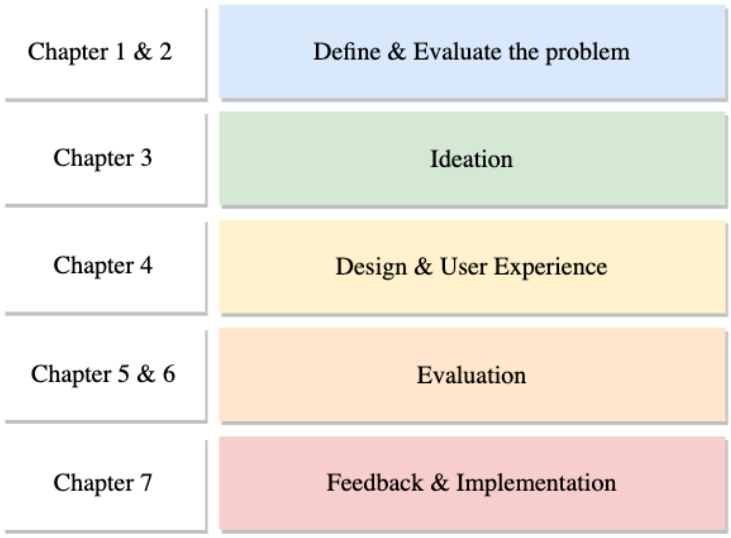


Figure 1.1: The Design Implementation Framework mapped to the structure of the thesis. The framework was proposed by [32] and comprises five phases: defining and evaluating the problem, ideation, design and user experience, experimental evaluations and feedback and implementation.

Experimental evaluations include testing the working prototype through laboratory experiments or real educational environments and assessing the impact on learning, e.g. motivation, interest and learning outcomes. Chapter 6 presents the results of the design of the evaluation framework for the application and the effects on the cognitive load and performance of the target group.

Finally, the prototype is deployed in authentic learning environments during the implementation and feedback phase. Chapter 7 summarises the conclusions and provides recommendations for future research and insights into the prototype.

Chapter 2

State-of-the-art

Section 2.2.3 is partially adapted from: S. Solmaz, J. L. Domínguez Alfaro, P. Santos, P. Van Puyvelde, and T. Van Gerven, "A practical development of engineering simulation-assisted educational AR environments," Education for Chemical Engineers, vol. 35, pp. 81-93, 4 2021.

The research presented in this thesis combines insights and methods from different disciplines. In this chapter, the relevant literature from each of these disciplines will be reviewed, focusing on the most influential works for the design and development of the prototype. The state of the art in immersive technologies, AR systems, and education will be explored, and the psychological and pedagogical principles that inform our approach to learning and instruction will be discussed. By providing this context, the aim is to situate this work within the broader research landscape in these fields and explain the rationale behind the design and development decisions.

2.1 Immersive Technologies

Immersive technologies, including augmented reality (AR) and virtual reality (VR), have gained widespread attention and have been explored in various fields, including education, business, marketing, and healthcare. These technologies create immersive experiences that combine, replace or blur the boundaries between the physical and virtual worlds [36].

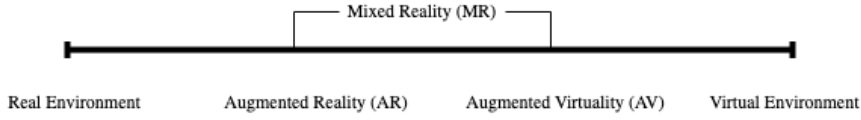


Figure 2.1: Reality-Virtuality Continuum proposed by Milgram et al. [37]

To explain the relationship and differences among these technologies, Milgram and Kishino's concept of the Reality-Virtuality Continuum (Figure 2.1) offers a classification based on the characteristics of the displays. According to these authors, AR is "any case in which an otherwise real environment is augmented by means of virtual (computer graphic) objects" [38]. The opposite case, in which virtual environments are augmented by reality, is known as augmented virtuality (AV). These authors also introduced the term mixed reality (MR) as a concept encompassing the combinations of AR and VR [38, 37]. However, many authors and practitioners have used and introduced new concepts, which have led to the use of inconsistent terminology. Recently, Rauschnabel et al. [39] proposed a new framework called "XReality Framework" that clarifies and expands upon these concepts based on the user experience, providing a classification that distinctions each of these technologies and using the term XR (xReality or XR) as an umbrella for the terms AR and VR and where the X replaces the type of format or reality [39].

According to the XReality Framework, AR is a hybrid experience integrating context-specific virtual content into a user's real-time perception of the physical environment via computing devices. The AR experience can range from assisted reality to mixed reality depending on the degree to which a user experiences the AR objects as being present in their own physical environment (also referred to as local presence). On the other hand, VR is defined as an artificial, virtual, and viewer-centred experience in which the user is fully immersed in a 3D space that is visually separate from the physical environment. The VR experiences can range from atomistic to holistic depending on the degree to which a user feels present in the virtual environment rather than the physical environment (also referred to as telepresence) [39]. This classification does not depend on the devices. It can accommodate the level of differentiation depending on the user experience, thus giving space to include potentially novel developments that may appear in the near future.

Both AR and VR can be situated along a continuum. The place on this spectrum is given, according to the framework, by differences that take into

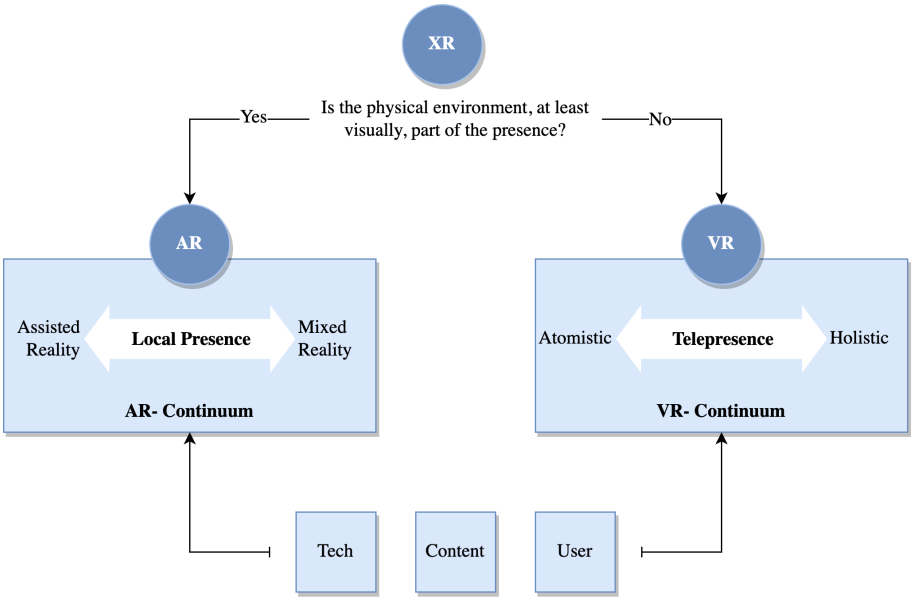


Figure 2.2: XReality (XR) Framework adapted from [39]

account the effects on the user experience, such as technology, content stability, content dimensionality, context, embodiment, interactivity, social experience and problems specific to each, such as augmentative control and motion sickness [39].

At one end of the AR continuum, one can find augmented reality (AR) applications where the user perceives the virtual content as artificial. For example, in the Google Maps application, the user is assisted in reaching a destination through virtual cues such as arrows and text projected on the screen, which are certainly perceived as artificial (Figure 2.3).

At the opposite end of the continuum, one can find mixed reality (MR) applications where the user perceives the virtual content as part of their physical environment. For example, an MR application might allow the user to see an armchair in their living room, even if there is no physical armchair (Figure 2.4).

Virtual reality (VR) experiences can also be placed along their own continuum. At one end, one might find VR applications designed for a specific purpose, such as learning simulations or visualizing molecules. In these cases, the feeling of presence may be less important than achieving the application's main goal.

At the other end of the spectrum, one might find holistic VR applications

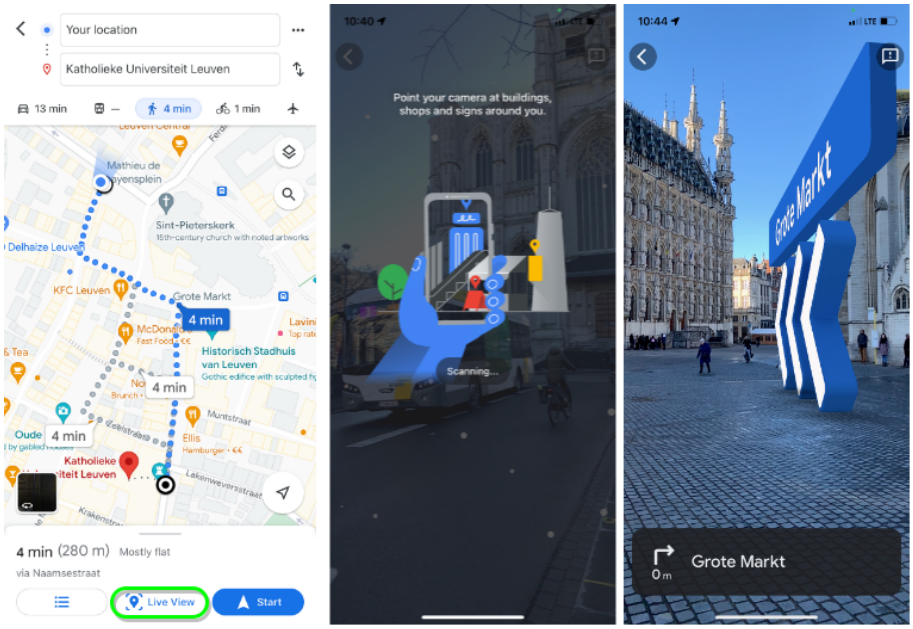


Figure 2.3: Google Maps Live View. Users can select the "Live View" option, scan the environment and get directions overlaid on their physical environment.

that provide users with experiences that are difficult to distinguish from real experiences. For example, a VR game or a National Geographic documentary might provide a realistic, immersive experience (Figure 2.5).

Overall, this XR framework allows the classification of technologies acknowledging important aspects that practitioners and researchers have clarified over the years. In this thesis, this classification is also used to understand AR systems in a broader sense and, when necessary, is extended with other perspectives to address different contexts.

2.2 Augmented Reality

Immersive technologies have grown hand in hand since their inception. In the 1960s, Ivan Sutherland proposed the idea of virtual reality (VR) to create a synthetic environment that could be viewed with a head-mounted device [40] (Figure 2.6). Later, the term Augmented Reality (AR) was introduced by Caudell & Mizell (1992) as a technology to improve the efficiency and quality

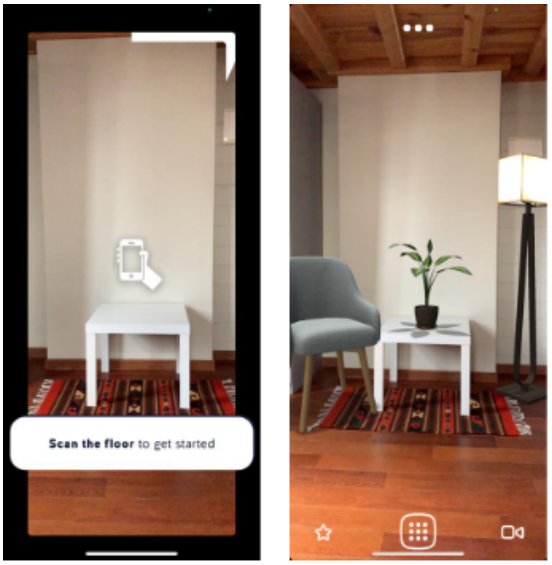


Figure 2.4: The Housecraft app allows users to scan their real-world environment and then virtually insert furniture, decorations, and other objects into the physical scene. The left image shows a user scanning the surroundings, while the right image shows the same scene with virtual items deployed into the environment.



Figure 2.5: National Geographic Explore VR is a virtual reality experience that allows users to explore photorealistic and immersive environments. This application is compatible with Oculus devices. Image from www.oculus.com/experiences/quest/2046607608728563.

of manufacturing activities in the aerospace industry. The main goal was "to augment the user's visual field with information necessary in the performance of the current task" [41] (Figure 2.7). Even though Caudell & Mizell were the first authors that coined AR as a term, a broad and widely used definition of AR was given by Azuma [42, 43].

Generally, AR technology is described as a system that "supplements the real world with virtual (computer-generated) objects that appear to coexist in the

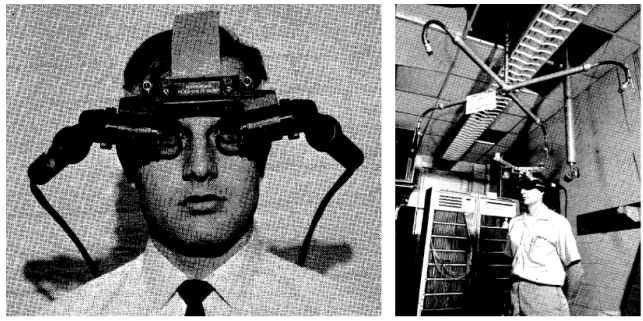


Figure 2.6: Head-mounted display (HMD) created by Ivan Sutherland. The left image shows the HMD in which the optical system magnifies the images on the miniature cathode ray tubes. The right image shows the system in use and the mechanical arm hanging from the ceiling [40].

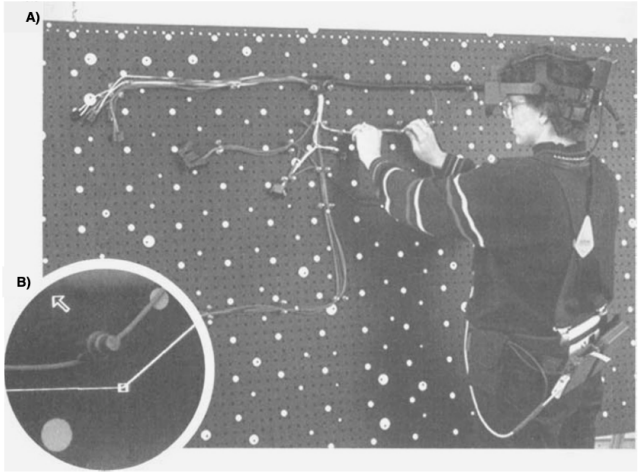


Figure 2.7: Tom Caudell's AR prototype was built by TriSen Systems. Figure A shows the system consisting of a lightweight head-based display and a waist-mounted computer. Figure B shows the enlarged view of the board. The white lines are computer graphic lines that indicate the next position of the wire ([44], p.516)

same space as the real world", and its main features are its capability to combine the real and virtual environments, the real-time interaction, and the alignment of digital content with the real world [43].

2.2.1 AR types

From a technological perspective, Augmented Reality (AR) can be classified into marker-based and markerless systems. To achieve the accurate geometric registration of the virtual content, the pose of the user's viewpoint (or camera) must be estimated with respect to a coordinate system. To do this, a tracking system such as a sensor or camera-based tracking system is required [45].

Marker-based AR utilizes artificial markers or image features to detect the position and display virtual objects [46]. A marker is a distinct visual cue composed of a specific pattern, which is recognized and compared to existing markers in a library containing associated information. By matching the pattern to its corresponding library entry, the system can accurately pinpoint the marker's location and orientation and access the stored information [47, 48]. For example, Figure 2.8 shows a paper bill serving as a marker, with the associated app displaying the 3D virtual object when the user scans the real bill.

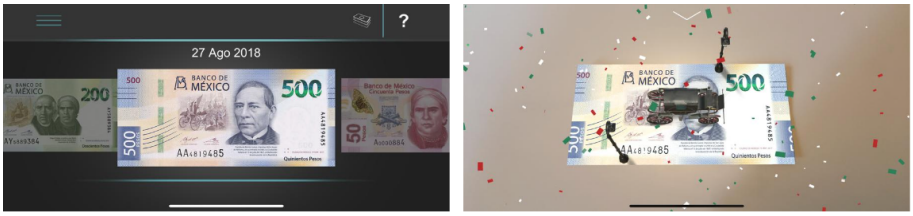


Figure 2.8: BilletesMx app as an example of marker-based AR. When the user scans the physical bill, the bill serves as a marker that triggers the app to display a 3D animated video of the character featured on the bill.

On the other hand, markerless AR utilizes the natural features of the environment to detect and deploy virtual content [45]. To track the unknown environment, simultaneous localization and mapping (SLAM) is normally used. This type of tracking relies on features and depth information [45, 46]. For instance, a smartphone app such as the Housecraft app can be used by users to view virtual objects in their real-world environment (Figure 2.4). This app uses the phone's camera to track the room's features, allowing users to place virtual furniture on the screen as if it were in the space. During the tracking process, an initial 3D point map is created and refined later with new feature points detected in subsequent camera frames [45].

A special case of markerless AR is location-based AR. This type of AR uses position data from devices determined by Global Positioning System (GPS) to deploy the virtual content [49, 29]. An example of this could be a mobile game

that uses GPS to create a virtual treasure hunt, like the application Pokémon Go (Figure 2.9).



Figure 2.9: Pokémon Go is an example of location-based augmented reality (AR) developed by Niantic. Using GPS, the application deploys Pokémons in various locations, allowing users to capture, train and battle the creatures when they get close to their location.

Mobile technology has expanded opportunities for learners to access learning content ubiquitously. Today, integrating other technologies, such as augmented reality (AR), has further enhanced this access. The combination of AR systems with handheld devices is known as mobile augmented reality (MAR) [20]. MAR utilizes smartphones or tablets with a built-in camera to display digital information on top of the real world. This allows users to simultaneously see and interact with the real and digital worlds without requiring specialized hardware such as Head-mounted displays (HMDs). The AR examples presented in Sections 2.1 and 2.2 correspond to mobile augmented reality.

2.2.2 MAR in Education

MAR is gaining popularity in education, as research has shown that AR has several affordances that support meaningful learning. These affordances include the ability to visualize abstract or unobservable phenomena using 3D objects,

annotate spaces with overlaid information, and combine digital and physical objects to create hybrid learning environments [22, 19, 23, 21].

In the context of science education, examples exploiting AR affordances are diverse and have been used to cover various educational needs. For instance, Tee et al. [18] developed a marker-based AR application to simulate a titration experiment in the real physical laboratory without using chemicals – this application rendered liquids by using two markers for the burette and the beaker's exact positions. After testing the application with ten students, their results showed that students could gain confidence when conducting experiments with the tool.

Similarly, Akçayir et al. [50] developed a marker-based AR application to study how students could enhance their laboratory skills in a physics laboratory. Their results suggested that the application helped foster positive student attitudes towards the laboratory. Other user-friendly platforms have been developed by Naese et al. [14], who used the HP reveal platform to support instruction in the laboratory by projecting videos after students scanned the equipment or QR codes with their mobile devices. This approach was found to be an effective way to provide additional information to students. Plunkett [13] used a similar approach to demonstrate the use of the equipment and visualize reaction mechanisms. Additionally, markerless and marker-based AR applications have been used to support the visualization of molecular structures [9, 51], and invisible phenomena such as the flow of electrons [52], or vapour-liquid equilibrium [53]. Overall, these researchers reported positive attitudes from students towards the support provided by the applications in understanding the phenomena.

AR affordances have been linked to various learning effects, including improved spatial abilities, memory retention, and motivation. Meta-analyses have also found that AR has a medium effect size on learning effectiveness, indicating that it can positively impact student's academic performance and achievement [27, 26, 24, 25].

Despite the evident evolution of AR, some limitations exist regarding its use in educational settings. These limitations are related to both pedagogical and technical aspects. Pedagogically, previous research has established that AR can be cumbersome to use and distracting for students. Concerning technological aspects, most reported limitations are related to the usability and design of AR apps. For instance, AR systems can have difficulties maintaining the virtual content and detecting the user's location, and can be technically challenging to implement by teachers as the virtual content is not easy to create [22, 19, 23, 21].

Table 2.1: Popular AR Platforms for creating AR environments

SDK & Developer	Operating System (OS)	Application Programming Interface (API)	Compatible game engine	Tracking target
Google ARCore	Android, iOS	Java, C#, C++	Unity, Unreal	Plane, marker, point cloud and face
Apple ARKit	iOS	Objective-C, Swift	Unity, Unreal	Plane, marker, point cloud and face
PTC Vuforia	Android, iOS, Windows	C#, JavaScript, C++, Java	Unity	Plane, marker, object

2.2.3 Development of AR applications

The popular spread of AR in mobile devices has led to various options for building AR applications. Table 2.1 presents the general information of the popular software for creating AR environments [54].

Most of the software available to create AR environments provides software development toolkits (SDKs) written in different programming languages. These programs facilitate a simple integration inside game engines such as Unity and Unreal. These game engines are similar in terms of core functionalities and support most AR hardware, however, they differ in terms of additional functionality. For example, Unity supports JavaScript and C# programming, while Unreal supports C++. Additionally, Unity is usually more user-friendly for beginners, as it has a large online community [46].

SDKs can be extremely useful for integrating AR capabilities with minimal effort. ARCore and ARKit are Google and Apple's respective AR SDKs, and both can utilize marker and markerless tracking for detecting the same features (planes, markers, face and point clouds) [55]. In terms of AR hardware, ARKit is available on devices with iOS 11 or higher, while ARCore is available on Android devices with Android 7.0 (Nougat) OS or higher [46]. Vuforia SDK, on the other hand, utilizes computer-vision technology to recognize images and 3D objects in real-time, allowing virtual objects to be accurately positioned. This SDK is compatible with Android, iOS, and Windows operating systems [54].

Leveraging the integrated AR resources available in Unity, this thesis presents the development of a prototype and its deployment for Android and iOS devices.

2.3 Theoretical Foundations

This section presents the most influential learning theories that inform the design and development of the application. These theories will also be used to explain or support observations in later chapters.

2.3.1 Learning Theories

There exist several theories to explain how people learn. These theories present differing perspectives on what constitutes learning and how it occurs. Learning theories can be broadly classified into several categories, with the most influential being Behaviourism, Constructivism, and Cognitivism [56].

Behaviourism posits that knowledge is acquired through the principles of conditioning and changes in behaviour. Learning is treated as the acquisition of behaviours disregarding mental processes. The roots of the theory and its variants are attributed to the works of Skinner, Pavlov and Thorndike [57, 58].

Cognitivism emerged as a response to behaviourism and focuses on the mental processes involved in learning, such as acquisition, assimilation, organization, and retrieval. This theory is attributed mainly to the works of Ausubel and Gagné. It defines learning as the change in mental structures (schemata) that are created, organized, assimilated, and reorganized by the learner [59, 58].

Constructivism proposes that individuals actively construct knowledge through their experiences and interactions with their environment relying on cognitive structures. This theory is based on cognitive psychology and rooted in the works of Piaget, and Vygotsky [60, 59, 58].

In the context of science education, the constructivist-cognitivist approaches are particularly influential and will be discussed in greater detail in subsequent sections.

2.3.2 Experiential Learning

The Experiential Learning Theory (ELT), proposed by David Kolb in 1984, is a comprehensive approach to learning that integrates experience, perception, cognition, and behaviour [61]. ELT is grounded in the constructivist ideas of Lewin, Dewey, and Piaget and emphasizes the role of experience as part of the learning process. Specifically, Kolb defines learning as:

> "Learning is the process whereby knowledge is created through the transformation of experience".

According to Kolb, experiential learning is characterized by six key features: (1) an emphasis on the learning process, (2) consideration of learners' prior experience, (3) dialectical conflict as a facilitator of learning, (4) a holistic perspective on human behaviour, (5) the transactions between the learner and the environment, and (6) an understanding of knowledge as a product of the transformation of experience [61].

For Kolb, learning is an **emergent process** rather than an accumulation or transmission of immutable thoughts, in which an individual creates new ideas through personal experience. Knowledge is not an independent entity that can be acquired or transmitted but is instead continuously created and recreated through the transformation process. Moreover, knowledge is derived from prior experiences of the learner, which influences future ones, i.e., "learning is relearning" [61]. Kolb acknowledges that the role of the educator is to recognize these previous experiences and ideas and build upon them [61].

According to Kolb, the learning process involves **tension and conflicts between opposing views of coping with the world**, such as accommodation-assimilation, concrete experience, and abstract conceptualization. Kolb's theory of learning identifies two dimensions in the learning process: one ranging from concrete experience to abstract conceptualization and another spanning from active experimentation to reflective observation. The level of learning is influenced by the learner's movement within each dimension and their ability to resolve conflicts between opposing viewpoints. If only one dimension is emphasized, learning may be hindered. To achieve higher levels of adaptation, such as creativity and personal growth, it is necessary to integrate all four ends of the two dimensions [61].

Kolb acknowledges that experiential learning is a concept that involves thinking, feeling, perceiving, and behaving of a person, and to understand human behaviour, a **holistic view** is more fruitful. Kolb views learning as the primary process of human adaptation, encompassing other adaptive concepts such as creativity, problem-solving, decision-making, and attitude change. It occurs in diverse settings and at all ages. It can take various forms depending on the context, ranging from reaction to immediate situations (performance) to adaptations throughout life (development)[61].

According to experiential learning theory, learning is a **dynamic interaction between the person and their environment**. This interaction is ubiquitous, i.e., happening everywhere at all times, and involves the mutual influence of personal characteristics, environmental factors, and behaviour. Kolb argues

that traditional teaching often overlooks this transaction and fails to recognize the complexity and bi-directionality of the interaction between individuals and their natural environments [61].

Finally, according to Kolb, **knowledge is the result of the transaction** between social (i.e. accumulated cultural experience) and personal or subjective experience. The process of creating knowledge can vary in sophistication, ranging from a child's discovery to scientific research. To effectively teach a subject, it is necessary to understand the nature of the subject matter and how knowledge is created. Kolb emphasizes the importance of considering these factors in order to support learners in their learning process [61].

Aligning with Kolb's idea, learning with the AR prototype is acknowledged as a process. A process that is supported by the tool. The tool considers that the students use the application as part of the learning experience and the interaction with it yields a rich transaction between what they experience with the application and what they know about the topic.

The virtual environment supports the learner with almost genuine experiments that reflect those of reality from which they can learn from.

2.3.3 Meaningful Learning in Science

Constructivist theories, such as Kolb's Experiential Learning, emphasize that knowledge is constructed through experience. This notion is especially applicable to the natural sciences, wherein learners bring with them pre-existing beliefs and ideas that inform the learning process. However, the process of constructing and incorporating new knowledge is often complicated by the presence of prior misconceptions arising from sources such as textbooks, personal beliefs, experiences, and even teachers [2]. These misconceptions can be difficult to address using traditional teaching methods.

Joseph D. Novak proposed Meaningful Learning, a theory to comprehend why instructional approaches fail to effectively support student learning in science and how to assist students in acquiring knowledge [2]. This theory is based on the research of David Ausubel [2].

According to Novak, successful science learners build comprehensive frameworks of related concepts as they construct meanings. In order to engage in meaningful learning, three conditions must be met. First, students must have prior knowledge that can be connected to new information in a non-arbitrary manner. Second, the material being learned must contain relevant concepts that can

be related to existing knowledge. Finally, students must consciously choose to incorporate non-arbitrary concepts [2].

Research has demonstrated that the structure of domain-specific knowledge [62] and level of expertise influence the ability to reason in natural sciences [63] and that conceptual change requires building or restructuring this knowledge framework [2]. To facilitate this process, Mintzes et al. propose providing students with strategies to assist them in developing their own skills and abilities for learning, such as planning, monitoring, controlling, and regulating their learning [2].

According to Novak,

> "Meaningful learning underlies the constructive integration of thought, feeling and action, leading to human empowerment for engagement and responsibility."

Novak classifies learning experiences into three domains. Cognitive, affective, and psychomotor domains. He proposes that for meaningful learning to occur, educational experiences must integrate the three domains that allow learners to connect their knowledge [64, 2]. For example, students learning acid-base theory should not only read concepts (e.g., strong acid-base, weak acid-base, equilibrium constants) but also design and perform laboratory experiments (e.g., titration, pH measurement) that allow them to connect these abstract concepts to decisions they have to make in their daily lives.

As mentioned in Chapter 1, learning resources are sometimes not equally accessible to students. For this reason, this project aims to provide a tool that allows educators to facilitate and students to complement the construction of theoretical knowledge through the use of a mobile augmented reality application. Specifically, the application aims to complement what they learn in their theoretical classes (cognitive domain) with a virtual experiment (psychomotor and affective).

From the general perspective, the use of the application aims to complement or support one part of three requirements for meaningful learning (psychomotor domain). However, the use of the tool can also impact other domains. For instance, modifying the students' attitudes towards the learning content (affective domain). The application is intended for students who do not have access to a laboratory in their curricula. For these students, the application can link the three domains. Additionally, the application can also help students with access to a laboratory by providing prior knowledge and experience that helps them be more efficient.

2.3.4 Cognitive Load Theory & Multimedia Learning

According to Kolb [61], it is necessary to understand both the nature of the subject matter and how knowledge is created to support learners. When learning science, students may struggle to comprehend abstract concepts, as the mental processes required to assign meaning to these concepts can be challenging [4].

The Cognitive Load Theory (CLT) proposes that human knowledge is stored in cognitive schemas, with a limited working memory capacity and unlimited long-term memory [65]. Working memory processes new information through visual and auditory channels, allowing for simultaneous information processing and increased effective working memory capacity [66, 67]. Long-term memory stores cognitive schemas, "chunking" the information and optimising the working memory. The cognitive schemas on long-term memory exist in different complexities and levels of automation [66].

The effects of different types of learning tasks and the associated cognitive load on learners' ability to process new information and construct knowledge are also addressed by CLT. Our human brain has a limited capacity to process information; thus, cognitive load increases when the amount of information exceeds its capacity, leading to impaired learning. Conversely, when cognitive load is low, the brain processes information effectively, resulting in better learning [65]. Instruction imposes three types of cognitive load for learners. The extraneous load, the intrinsic load, and the germane load. The intrinsic load depends on the task complexity and learners' prior knowledge, extraneous load comes from instructional features that hamper learning, and germane load acts as a function that redistributes the memory resources from extraneous load to intrinsic load [65].

Similarly to CLT, the Cognitive Theory of Multimedia Learning (CTML) proposed by Mayer [67] consider the cognitive architecture and processing at the core of designing multimedia environments. For Mayer, active learning happens when the learner effectively organises, selects, and integrates the information acquired through visual and auditory channels, which are limited in their capacity, and builds a coherent mental representation (Figure 2.10). Therefore, when learning, multimedia environments should be designed to guide the cognitive processes without burdening the learner's cognitive system.

When designing learning environments, instructional design strategies aim to reduce extraneous load and align the learner's pre-existing knowledge with the intrinsic load to increase engagement with the learning task. The following section will describe the instructional design model used to design the AR prototype.

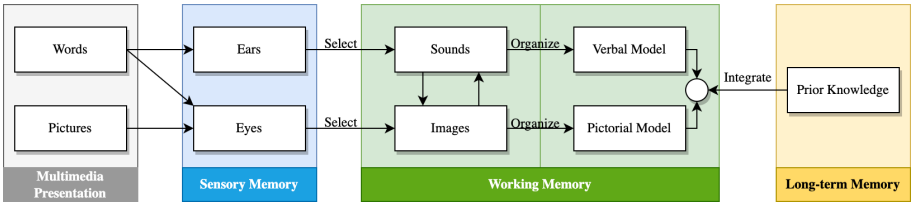


Figure 2.10: Cognitive theory of multimedia learning. Scheme adapted from [67] represents the human information processing system. It is composed of the sensory memory (blue box), which records momentarily words and pictures coming from the multimedia environment (grey box); the working memory (left green box), which is responsible for manipulating the information and transforming it into representations (right green box) and long-term memory (yellow box) which main function is to store the information permanently.

2.3.5 Complex Learning & 4C/ID model

The four-component instructional design (4C/ID) model is based on the principles of Cognitive Load Theory (CLT) and is aimed at aiding practitioners in constructing instructional approaches that facilitate complex learning. Complex learning involves the integration of knowledge, skills, and attitudes, as well as the application of learning to real-world contexts [68, 65].

Developing complex learning skills relies on two cognitive processes: schema construction and automation. When schema construction occurs, it allows the individual to create more intricate schemas by joining together smaller schemas. This helps to organize and store knowledge while reducing the strain on working memory. The automation of schemas happens due to a person utilizing a certain cognitive schema several times and successfully. Doing so reduces the burden on working memory and allows the individual to learn and develop more complex skills [66]. The 4C/ID model offers an instructional design approach to foster the creation and automation of schemas of complex learning.

According to 4C/ID, complex learning comprises four interrelated components: learning tasks, supportive information, procedural information, and part-task practice. Learners can form cognitive schemas by working on learning tasks and studying supportive information. They can automate these schemas by consulting procedural information, repeatedly performing routine aspects of learning tasks, and drilling on part-task practice (Figure 2.11).

Tasks that are intended to teach a specific skill, knowledge, or attitude are referred to as **learning tasks** (black triangle in Figure 2.11). These tasks vary in difficulty and complexity and can be organized from easiest to most difficult. To manage intrinsic cognitive load, these tasks should be organised into levels of increasing complexity. Additionally, to manage extraneous cognitive load,

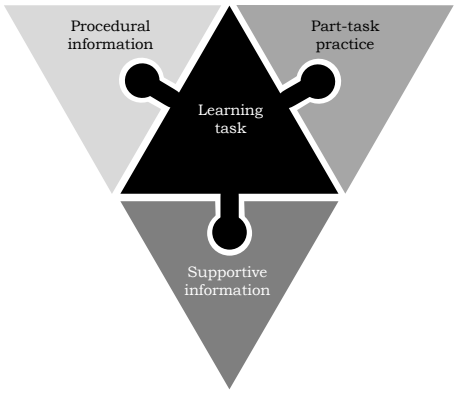


Figure 2.11: Scheme representing the four components of the 4C/ID model [68]: the learning task in the centre (in black), the part-task practice (in medium grey) at the top right, the procedural information (in light grey) at the top left, and the supportive information (in dark grey) at the bottom. The components do not have any specific hierarchy. The learning task is a whole-task experience based on a real-life task, aiming to integrate skills and knowledge; if the learning experience contains more than one learning task, variability should be present, and support and difficulty should be managed inversely proportional. The part-task practice involves the additional practice of routine aspects of the task. Procedural information shows the learner how to perform routine aspects of the task and is presented in a timely manner. Finally, the supportive information component is aimed at supporting the learners with information non-related to the routine aspects (i.e., problem-solving), and it should be available to the learner at all times.

learner support and guidance should gradually decrease as the learners move through the levels, decreasing as they become more comfortable and confident in the tasks until they can perform the learning tasks at a particular level of complexity without support. This creates a cycle where learners gradually move from more support to less and from less complex tasks to more complex ones until they can independently tackle more difficult tasks. To encourage germane load, all learning tasks have a high level of variability. This encourages learners to compare and contrast the tasks with one another [65].

Supportive information (bottom dark-grey triangle in Figure 2.11) is provided to link what learners already know to what is required for a task, helping them perform non-routine aspects such as problem-solving or decision-making. This information is available at the start of a task class and remains accessible throughout.

Meanwhile, **procedural information** (indicated by the light grey triangle in Figure 2.11) is given to help students complete routine aspects, with step-by-step instructions provided when needed. To attain a high level of automaticity, **part-task practice** (indicated by the medium grey triangle in Figure 2.11)

is given in addition to learning tasks consisting of a high volume of repetition for selected recurrent aspects to release processing resources for performing non-recurring aspects.

To design the four components of the 4C/ID model for multimedia environments, different principles have been described by van Merriënboer et al. [66] for each component.

Designing learning tasks

Sequencing principle The sequencing principle states that teaching complex pieces of information is more effective by starting with simple concepts and building up to the full complexity. Studies have shown that this can lead to better transfer test performance. The 4C/ID model is based on this principle and suggests starting with a task class that can be solved based on a simple domain model and progressing through increasingly complex models or learning content [66].

Physical fidelity principle The physical fidelity principle suggests that a high-fidelity task environment can contain irrelevant details that may be detrimental to learning for novice learners. According to the 4C/ID model, training should begin with tasks performed in a low-fidelity environment, which only includes the necessary aspects of the real environment. As learners become more advanced, it becomes necessary to perform the tasks in a high-fidelity or real task environment [66].

Training-wheels principle The training-wheels principle is a way to support novice learners by providing constraints to make sure they cannot perform actions that are not necessary to reach performance goals. Studies have found that it can positively affect learning and task performance, but its effectiveness depends on how it is designed. It is included as one way to decrease guidance for learning tasks within one class in the 4C/ID model [66].

Variability principle The variability principle states that learning tasks should be sufficiently different from each other to construct abstract schemas that enable the transfer of learning. Studies have shown that higher variability across learning tasks leads to higher transfer test performance [66].

Collaboration principle The collaboration principle suggests that complex learning tasks should be assigned to groups rather than individuals to increase learning outcomes. The 4C/ID model incorporates this principle by assigning complex tasks to groups. Group work can also provide scaffolding, which helps learners acquire domain-specific and self-regulated and self-directed learning skills [66].

Designing supportive information

Prior-knowledge activation principle The prior-knowledge activation principle states that subsequent learning is improved when prior knowledge is activated beforehand. This can be done in various ways, such as problem analysis, perspective-taking, and concept mapping. Pictorial representations can also be used, and the learner's prior knowledge influences the effectiveness of these. The 4C/ID model promotes an inductive approach that stimulates pre-existing knowledge, making it easier for learners to comprehend general and abstract ideas. This technique facilitates the creation of links between recently acquired knowledge and what is already known, leading to increased memory retention and understanding [66].

Multimedia principle The multimedia principle suggests that learning is enhanced when text is accompanied by visuals [67] but only if they are relevant to the task [69]. The different cognitive processes that are induced lead to a more comprehensive understanding of the content. This principle has been researched and supported; however, its effectiveness depends on factors such as the learning content, type of graphics (e.g., instructive, seductive or decorative [69]), and type of task (e.g., conceptual, procedural, or casual [70]). The 4C/ID model acknowledges the importance of the multimedia principle in schema construction and elaboration and, therefore, typically suggests presenting supportive information in a multimedia format [66].

Dynamic visualizations principle Dynamic visualizations (animations, videos, etc.) that illustrate processes and mechanisms that change over time can positively impact learning and transfer when they are designed accordingly and depict movements. The effect depends on one's prior knowledge; those with lower prior knowledge learn better from static visualizations, while those with higher prior knowledge benefit more from dynamic visualizations. The 4C/ID model stresses the use of dynamic visualizations, such as videos, to illustrate task performance. This is especially pertinent when demonstrating locomotor

behaviour, which provides helpful examples of how the task is meant to be performed [66].

Redundancy principle The redundancy principle suggests that presenting the same information multiple times hampers learning. However, in some cases, redundant information can be beneficial if it directs learners' attention to important information without causing an extra cognitive load. The 4C/ID model explains that information should be presented supplementary to build on previous knowledge but not repeat the same information [66].

Designing procedural information

Modality principle The modality principle suggests that dual-mode presentations (incorporating audio and visual elements) are more effective than single-mode presentations (using only visual information) when teaching. In the 4C/ID model, procedural information is recommended to be spoken by a teacher or other pedagogical agent if the learning task contains visual elements. In contrast, visual just-in-time instructions may be better for tasks without visual elements [66].

Temporal split-attention principle The temporal split-attention principle states that learning from related sources is more effective when they are presented together rather than separated in time. In the 4C/ID model, this principle is particularly important for presenting procedural information for performing tasks. Simultaneous presentation of this information and the task itself has been shown to positively affect transfer test performance [66].

Spatial split-attention principle The spatial split-attention principle suggests that students who need to learn information linked to visuals, such as pictures or animations, will benefit from having the text physically integrated into the visual. In the 4C/ID model, it is beneficial to present procedural information in a way that is optimally integrated with the learning tasks and the task environment [66].

Signalling principle The signalling principle suggests that learning can be improved when a learner focuses on the key elements of the task or presented information. This reduces the need to search for information, frees cognitive resources and can lead to better transfer test performance. The 4C/ID

model states that signalling is especially useful when dealing with procedural information and routine aspects of a task [66].

Segmentation principle The segmentation principle states that breaking up dynamic visualizations (animations, videos, etc.) into meaningful segments helps with learning and transfer. This is especially helpful for learners with low prior knowledge and works best when the learners control the segmentation. The 4C/ID model uses the segmentation principle to present just-in-time information step-by-step, thus avoiding overloading the learner with transient information [66].

Designing part-task practice

Component-fluency principle The component-fluency principle suggests that repeating and practising one or more parts of a task can positively affect learning and performing the entire task. This is because by strengthening the routine aspects of a task and making them automatic, cognitive resources can be freed up and allocated to the problem-solving aspects of the whole task. The 4C/ID model advises only applying part-task practice on particular routine aspects after learners have been exposed to them in the context of the whole task. This allows them to understand how the routine parts can be integrated into the whole [66].

Limitations of the 4C/ID model The 4C/ID model has some limitations. While it can be utilized to create multimedia learning, it does not provide instructions on which media to select based on restrictions, task requirements, or target group characteristics. Instead, it concentrates on analysis and design rather than the multimedia environment's development, production, implementation, and evaluation. Furthermore, although certain principles are discussed for one of the four components, each of the same principles may be applied to various components [66].

The AR prototype described in this thesis uses the 4C/ID model to design learning tasks across different application levels. Specifically, the four components of the model and the design principles outlined in this section are employed on each level to facilitate schema formation and the automation of the titration experiment. However, not all the design principles are explicitly explained in this section. Further chapters will discuss the specific principles used to design the application.

2.3.6 Inquired based learning

How to create a learning experience that helps learners understand a complex subject?

Inquiry-based learning is a teaching method in which students learn by actively participating in discovering new knowledge. This approach is based on the constructivist perspective, which emphasizes the role of the learner in constructing their understanding of the world. In inquiry-based learning, students engage in a self-directed process of discovering new causal relationships by formulating hypotheses and testing them through experiments or observations. This approach is particularly effective in science education, allowing students to engage with the complex scientific process through smaller, logical learning units [71].

There are various frameworks for understanding the inquiry-based learning process, but one popular model is the five-phase model proposed by Pedaste et al. [71]. In this model, the phases of inquiry-based learning are orientation, conceptualization, investigation, conclusion, and discussion. In the orientation phase, students are introduced to a problem or question, stimulating their curiosity. During the conceptualization phase, they generate research questions or hypotheses. In the investigation phase, they actively seek answers to these questions by exploring, observing, designing experiments, and interpreting data. In the conclusion phase, they address the research questions and potentially gain new insights. Finally, in the discussion phase, students share and reflect on their findings with others. It is worth noting that the discussion phase may not be necessary for all inquiry-based learning cycles, but it can be beneficial for improving the quality of learning outcomes.

Pedaste et al. also studied the inquiry-based cycle for AR learning environments finding that AR affordances are mainly applied in the conceptualization phase and, to a lesser extent, in the investigation phase [72]. In this thesis, the model proposed by Pedaste et al. [71] is used to design the learning environment precisely as the blueprint of each level or learning task.

2.3.7 Embodied Cognition

Classical learning theories, such as cognitivism, account for the role of the mind and environment in learning. However, new approaches, such as embodied cognition, extend the understanding of the learning process by studying the relationship between the mind, environment and body. Researchers in this area

proposed that "an organism's body is, in some sense, integrated into cognitive processing". However, the mechanisms of how this occurs remain unclear [73].

According to embodied cognition research, the body plays an important role in the learning process through several proposed mechanisms, such as creating analogies between physical and abstract concepts, using gestures to link and abstract, enhancing cognitive skills, offloading cognitive processes into the body or environment, and interpreting visual representations [74].

Research in immersive technologies such as VR and AR has particularly been interested in embodied cognition due to their ability to engage the user in the learning process. AR, in particular, provides users with spatial experiences that allow them to move around 3D simulations and gain new perspectives by rotating or walking around the object. This allows them to store information in the physical world, leading to better learning outcomes [75]. Additionally, AR environments enable users to make gestures in relation to re-enacting 3D simulated phenomena, which can represent abstract concepts in physical form. According to Radu et al., [76], this can strengthen learning.

In line with embodied cognition, using AR as the medium to convey learning content is chosen in this work in order to enable the exploration and physical reenactment of typical actions performed in a real laboratory.

2.4 Common Perspectives of Learning Theories

The different theories presented in this chapter share common views on learning and its construction. Experiential Learning [61] explains learning as a process in which learners (re)construct and transform their knowledge as they acquire more experiences. It also emphasizes the understanding of learning as a holistic experience which accounts for different human behaviour (thinking, feeling, perceiving etc.). This is in line with the 4C/ID model [68, 66], which calls for "complex learning" - an integration of knowledge, skills, and attitudes, as well as what Novak proposed as necessary to yield meaningful learning - the integration of different domains (i.e., cognitive, affective, and psychomotor). However, this integration is up to the learner, so teachers must provide a range of experiences to choose from that can support students in their knowledge construction [2].

In science education, this process can be difficult due to misconceptions and the complexity of constructing interrelated knowledge frameworks [2]. To help in this endeavour, teaching methods such as inquiry-based cycles can be used. This approach allows students to engage with the scientific process through

smaller, logical units [71] while still maintaining a holistic view of performing a complete experiment.

Further, understanding how people store and manage information in the brain can be used to decide how to design learning experiences - as demonstrated by Cognitive Load Theory (CLT), which explains the mental load learner experiences depending on factors such as prior experience, task, and instructional material [65]. Furthermore, the Cognitive Theory of Multimedia Learning (CTML), proposed by Mayer [67], provides an explanation of how people process information acquired through multimedia as well as offers prescriptive advice regarding what multimedia features should be included to aim for better learning. The 4C/ID model relies on the principles of multimedia learning to provide a (holistic) model for complex learning that supports the integration of knowledge, skills, and transfer of learning to future contexts. The holistic design this model emphasizes helps to address such problems as compartmentalization (focusing on one particular domain of learning), fragmentation (breaking content into parts), and the transfer paradox (when efficiency is preferred over transfer), which challenge educational practice [68, 66].

Altogether, these theories present a common perspective of the learning process, emphasizing the need for a holistic approach to education and for educational designs that promote the integration of multiple domains and the transfer of learning to novel situations.

2.5 Conclusion

AR is gaining attention as a possible resource for providing learners with a scalable and accessible virtual learning environment. As more and more software and hardware become available, the deployment of AR is becoming more accessible. Mobile augmented reality is particularly attractive in education due to its accessibility. The potential learning benefits of AR are being explored, and its environment design is becoming more common. Established instructional frameworks such as the 4C/ID model support these new endeavours for designers. Designing AR environments for teaching scientific concepts pose additional challenges because of the abstract nature of the subject matter, requiring a holistic approach that combines theory, experimentation, and real-life examples. Constructivist approaches to learning offer a methodological way to aid students in meaningfully engaging with the learning content. In this thesis, the 4C/ID model is used as a prescription to design an AR environment deployed on handheld devices. The inquiry-based cycle is used to guide the design of the specific activities of each level. The goal of the experience is for learners

to interact with their context, creating meaning for the scientific concept of titration. Additionally, the user can improve their learning by offloading the cognitive processes onto the physical environment through enacting physical actions.

Chapter 3

Mobile Augmented Reality Apps in Education

This chapter is adapted from: J. L. Domínguez Alfaro and P. Van Puyvelde, "Mobile Augmented Reality Apps in Education: Exploring the User Experience Through Large-Scale Public Reviews," in Lecture Notes in Computer Science, vol. 12980, L. T. De Paolis, P. Arpaia, and P. Bourdot, Eds. Cham: Springer International Publishing, 2021, pp. 428-450.

3.1 Introduction

With the advances in mobile computing, Augmented Reality (AR) capability has been incorporated in wireless devices such as smartphones and tablets [77], leading to the emergence of Mobile Augmented Reality (MAR) [20]. This technology has been deployed across various applications and sectors, e.g. entertainment, industry, gaming, tourism, and education [78].

Several systematic reviews have examined AR's potential in terms of its usability, advantages and disadvantages in educational settings [22, 19, 23, 79, 80]. Reportedly, AR applications are generally well-received by users, although technical features and the risk-benefit ratio of these apps affects user satisfaction and acceptance [81, 20, 82, 83]. However, user experience within public apps specifically designed for educational purposes remains largely unaddressed.

This chapter presents the analysis of educational MAR applications from the Google Play Store, extracted on April 2020, and synthesises the user reviews into insights for improvement. Due to the number of applications and user reviews, the analysis is a challenging task. Therefore, an automatic tool derived from data mining techniques is used to aid in qualitative and quantitative analysis.

In addition to user reviews, this research analyses some of the indicators that users consider before downloading an app, such as ratings, number of installs, release dates, app sizes, and AR type. Since the user experience starts from the moment the user selects and downloads an application [84], the study of such metrics can provide valuable insights.

Additionally, the levels and fields of education are analyzed using the International Standard Classification of Education (ISCED) [85] used by other researchers [19, 27]. This serves to identify the educational sector that an AR application targets.

In this chapter, the following research questions are explored in detail:

- RQ1: What are the main problems, feature demands, and opinions expressed by the users of educational MAR apps published in the Google Play Store?
- RQ2: What are the main characteristics of price, size, rating score, installs, release dates, and AR types of educational MAR apps?
- RQ3: Which are the levels and fields of education that educational MAR apps published in the Google Play Store target?

3.1.1 Background and Related Work

Despite the advancement of AR technology, there are still limitations to its use in education. These limitations are related to teaching and learning approaches and technical constraints. Some researchers have emphasized the importance of designing user-friendly experiences to overcome these challenges and successfully integrate AR into educational settings [49, 27]. This chapter aims to explore the user experience of AR in education by analyzing the technological challenges reported by users of public educational AR apps. This information can be useful for instructional designers to improve the user experience of their AR systems.

3.1.2 User Experience of MAR Applications

According to ISO 9241-210, user experience (UX) is defined as the "person's perceptions and responses resulting from the use and anticipated use of a product, system or service" [86]. Several studies have explored the UX of mobile augmented reality (MAR) applications and have concluded that participants value useful content, ease of use, and absence of technical difficulties [81, 83]. Additionally, Dirin and Laine [81] observed that emotional perception could also influence the experience, with people reporting positive emotions such as encouragement, excitement, and interest. Moreover, Li et al. [82] found that users' satisfaction can be a motivation to continue or discontinue using MAR applications. In the context of science education, Salar et al. [87] reported that the usability of AR environments positively influences students' interest. This work builds upon previous researchers' efforts by analysing the user experience of MAR users within the context of educational apps.

3.1.3 User Reviews and App Data as a Source of Information

Application distribution platforms or app stores contain rich information that can help software engineers and designers during the design process [88, 89]. For example, they have public reviews reflecting users' experience and app metadata such as price, number of installs, release dates, rating, app size, etc., which provides an overview of the app characteristics.

User review analysis can be used to quantitatively summarize the topics contained in the reviews and to gain insights from a qualitative examination of the content [90]. Nicholas et al. [91] used the user reviews of 48 apps for bipolar disorder to analyze qualitatively the user perspectives of these apps. Similarly, Alqahtani and Orji [92] analyzed reviews of mental health apps to reveal what users of such apps like and dislike.

However, using app reviews as sources of information comes with some challenges. App reviews can be short, unstructured, and grammatically incorrect. Depending on the app, they can be abundant, making it a complex and time-consuming process to extract information from them [88, 93].

Regardless of these challenges, compared to other resources such as focus groups or surveys, the reviews provide direct and numerous feedback for the stakeholders involved in the app development [94]. Moreover, the topics included in such reviews contain relevant information from the user experience, such as the ideas, the needs, and the experience with a certain app that may potentially inform scholars interested in understanding the context of a particular app [89].

In this chapter, user reviews are utilized as a source of information (e.g., identify problems of existing apps, features requested from the users, etc.). Unlike other scholars, we used an automatic and public tool developed by Panichella et al. [93] to classify the content of the reviews into relevant categories that allow the detailed analysis of the requests and information procedurally with low processing time.

3.2 Methodology

The general data processing methodology comprises three phases: app extraction, review extraction, and review classification (Figure 3.1). The Google Play Store was chosen as the data source. This decision was made due to a relatively larger amount of apps in Google's app store compared to the Apple Store [1].

3.2.1 App Extraction

The first step was to extract a list of apps from Google Play using an open-source scraping tool called google-play-scraper (version 7.1.2) [2]. The "search" function provided by this library allows the extraction of the app identifiers corresponding to the results from a search in the Google Play store using any term. The search terms used are: "augmented reality education", "augmented reality education apps", "augmented reality educational apps", "ar education", "ar education app", and "ar application for education". These were chosen by using a keyword tool [3]

For each search term, a maximum of 250 apps was retrieved, which gave 1500 apps from all search terms. Of these 1500 apps, only 440 were unique, i.e. not repeated in the list. The metadata from each app was retrieved using the "app" function from the scraping tool mentioned earlier from the unique list.

This function extracts the metadata of each application (A.1) and uses language and country as parameters for the extraction. The parameters "English" and "Belgium" were used to ensure that the apps were available for download in the country where the research occurred.

Finally, an additional cleaning step to remove non-AR apps was performed by assuming that AR apps must contain the keywords "ar" or "augmented

[1] https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/, accessed: 11-08-2020

[2] https://github.com/facundoolano/google-play-scraper#search

[3] https://keywordtool.io/app-store

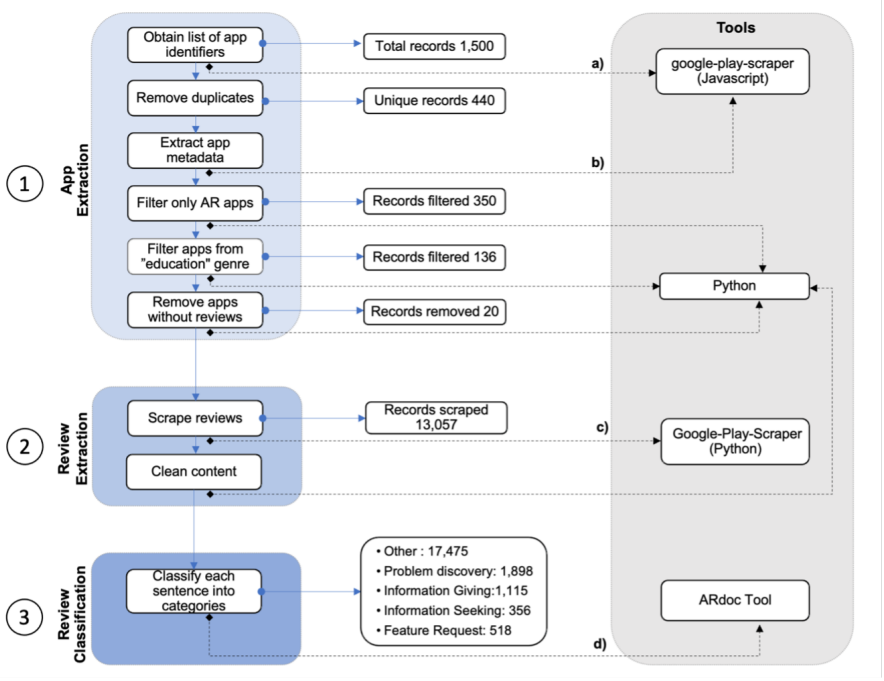


Figure 3.1: The General methodology for data processing. The main phases are 1) app extraction, 2) review extraction, and 3) review classification. This figure summarizes the steps followed in each phase, the results, and the tools used. a) "search" function[2], b) "app" function[2], c) "reviews" function[4] and d) automatic data mining tool [93].

reality" in their title or description. Applications that did not have any of these non-case-sensitive keywords in the title or the description were removed from the database. As a result, the data set was reduced to 350 apps.

3.2.2 Review Extraction

From the previously cleaned app's database, a subset was created from apps that fall under the app store "Education" class; only 136 apps satisfied this criterion. The user reviews were mined using a different open-source scraping tool called Google-Play-Scraper (version 0.0.2.2) [4] coded in Python. The reviews from each app were extracted using the function "reviews". In total, 13,057 single reviews were obtained from 116 apps. It is important to mention that even though 136 apps were mined, only 116 apps contained reviews.

[4] https://github.com/JoMingyu/google-play-scraper

3.2.3 Review Classification

To automatically classify the content of the reviews, the App Reviews Development Oriented Classifier (ARdoc) developed by Panichella et al. [93] was used. This tool combines three feature extraction techniques: Natural Language Processing (NLP), Text Analysis (TA), and Sentiment Analysis (SA). Using a machine learning (ML) algorithm, ARdoc automatically classifies the sentences of the reviews into a predefined taxonomy. The taxonomy includes five categories: information giving, information seeking, feature request, problem discovery, and others (Appendix A.1). The taxonomy emerged from previous research, which combined common topics from user reviews and comments given by developers [95, 93].

In this study, the Java API version of the ARdoc tool was used. In addition, the feature combination of text structures and sentiment analysis ("NLP+SA") was employed since, according to Panichella et al. [93], this combination of features allows a higher classification precision, recall, and F-measure (all ranging between 84-89%). The architecture of the ARdoc tool is described in Appendix A.1.

In total, 21,362 sentences from 13,057 reviews of 116 apps were classified. However, it is important to note that most reviews contain more than one sentence; therefore, the number of reviews is smaller than the number of sentences.

The extraction of informative reviews was performed to analyze the content of each sentence and derive conclusions. For this purpose, it was assumed that informative reviews are those which contain sentences classified into "problem discovery", "information giving", "information seeking", and "feature request". As a result, the filtering step consisted of removing reviews with sentences that had all been classified as "other" by the ARdoc tool; the total informative reviews were 3,065. It is important to note that not all the apps contained informative reviews; therefore, the reviews used for further analysis only came from 86 apps. The complete data set and classification can be found in the Supplementary Material of [96].

The informative reviews were inspected manually to derive more specific and detailed conclusions. To ensure an impartial assessment of all apps, the application "Sky View Lite" was deliberately excluded due to the large number of reviews (1,313). In total, 1,752 reviews were further analyzed (Table 3.1).

It is important to mention that all reviews were divided manually into mutually exclusive groups corresponding to the app categories from ARdoc. Therefore, most reviews contained more than one sentence that could belong to different

categories. If a review had sentences classified into different classes, the priority was given based on the review context. In addition, another class, denominated "Irrelevant", was added to reviews with vague or unclear statements.

To evaluate the reliability of the manual classification, as suggested by [97], the intracoder reliability was calculated at two different times during the research; the second test was done after two months for the mutually exclusive categories. Cohen's Kappa (k) values were obtained, and according to the interpretation of the results, an almost perfect agreement was observed for feature request (k=0.94) and information seeking (k=0.96). While substantial intracoder agreement was found for problem discovery (k=0.77), information giving (k=0.68) and Irrelevant (k=0.74). Additionally, the percentage of agreement between the mutually exclusive category manually assigned and the classification from ARdoc was obtained to be 76% (the manual classification is considered an agreement if at least one of the sentences in the review classified by ARdoc has the same category).

Finally, each informative review was manually analyzed and assigned to a broad topic. These topics were derived from those defined by Pagano and Maalej [89] and subsequently used by Panichella et al. [95] to define the taxonomy of ARdoc [93]. If the content of the review was not related to any of these topics, a new broad topic was created.

3.2.4 App Characteristics Extraction

All metadata of each application obtained in Section 3.2.1 was formatted in a table for further analysis (see Supplementary Material [96]).

3.2.5 Classification under Fields and Levels of Education

To classify the apps according to the levels of education, the International Standard Classification of Education (ISCED) [85] was used in its aggregated level (A.2). The app developers' description was manually examined to identify the level and field of education targeted by the app. Each app can be classified into several levels. If the target level was not explicitly expressed, the app was assigned, considering other features such as images, videos, and educational content reported in the description. The apps were labelled as "Not Classified" if this information did not provide clear indications. On the other hand, if the fields of education were not explicitly expressed, the apps were classified as "N/A".

3.3 Results and Discussion

This section summarizes the findings related to the user reviews and their classification using the ARdoc tool, the AR apps' characteristics obtained from each application's metadata and the survey of the educational levels and fields.

3.3.1 What are the main problems, feature demands, and opinions expressed by the users of educational MAR apps published in the Google Play Store?

As described in Section 3.2.3, the reviews extracted from 116 educational apps were classified by the ARdoc tool. As a result, 1,752 reviews were analyzed.

Table 3.1: Classification of informative reviews as described in Section 3.3. This classification contains the number of reviews per category, the percentage in relation to the total of reviews, the average app score rating per category, and the number of apps in each category

Category	Number of reviews	Percentage of total reviews ($N = 1,752$)	Average score per category	Number of apps
Problem Discovery	1,142	65%	1.6	70
Information Giving	317	18%	3.8	56
Feature Request	135	8%	3.9	29
Information Seeking	125	7%	3.0	38
Irrelevant	33	2%	2.9	17

Table 3.1 shows that the largest number of reviews correspond to problem discovery (1,142/1,752). These reviews come from 70 apps, meaning that from the total of apps with informative reviews, 81% (70/86) of apps have reviews that report problems. Additionally, according to these reviews, the average score rating of the app is 1.6 out of 5 stars, which is the lowest among the categories.

Problem Discovery

Six broad topics were identified in the "problem discovery" category. Specifically, it was found that users have problems with the app's functionality, specific features, general service, hardware, software, and content (Figure 3.2).

In this category, most problems with AR apps were related to functionality. Although some reviews were not descriptive, for example, some users provided

reviews that indicated a problem, such as "app is not working" or "app did not work". The high number of reviews on this topic suggests that some users cannot install AR apps or even use basic functionality. For example, commenting on AR functionality, one of the users said, "only concern is that ar does not work it is [*sic*] just freezes or shows a black screen" (from the Physics Lab app).

Hardware-related problems were found on 14 apps. The most common problems related to hardware are camera autofocus, battery consumption, and phone overheating. Users from different apps expressed these observations. For example, commenting on phone overheating, one user said: "It's pretty good and awesome for children even for adult it's fun! but it makes my phone overheat and shutdown always ..." (from the Animal 4D+ app). One user commented about battery consumption: "Expeditions is dramatically draining my phone battery. I am wondering why my phone is dead. Looked at history and found that Expeditions was responsible for using up 40% of my battery even though I have not used it in several days..." (from the Expeditions app).

Software-related problems were classified as those reviews in which users explicitly express the appearance of an error message, issues with the mobile data, or software compatibility. For example, one user stated: "tried in 3 devices. failed to initialize Vuforia" (from Human Anatomy AR app).

Another recurrent problem expressed by users was content-related. Either the AR content was unavailable, it was not working, or its access was absent. For example, a user said: "This is not a good app the animations are really too short, and, in some topics, the audio is not available..." (from the Cambridge Explore app).

Thus, from the analysis of the "problem discovery" category, users of educational MAR apps experience diverse issues related to several factors. However, the technical hurdles related to the AR technology, such as power consumption, AR functionality, and real-time interaction or unavailability of content, must be a high priority to developers as these problems will cause users to stop using the applications [82, 83]. For example, power consumption can be solved through energy-saving techniques, reducing power consumption from GPU, and using active cooling methods to control the temperature of handheld devices [98]. To improve AR functionality, developers can leverage improved tracking methods. For instance, hybrid tracking technology tackles the computational inefficiency of handheld devices and network performance [99].

Finally, web services and the new implementation of emerging 5G networks can help to reduce the issues related to content and real-time interaction by improving the data transmission rate [100]. As mobile device computational efficiency tends to be limited, especially in low-end devices, apps that utilize

web services have potential to deliver an enhanced experience. Rather than requiring the user to download an app to their smartphone, content can be accessed directly through a web browser.

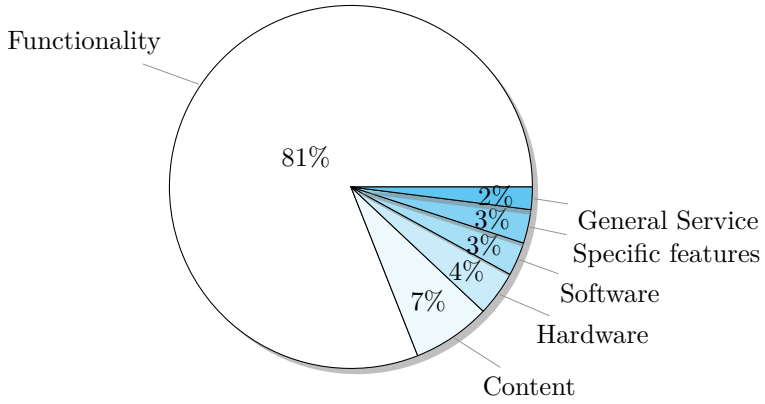


Figure 3.2: Identified broad topics on the problem discovery category and its percentage relative to the total of reviews in this category (1,142). The topics cover issues related to functionality (81%), including installation, interaction, and scanning issues; Content (7%) covers unavailable, limited, or locked multimedia; Hardware (4%) includes the sensor, power consumption, and camera issues; Software (3%) covers error messages, mobile data, and compatibility; Specific features (3%) covers exclusive features like speech recognition and the user interface; General service (2%) covers registration, payment, and privacy.

Information Giving

The reviews in this category account for 18% (317/1,752, see Table 3.1) of the total informative reviews from 45 apps. According to these reviews, the average score rating is 3.8 out of 5 stars.

Five broad topics were identified from the content analysis of this group of reviews. Some of these topics were also identified by [89, 95]. The most common issues are praise, dispraise, helpfulness, shortcoming, and other feedback (Figure 3.3).

The following reviews illustrate some examples of praises found in the data set. For example, one user of the AR Mondly app stated: "unique ar experience! mondly ar proves augmented reality can be the future of language learning!". Another user from the Expeditions app commented: "I love using this app with my students. I used a google expeditions kit from @aquilaeducation, and my students absolutely loved traveling all over the world!".

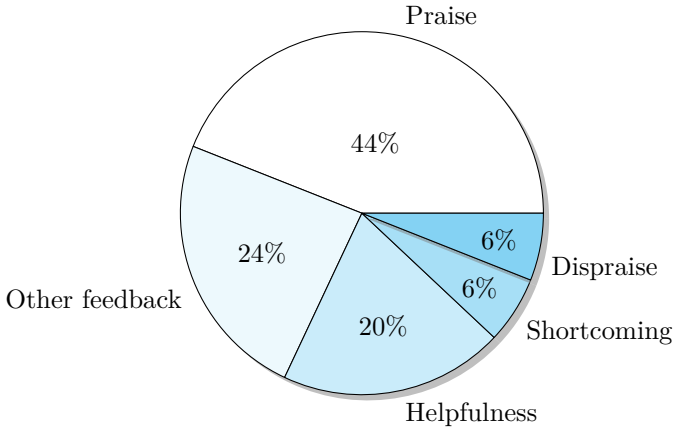


Figure 3.3: Identified broad topics within the "information giving" category and its percentage relative to the total of reviews in this category (317 reviews); these topics include praise (44%), dispraise (6%), helpfulness (20%) (e.g., reviews expressing how the app has been beneficial), shortcomings (6%) (e.g., reviews citing payment, privacy, or instruction issues), and other feedback (24%) (e.g., reviews expressing other reactions from users).

On the other hand, users also expressed their discontent. It is important to note that most of the criticisms made by users mostly come together with a "problem discovery". Therefore it was not usual to find dispraise sentences alone but combined with a specific issue or a dissuasion statement. For example, one user expressed: "I didn't even liked [*sic*] to give one star it is not a good app don't download it I hate it" (from Space 4D+ app). Another example of a review is "the worst app ever. waste of time. do not download this app" (from Physics Lab app).

Another common topic in the "information giving" category was "helpfulness". These reviews were generally classified as such when a particular application was practical or helpful. Of the total of apps, 29 have reviews related to this topic. For example, one user commented on the learning experience: "A great app to learn and practice chemistry for grade 10 students! You should try using it" (from the Dat Thin Pone app). And another user stated: "Thanks a lot for the app its [*sic*] awesome and very helpful for the students who cannot afford physical lab" (from the Physics Lab app).

Some users provide information about concrete aspects of apps they are unsatisfied with. These types of reviews were classified under the "shortcoming" topic. Even though shortcomings seemed similar to problems classified under the "problem discovery category", they differ in the intention expressed by the

user. While a problem is a fact that the user experience, the shortcoming derives from subjective opinion. For example, users unsatisfied with markers expressed, "maybe works but need to print markers. no thanks! only free trial" (from the Predators AR app). This review not necessarily implies that the marker does not work but only that the user is not happy with printing it. Similarly, another user said, "please provide instructions. ok, I downloaded the file and yes it works and works well. but having to have the target image is not practical" (from the Augmented Reality Solar System app).

The most common shortcomings include payment options, privacy concerns, and instructions provided by developers. For example, a Quiver 3D coloring app user shared: "I really like your app, but I wish you did not have to buy everything separately, and you could buy a membership or whatever. I understand that you have to make money, but I would also like if there were more free pages. thanks!". Regarding privacy concerns, a user said: "it requests to spy on you", and another expressed: "the app wants to access my contacts and to make and manage phone calls. This is a coloring app. I do not see why it needs the ability to make calls..." (from the Quiver 3D coloring app). Finally, related to the app instructions, a user from the Physics lab app expressed, "It is alright if you understand physics, but for [*sic*] people for do not understand it needs more instructions for people who don't understand. the instructions that you currently have a very hard to understand".

Overall, the results indicate that most users tend to write reviews when they are extremely satisfied or highly dissatisfied (praise or dispraise), but they do not provide specific insights unless the user gives exact details. In addition, users share the helpfulness of certain apps and some shortcomings that they experienced, such as lack of instructions, dislike towards markers, payment issues, and privacy concerns.

Considering these observations, developers can improve the user experience by explaining why the AR systems require specific permissions. This information can be provided before and during the use of each application. Although not all users may understand the justification of the system requirements, providing the data will allow them to choose with an informed view [101].

Users also provide information about the real-life scenarios in which MAR educational apps were found helpful. According to these findings, MAR apps can be useful when students cannot be in a particular place, e.g., a classroom, laboratory, or historical place. Therefore, developers and practitioners can concentrate on developing experiences for users that cannot be physically present in the same learning environment of their interest. Nowadays, this will not be hard to realize after the recent pandemic, where these scenarios are more "normal"[102].

Feature Request

The "feature request" category accounts only for 8% (135/1,752) of the reviews analyzed coming from 29 apps (Table 3.1). Interestingly, the users providing these reviews gave a higher rating score (3.9 stars) than users sharing other reviews. Three types of requests were identified among this group of reviews: 1) content, 2) features, and 3) general improvements.

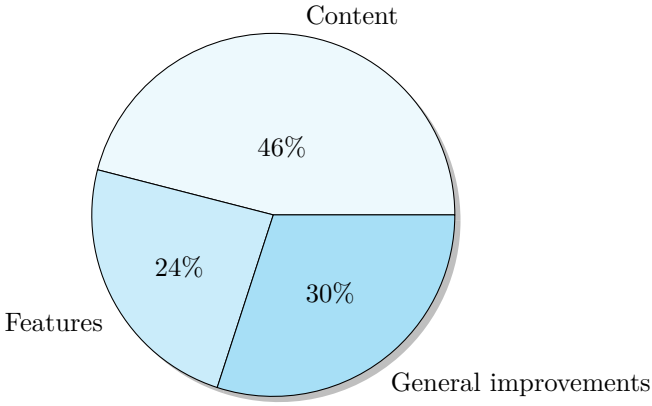


Figure 3.4: The "feature request" category of reviews contained 135 reviews in total, with 46% of the reviews requesting new content, 24% requesting additional features, and 30% requesting general improvements.

The content request was one of the main demands of users. A user stated: "Could you please add more information in de [*sic*] app when you peel off the layers? (...) students will learn even more about it. (...) Nicely done!" (from the Human Eye- Augmented Reality app).

Commenting on adding features, one user said, "fun app. (...) for kids and adults alike (...) their comments are: would be good if we could draw our own pictures! (...) maybe for a future update?" (from the Quiver 3D coloring app). Finally, concerning the request for improvements, one user stated: "The first app to use ar core and does it in a brilliant way. [*sic*] would be great if you could expand the app to include placement options (placing on a surface) and have the atoms react to light" (from the Atom Visualizer for ARCore app).

The current request of users of AR apps may be partly explained by the fact that the developers are exploring the technology; thus, they might be showing only some examples showcasing how the technology/app can be used but not necessarily offering a complete experience. Therefore, future developments

may dive into further educational content integration while maintaining good design, which can be achieved through a multidisciplinary collaboration of scholars. For instance, app developers can leverage multimedia instructional design techniques to enhance learning, and practitioners can validate students' content and learning outcomes through evaluation studies.

Information Seeking

The "information seeking" category is defined by sentences expressed by users attempting to obtain information [93]. One hundred twenty-five reviews from 38 apps were classified under this category, accounting for 7% of the sum of informative reviews. The average rating score provided by users in this category was 3.0 out of 5 stars (Table 3.1).

Among the reviews in this group, the most popular questions were regarding how to use and how or where to obtain content/markers. Regarding the question of "how to use?" a user expressed: "do not understand how to use it (...) the screen say [*sic*] scan the flat surface. What flat surface? Do I need one of those printed qr cards or what someone help [*sic*] me" (from the Apollo's Moon Shoot AR app). Content-related, one user asked, "From where I can get the image?" (from the AR Dino Roar app).

Other questions were related to AR elements. For example, a user asked, "What is an AR marker" (from the iBugs AR app). Another one said, "I do not have an ar maker (...) why is it that you need a [*sic*] AR maker. Improvment [*sic*] I want for my child that you do not need a [*sic*] AR maker." (from the Predators AR app).

Although less common, privacy concerns were also found under this category. This topic was identified under "information giving"; however, some users express this concern as questions rather than statements. For example, one user said: "why do I need to give permissions to make calls?" (from the AR animals app), and another stated: "a great way to make your pictures come alive and an underused technology. My kids (and me) love it. but why is there a permission to know my precise location?" (from the Quiver 3D coloring app).

In summary, these results indicate that the most frequent queries of MAR app users are related to the application's purpose and usage. It appears that some users may not understand how to use the app, what a marker is, or where to obtain content. This lack of understanding could be attributed to either developers providing inadequate information about the app or users not being familiar enough with AR. To address this, developers should consider including

clear instructions, intuitive interactions, or familiar cues in the system to help users become more familiar with the technology.

3.3.2 What are the main characteristics of price, size, rating score, installs, release dates, and AR types of educational MAR apps?

The metadata extracted in Section 3.2.1 was used to answer this question. The data includes price, app size, ratings, number of installs, and release dates.

Price

Regarding price, most apps were reported as "free" except for the app called "Quiver Education". This result suggests that most users can access AR apps for free. However, from the data extracted, it was impossible to determine whether extra paid content was available to the user after downloading.

This is a positive result for educational purposes as students would have access to the AR content. Even if the "in-app purchases" are present, the students may have a higher chance to access the content if the practitioners or institutions can assess the content for free and provide students with the resources.

App Size

Regarding the app size, which is the amount of memory in megabytes (MB) each app occupies in the phone, only 105 apps declared this data. The other 11 apps contained the legend "varies with device". This information, as noted by D'Heureuse et al. [103], does not include the "in-app downloads", which refers to additional data (e.g., music, videos, 3D models).

The estimated average app size is 51.6 MB, with a variance of 25.6 MB. This result is larger than "all apps" in the Google Play Store, which was reported to be 15MB in 2017 [5]. A possible explanation for this result may be due to multimedia content such as video, images, and 3D models. These common resources in regular apps account for most of their size, which may suggest that for AR apps, this also applies. Generally, when developers optimize apps, the size is reduced. However, in the case of MAR educational apps, the optimization process may not be a priority due to the novelty of the technology.

[5] https://sweetpricing.com/blog/2017/02/average-app-file-size/, accessed: 15-06-2020

The app size is a factor developers consider since it can be related to how often an app is installed [104] or, as Mahmood [105] suggested, it may impact how users rate the app. For users, this information is vital to assess how much memory their device needs to run AR apps. Furthermore, app size can also be relevant for scholars and practitioners looking for the successful use of AR apps among their audience. The use of cloud services and Web-based AR applications may overcome this issue. These approaches can be suitable since users will not have to download any smartphone application, and the digital content can be displayed in the cloud. However, these applications are still scarce and experience other challenges related to mobile networks [100].

Rating Score

Regarding ratings, Google Play Store employs a five-star system, ranging from 1 star (lowest) to 5 stars (highest), based on the perceived quality of the app as reported by its users. The average rating of all apps ($N = 116$) is 3.2 stars with a standard deviation of 1.5 stars. According to the stats reported in AppBrain[6], the average rating for all apps in the Google Play store is 4.1 stars. Therefore, the educational MAR apps from the data set are, on average, rated lower than other apps. It is difficult to explain this result since the rating can be associated with multiple factors. According to Tian et al., [106], the app's size mainly influences the app's rating, the number of promotional images on the app store, and the target software version. In general, this finding reflects that the user experience of educational MAR apps is not optimal as compared to conventional apps.

Number of Installs

In the case of the number of installs in the Google Play Store, they are expressed in terms of ranges rather than exact numbers. The ranges are defined as follows: (0-5), (5-10), (10-50), (50-100), (100-500), (500-1000), etc. The install range was compared with the weighted average score rating to explore whether the rating correlates with how much an app is downloaded. Figure 3.5 shows the relationship between the weighted average score versus the number of installs or download ranges.

As shown in Figure 3.5, apps with a small number of installs (e.g., 50-100) have a higher rating score than those with more downloads; this score drops as the number of installs increases until the range of 10,000-50,000. At this point, the app score appears to increase again. This behaviour can be explained by the

[6] https://www.appbrain.com/stats, accessed: 15-06-2020

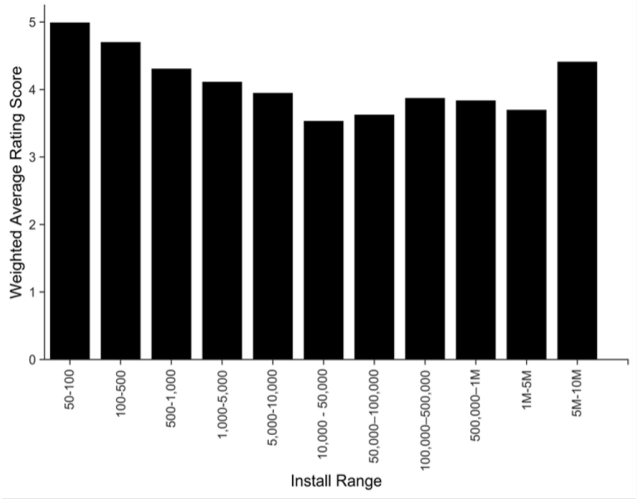


Figure 3.5: Weighted average score per install range

hypothesis proposed by D'Heureuse [103]. Apps with few downloads tend to have a rating score heavily influenced by a few users (probably from people known by the developer). In contrast, larger install ranges (more than 10,000 downloads) receive feedback from more users, and the rating scores tend to increase, which reflects user satisfaction and increased popularity.

The number of apps in the large install range (1M-5M and 5M-10M) is very low for educational MAR apps. For example, only four apps in the 1M-5M range were found (Animal 4D, Expeditions, Devar, and Quiver), and only one (Skyview) was in the 5M-10M range. A possible explanation for this might be that the novelty of AR among users and developers is still present. Likewise, the use of the technology as an educational tool has not reached its maturity, thus leading to limited use, disinterest in installing the apps, or dissatisfaction among the users. Additionally, considering that the number of installs can be related to the success of the apps [103], these findings suggest that educational MAR apps are still not very successful.

Release Dates

The release dates were also analyzed to explore the evolution of MAR in the public market. According to Figure 3.6, the apps in the data set have been released since 2013, having a higher number of releases between 2017 and 2019. Since 2017, Google LLC and Apple Inc. have released their application

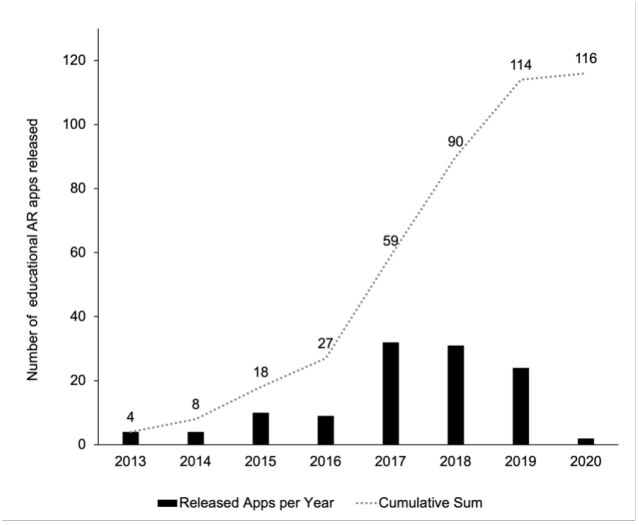


Figure 3.6: Total educational MAR apps released per year since 2013 and its cumulative sum

programming interfaces or software development kits (SDKs): ARCore [7] and ARKit [8] , respectively. These platforms allow developers and users of iOS and Android to create AR apps for mobile devices [107]. In addition, other SDKs such as Vuforia have eased the implementation of AR capabilities regardless of the device's primary operating system [9] . Overall, the release of these SDKs may explain the increase in the number of app releases.

AR-types

Finally, to explore the type of AR technology used by the mobile applications, the description given by the developers was examined manually. If the technology was not explicitly expressed, the application was downloaded and subsequently assigned to one of the AR types.

As it can be observed from Table 3.2, marker-based apps account for the majority of the apps (66%). On the other hand, markerless apps account for 23% of the apps and hybrid types only for 11%. Interestingly, nine apps were found to mix markerless technology and virtual reality. Finally, no applications that used location-based AR were found in the data set.

[7] https://developers.google.com/ar
[8] https://developer.apple.com/augmented-reality/
[9] https://library.vuforia.com/getting-started/overview.html

As mentioned, the majority of educational apps were classified as marker-based apps. This result may be explained by the fact that marker-based technology is easier to implement, the tracking is more robust, and it is less computationally expensive when compared to markerless AR [19, 108]. Altogether, these results suggest opportunities for scholars and developers to create more markerless apps and apps using combinations of AR types. However, for educational purposes, each type of AR may provide different outcomes; therefore, it would be interesting to have more explicit guidelines on the different learning affordances that each tracking technology has to offer per field and level of education so that users and practitioners can choose the application according to their needs.

Table 3.2: The number of apps per type of AR technology

AR-type	Number of apps	Percentage of total ($N = 116$)
Marker	76	66%
Markerless	27	23%
Markerless and VR	9	8%
Marker and markerless	4	3%

3.3.3 Which are the levels and fields of education that educational MAR apps published in the Google Play Store target?

To understand the educational target (levels and fields of education) of the MAR apps in the Google Play Store, each application was assigned to a field and level of education of the corresponding ISCED classification according to the information identified in the developer's description.

Table 3.3: The number of apps on each aggregated level of education according to the ISCED [85]

Aggregated level of education	Number of apps	Percentage of total ($N = 116$)
Levels 0-2	78	67.2%
Levels 3-4	4	3.4%
Levels 3-4 & or 5-6	22	19%
Not Classified	12	10.3%

Table 3.3 shows that most apps target the aggregated levels 0-2 (67.2%). This observation aligns with previous educational app studies, which have found that most developers target preschool and primary school-age children [109]. Interestingly, this finding is contrary to that of Bacca et al. [19], who found that most AR studies target higher levels of education. One possible explanation

is that the app market locates the "kids app market" into several industries. Developers use the term "educational" or tag words like "kindergarten", "middle school", etc., to sell apps for the children's audience. Still, they do not necessarily develop the apps with an educational purpose or this specific audience[10].

It is important to note that for some apps (10%), the target level was not explicitly expressed. This outcome may be relevant for future developers who aim to interest students in their apps. Adding this information may also give practitioners more confidence in using a particular app in a formal setting.

Table 3.4 shows that most apps were classified in "Generic programs and qualifications" and "Natural Sciences, Mathematics and Statistics". Overall, these results reflect those of Bacca et al. [19], who also found that science was the most popular field for AR studies, suggesting that this is due to AR's advantages when teaching abstract concepts. However, further studies and app development in different fields of education are recommended to expand and leverage the advantages that AR may bring in various educational areas. For example, the public market of educational MAR apps can exploit more apps in the context of science and engineering but for higher levels of education with validated educational content. Also, new apps can target adult learners who may not be considered under the International Standard Classification of Education (ISCED). Furthermore, developers can venture into new fields, such as social sciences and business or administration, which may have new interesting affordances. Moreover, public institutions can collaborate with developers in making free, content-validated, and relevant for particular curricula educational MAR apps for a wide range of students.

3.4 Limitations

The design of the current research is subject to limitations. First, the search of the apps through the Google Play Store is subjected to a limit of 250 apps per search term. Although the method ensures that the considered apps are the ones the user can find by using the same words, the data set can be extended if other search forms are employed, for instance doing searches from different locations and different languages or adding results from websites. Second, utilizing only six search terms can be a source of bias when creating the data set; a more exhaustive list of terms might improve the broadness of the apps found. Even though the criteria for selecting the terms were based on users' most commonly searched terms, there may be other words that could also yield results for AR

[10] https://www.bjornjeffery.com/2019/05/31/the-kids-app-market-a-strategic-overview/, accessed: 01-06-2020

Table 3.4: Classification of the apps according to the aggregated level of ISCED of broad fields of education [85] in descending order and the corresponding general topics

ISCED Broad Fields	General Topics	Number of Apps	Percentage (N=116)
00 Generic programs & qualifications		79	68%
	Natural Science, animals	16	14%
	Multiple Areas	15	13%
	Natural Science, space	14	12%
	Literacy	12	10%
	Art	7	6%
	Natural Sciences	6	5%
	History	3	3%
	Natural Science, biology	2	2%
	Foreign Language	1	1%
	Geography	1	1%
	Mathematics	1	1%
	Social Skills	1	1%
05 Natural Sciences, Maths & Statistics		13	11%
	Chemistry	7	6%
	Physics	4	3%
	Mathematics	2	2%
N/A		12	10%
	Visualization Platform	6	5%
	Museum	4	3%
	Game	1	1%
	Marketing	1	1%
09 Health and welfare		9	8%
	Anatomy	9	8%
07 Engineering, Manufacturing & Construction		3	3%
01 Education 02 Arts and humanities 03 Social sciences, journalism & information 04 Business, administration & law 06 Information & Communication Technologies (ICTs) 08 Agriculture, forestry, fisheries & veterinary		0	0%

apps, which were not considered in this study. Third, selecting mobile apps only from the Google Play Store limits the spectrum of apps. This opens the possibility for future research regarding an extension of MAR educational apps from the iOS store and the possible comparison with Google Play Store apps. Another limitation comes from using ARdoc to classify the reviews. The tool has its own "accuracy" and intrinsic biases derived from its assumptions and characteristics. In addition, even though the taxonomy used by [93] is in line with the objective of this chapter, the exclusion of reviews with the "Other" tag may have excluded sentences with relevant data. Finally, the extracted reviews may not represent the total number of users of the MAR apps. In general, only users more inclined to share their thoughts are described in the data set. However, the group can still be considered diverse as different target audiences use the apps, and the reviews show diverse perspectives.

3.5 Conclusions

The present chapter aimed to extract general information about educational MAR apps published in the Google Play Store and examine the user experience by classifying and analysing user reviews.

This research demonstrated that most users of educational MAR apps encountered various problems that hamper their experience (65% of reviews classified as problem discovery). Only a few apps appeared in the "high install" range, suggesting that the adoption of educational MAR apps is at an early stage. However, the emergence of Google and Apple's software development kits has enabled more AR apps to be released since 2017, which may suggest that the technology could soon reach its potential if users favour the apps and developers leverage the recent advances of AR.

Although marker-based AR was the most popular technology used in the apps, its practical shortcomings may block AR adoption, as expressed by some users. Improving upon these shortcomings and focusing on future developments of AR, such as markerless technologies, may improve the user experience and increase the acceptance of the technology.

Among the problems experienced by users of the current MAR apps, the application's functionality was the most popular. Therefore, exploring the causes of these problems in more detail may provide a better chance to enhance technology adoption. However, since users often do not provide detailed information, these results must be interpreted cautiously.

Regarding the "helpfulness" of AR apps for users, the results showed that some

users, regardless of the limitations of the apps, can recognize specific scenarios where AR apps are helpful. These observations align with what several AR studies in educational settings have acknowledged regarding the advantages of education. However, unlike these controlled studies, the reviews showed some real-life examples from the subjective perspective of the user.

The most popular demands among users were content-related. Users generally request additional content as most apps only provide a few models for visualization, target images, or limited free content. Therefore, future app developments can be concentrated on creating apps with a more comprehensive experience other than merely displaying a limited amount of multimedia.

Finally, the results showed that most apps in the data-set target childhood education and science is the most popular field of exercise. These findings may offer opportunities for developers and scholars interested in expanding the use of AR in education. Still, continuous efforts are needed to make AR more accessible to diverse audiences. This research did not assess the "educational" value of the apps; therefore, further research with a stronger focus on educational purposes is suggested. Future studies can include, for example, the assessment of multimedia learning principles inside the public MAR apps. On the other hand, developers could focus on providing detailed explanations about each application and its use and certainly validated content to improve the user experience of educational MAR app users.

Undoubtedly, future research towards addressing technical issues must be a priority for researchers to ensure that AR can be well-integrated with society, regardless of its use. More significant multidisciplinary efforts are needed to ensure that well-develop MAR apps with relevant educational content are available for more people no matter their level of education, country, or current device.

To aim for increased AR adoption as an educational tool, developers, educators, designers, and scholars may understand and address user concerns while boosting learners' educational value and experience. Altogether, academic research and public sources, such as user reviews explored, may provide the relevant information needed to achieve this goal.

The research questions answered in this chapter enabled the ideation and definition of the general user need when developing a mobile AR application (Section 1.4). It was established from this study that educational AR applications could be beneficial for students, especially if the learning content is validated and focuses on higher levels of education. The markerless type of AR is relatively unexplored and is preferred by users. To ensure accessibility, specific technical issues must be considered, such as the compatibility of devices. Furthermore,

free content, interactivity, and user-friendly interfaces must be provided for users to download and use the application easily. Based on these results, the basic requirements for the prototype were established.

To this end, MAR Lab should:

1. use validated learning content to ensure educational effectiveness.
2. regard a scientific topic for higher education to address the gap of this level of education.
3. use markerless AR technology to improve user experience.
4. be compatible with various devices to ensure accessibility.
5. have interactive content to engage users.
6. have a user-friendly design to improve usability.

These characteristics were considered while designing and developing the MAR lab, the details of which will be presented in the following chapters.

Chapter 4

Design and Development

This chapter is adapted from: J. L. Domínguez Alfaro, S. Gantois, J. Blattgerste, R. De Croon, K. Verbert, T. Pfeiffer, and P. Van Puyvelde, "Mobile Augmented Reality Laboratory for Learning Acid-Base Titration," Journal of Chemical Education, vol. 99, pp. 531-537, 2 2022.

The research conducted in Chapter 3 enabled the ideation and definition of the users' needs for the development of a mobile AR application (Section 1.4). It was established from this study that educational AR applications could benefit students, given they are validated and use the relatively unexplored markerless type of AR. To ensure accessibility, technical issues such as device compatibility must be considered. Additionally, free content, interactivity, and user-friendly interfaces must be provided to enable users to download and use the application easily. All these requirements were taken into account during the design and development of MAR Lab. This chapter presents the design, development, and testing of its first digital prototype, which was developed using rapid prototyping methodology.

4.1 MAR Lab Design and Development

4.1.1 Research Methodology

Rapid prototyping (RP) was used in this thesis as a design methodology to design and develop MAR Lab. The advantages of rapid prototyping include

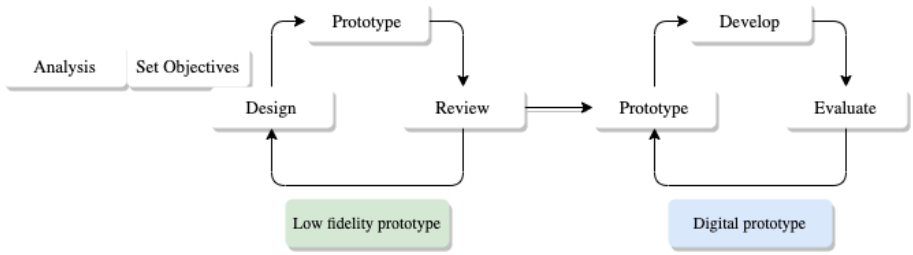


Figure 4.1: Rapid Prototyping Model adapted from [33, 35, 34] to show the steps followed to develop MAR Lab. The analysis and objectives definition concurred with the design phase. Each cycle corresponds to the low-fidelity and digital prototypes, respectively.

ease for users to try systems, find problems and provide feedback on each step of the process. It also reduces development time, cost, and communication issues and, more importantly, yields a suitable system for the designated task. However, rapid prototyping assumes that a design environment can be modified and synthesized quickly. Therefore, it requires tools that allow modularity and plasticity. Modularity means that different units can be modified without affecting the whole product. On the other hand, plasticity refers to the ability to change aspects of the units with little time and cost drawbacks [33, 35, 34].

The iterative process for creating MAR Lab began with defining the educational need with the expert panel and brainstorming about the prototype. Note from Figure 4.1 that these phases did not occur linearly. According to RP, the complete understanding of the needs and objectives results from the design process rather than as input. In this case, the objectives defined were meant to guide but not be exhaustive; some changed over time.

Subsequently, a low-fidelity prototype was created and evaluated using a think-aloud study and a survey. The results from the evaluation served as the basis for developing the digital prototype of the MAR Lab app. Each of these cycles is described separately in the thesis.

4.2 Low Fidelity Prototype

This section describes the first phases of the RP model (4.1) and the development cycle of the low-fidelity prototype.

4.2.1 General Concept

The MAR Lab application is designed as a markerless, interactive AR tool for training students in titration experiments. It has minimal physical requirements (e.g., users do not need to print a marker or need the real object), making it accessible to a wide range of users with handheld devices capable of running AR technology. This allows for an immersive learning experience without requiring specialized equipment or materials.

The application utilizes the affordances of augmented reality to allow the learner a more "realistic" environment than a simple smartphone application, meaning that the user interacts with 3D models of materials generally present in a real laboratory.

Currently, most AR research is focused on the affordance of "augmenting" the physical world with virtual content and blending both worlds through tangible interactions. However, this prototype stresses an inverted view: the virtual laboratory "augments" the user's physical world and contextualizes the content into the physical reality without tangible markers.

4.2.2 Analysis Needs and Educational Content

The case study developed for the prototype relates to acid-base titration. This topic was chosen as it is one of the typical experiments done in introductory courses of both high school and undergraduate curricula. Furthermore, as reported by other scholars [110], practitioners at KU Leuven appointed that students struggle to understand the underlying chemistry behind the titration experiment and thus hypothesize that a "practical" experience will remedy this difficulty.

Titration is a method of quantitative chemical analysis used to determine the concentration of an analyte in solution by allowing a carefully measured volume to react with a standard solution of another substance (titrant) whose concentration is known. The unknown concentration of the analyte can be calculated by finding the volume of the titrant that reacts stoichiometrically with the measured volume of the analyte at the equivalence point [111].

The titration set-up comprises a long-graduated tube called a burette with a stopcock at the bottom, which controls the titrant's flow (Figure 4.2). Indicators are used to indicate when the acid-base reaction is complete. The point at which the indicator changes colour is known as the endpoint of the titration. It is important to differentiate between the equivalence point and the endpoint. The equivalence point is when equal amounts of acid and base have been combined.

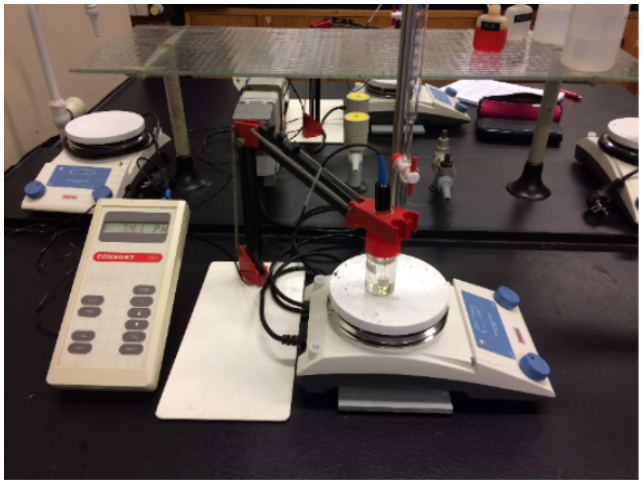


Figure 4.2: The hand-operated titration set-up generally consists of a burette, a pH meter, a magnetic stirrer, a magnet and the beaker with the analyte to be tested.

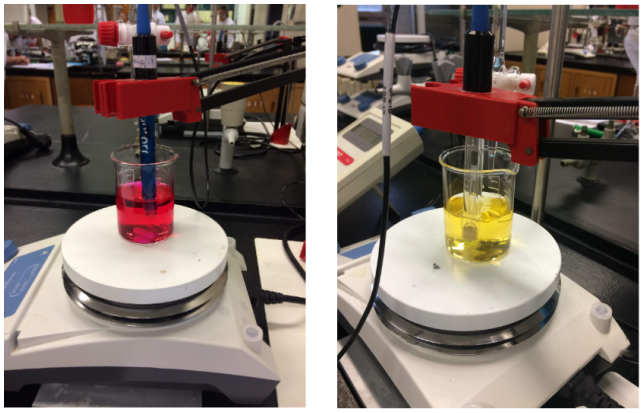


Figure 4.3: The indicator changes colour due to a molecular structural change that happens when the pH of the unknown solution changes.

At this point, the acid-base reaction is complete, the addition of the titrant from the burette is stopped, and the volume added is recorded (Figure 4.3). The endpoint, on the other hand, is when the indicator changes colour. It is an experimental estimate of the equivalence point [112].

To achieve accurate results, it is crucial to select an indicator that closely aligns with the equivalence point. More accurate values can be determined with an electronic instrument such as a pH measurement device [113].

The difference in volume added between the equivalence point and the endpoint is known as the titration error. This error arises due to the inability to determine the exact equivalence point and it is additional to the errors that may occur during volume measurements or weighing [114].

The values of pH measured at different stages of the titration can be plotted against the volume of the titrant added. As a result, the titration curve is obtained (Figure 4.4). In the laboratory, these values are measured and plotted, whereas, in theoretical courses, the stages of the titration are calculated [113].

One can determine the concentration of a strong acid or base as well as a weak one. In each case, different calculation considerations are required. The first prototype comprises one level in which the learner is tasked to "perform" a titration of a weak acid (acetic acid) with a strong base (sodium hydroxide). The user records the result using a pH meter and identifies the equivalence point with phenolphthalein as an indicator.

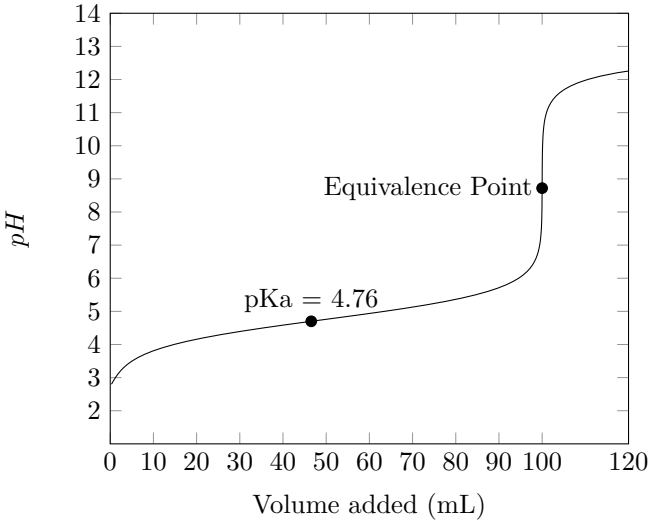


Figure 4.4: Titration curve for the titration of 100mL of 0.2M acetic acid (CH_3COOH), a weak acid, with 0.2M of sodium hydroxide ($NaOH$), a strong base. The slightly flat region of the curve between 10mL and 90mL of added base indicates the buffer region ($pK_a - 1 \leqslant pH \leqslant pK_a + 1$). The pH value at the midpoint of this region equals the pK_a value of the acid. The equivalence point for this titration is at a pH of 8.72. Phenolphthalein is a suitable indicator because it changes from colourless to pink between a pH of 8.2 and 10, making it easy to detect the equivalence point.

4.2.3 Defining Learning Objectives

The specific learning objectives of the application were defined under the guidance and supervision of two lecturers of the General and Technical Chemistry (Algemene en Technische Scheikunde B-KUL-H01A8A) course at KU Leuven. These learning objectives were realized using the revised Bloom's Taxonomy [115], which can then be matched with a specific type of assessment. The goals include both practical and cognitive skills, and they are:

- To identify and remember the materials of a titration set-up
- To explain the function of the different components of a titration set-up
- To assemble and operate the experimental set-up
- To interpret, illustrate and assess qualitatively the titration curve
- To calculate the unknown concentration of an analyte in a solution

The objectives are integrated into the different levels of the application and are represented in Figure 4.5 as white rectangles on each level or learning task.

4.2.4 Learning Experience Design

In general, it is assumed that learning chemical concepts is a complex task due to the abstract nature of the subject matter. For this reason, the design draws on complex learning instructional design principles to minimize the cognitive load outside the system and ensure that interaction with the application does not hinder users' learning.

The learning experience has been designed with four levels. It incorporates the four components of the 4CID model: learning tasks, procedural information, supporting information, and part-task practice (Figure 4.5).

Each level is proposed as a complete task; this is the basis of complex learning, which moves from an atomistic view (separated into parts) to a more holistic view of the task. The intention is that users can train at each stage with increasing levels of complexity (4C/ID sequencing principle 2.3.5). This complexity is reflected in the design through the different learning objectives that students are expected to be able to achieve at each level.



Figure 4.5: a) The MAR Lab is designed using the four components of the 4C/ID model; learning task (**LT**, black triangle), procedural information (**PI**, light grey triangle), part-task practice (**TP**, medium grey triangle) and supportive information (**SI**, dark triangle). Each component shows linked example screens from the prototype. The **LT** is represented by a list of tasks (titration experiments) that appear in the side menu. **PI** (light grey triangle) is provided through both timely messages when completing a step and hints on the procedure. **TP** is found in the application in the form of instructions which can be viewed in the logbook and 2D screen. Finally, **SI** is accessible through the side menu on each level and presented to the user on a 2D screen. b) MAR Lab has four levels which are shown through blocks of triangles. Each block represents a different level of task complexity and includes each of the 4C/ID components.

4.2.5 Level Structure

The MAR lab application is also based on experiential learning [61] and embodied learning theories [73], which aim to enhance a deeper understanding and improve students' participation in learning by eliciting bodily actions (e.g., the user must move their hands and their bodies to interact with the lab materials). The hands-on experience through concrete activities with virtual objects emphasizes the concretization of actions that may be suited to teach complex scientific concepts [66]. Additionally, the actions performed in the MAR lab application aim to improve the sense of control while using the application. By providing control to the user, the app aims to enhance the enjoyment of the experience related to the learner's engagement [4].

The interaction with the objects simulates the natural actions such as grabbing, reading, pouring, observing, mixing, etc., done in real settings. Thus, reducing the need of the user to construct a context in real-time and allowing engagement with learning the concept. Additionally, the interface guides the users through the experiment, provides contextualized chemical information (4C/ID temporal and spatial- attention principles 2.3.5), and additional forms of representation of the phenomena observed during the titration experiment (e.g., change in pH, dissociation of acids and bases) (4C/ID redundancy principle 2.3.5), which they can subsequently reflect upon and complement the conceptualization of the abstract concepts of the learning module [72].

Each level elicits four of the five general phases of the inquiry-based learning cycle (Figure 4.6).

1. The orientation phase is depicted at the start with the learning module's instruction.
2. The conceptualization phase includes multiple-choice questions posed to the learners by a cartoon character before the experiment (Figure 4.15.5, 4.15.6). These questions aim to activate prior knowledge of the user and test their understanding during the experience [66] (4C/ID prior-knowledge activation principle 2.3.5).
3. The investigation phase comprises sixteen steps in total. These steps include assembling the materials, titrating, and "cleaning" the environment. The cartoon character provides hints in this phase.
4. The conclusion phase is a guided analysis by the cartoon character after completing the experiment through multiple-choice questions. A review screen is made available at the end of the training.



Figure 4.6: The stages of the application mapped into the inquiry-based learning cycle [71, 72]. All phases of the cycle correspond to a stage of the learning experience, except for the discussion phase.

4.2.6 User Evaluation

The low-fidelity prototype was developed in two stages. First, paper sketches were created to explore the overall design of the prototype. These sketches were then translated into an interactive PowerPoint (Figure 4.7). To evaluate the prototype, a think-aloud study was conducted to identify areas of improvement. To gain a wider perspective, a large-scale online survey was conducted to gather feedback from potential users.

Participants

Fifteen students from the General and Technical Chemistry (Algemene en Technische Scheikunde B-KUL-H01A8A) course of KU Leuven were recruited. The sample comprised 12 males and 3 females; all volunteered to participate. In addition, an online survey was distributed to 85 students from the same course.

4.2.7 Results

Think-aloud study

The think-aloud study was conducted to analyze the user experience of potential users and detect any ambiguity in the original design. Fifteen first-year engineering students volunteered to participate in this study and expressed their enthusiasm for the concept. As a result of the study, different issues were identified, which provided an opportunity to refine the design. Table 4.1 presents the issues identified from the evaluation of the prototype. The most recurrent problems are discussed in detail below.

The results from the study suggest that the two most recurring ambiguities were the role of the logbook and the "stop" titration button. Most participants were unfamiliar with the logbook concept, resulting in them not using or accessing it. Initially, the logbook was placed at the bottom of the interface (Figure 4.7, blue layout). Still, this design element was changed to make it more intuitive

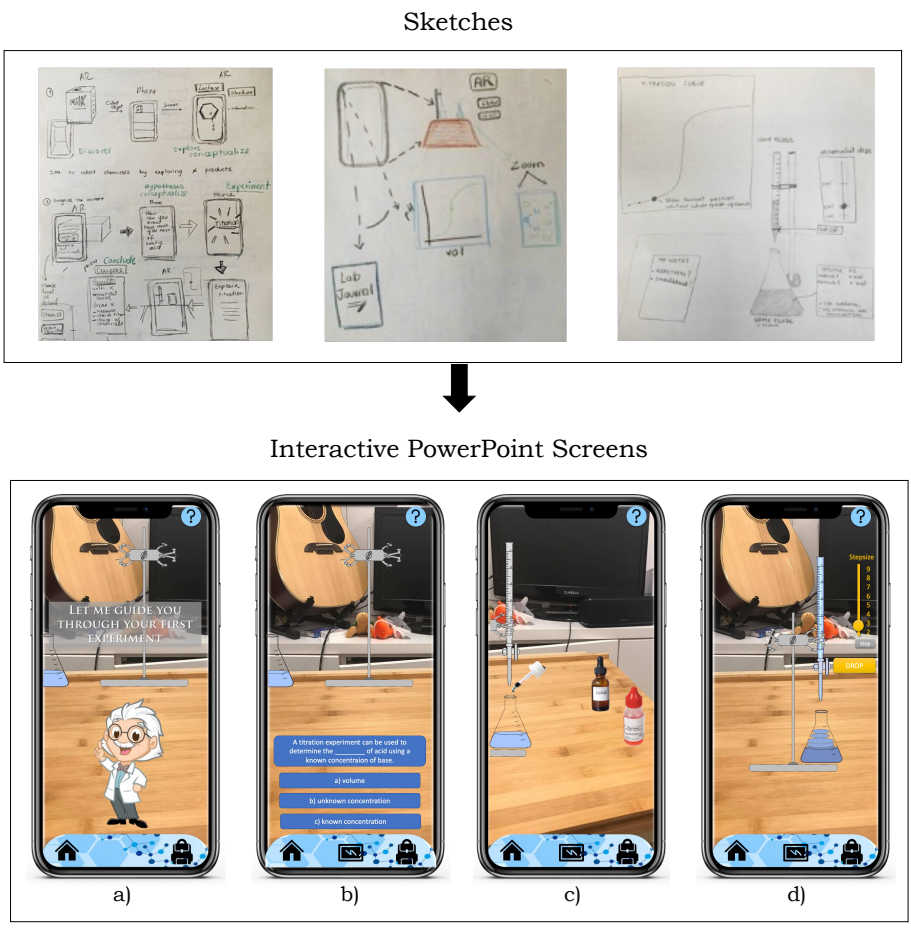


Figure 4.7: A low-fidelity prototype was created as an interactive PowerPoint from initial sketches generated during the ideation process. In the study, participants were shown different interactive screens including (a) cartoon-based introduction, (b) multiple-choice questions in the AR scenario, (c) assembly of materials for experimentation and (d) the titration step where the use of the slider was needed to handle the volume from the burette. The images were adapted from [116].

for the participants. The new design used an icon at the top of the screen that users could click on (Figure 4.8a).

Regarding the "stop" titration button, users initially had trouble understanding the objective of the button (Figure 4.7d). Therefore, the text in the button was modified to "finishing titrating." (Figure 4.8b).

Table 4.1: The most popular issues mentioned by participants and the solutions proposed.

Issue	Number of participants that mentioned this issue	Possible solutions
The function of the button "stop" titration is unclear	8 /15	Explain the function at the start or change the title, e.g., "finished."
The role of "logbook" unclear or unused	8/15	Animate the flow of information or show an example at the start
Location of the logbook	3/15	Use an icon instead to access the function
Information arrows not visible	2/15	Make sure that enough contrast is present
Icon "backpack" unclear	4/15	Use another icon, e.g., book
Unclear about the order of levels	3/15	Use different colors per level and add numbers

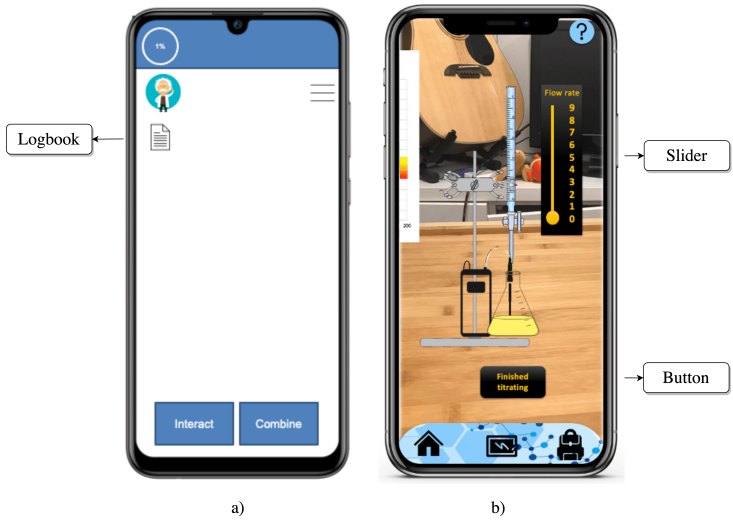


Figure 4.8: The proposed modifications to the prototype include (a) placing the icon to access the logbook at the top of the screen instead of requiring the user to slide the screen from the bottom to interact with it; (b) adding a button to stop titration. Furthermore, it is proposed that the user interacts with the titration step by using the slider (to control the flow from the burette) and the button to check if the reaction is finalized and stop the titration.

It is important to note that the issues found in the think-aloud study could result from the Powerpoint presentation's limited interactivity compared to a smartphone app. Nevertheless, these observations were useful in identifying potential challenges caused by the prototype's function design.

Online Survey

The purpose of the online survey was to capture the participants' opinions on the major design decisions for the application, specifically on the "look" of the application. Participants were asked to indicate which choices they preferred and provide their opinions on features such as icons, functionalities, use of a cartoon, language, and multiple-choice questions (MCQs). The survey is available in Appendix B.1, and the results are discussed below.

In the survey, 85% of the participants indicated that they had no previous experience with AR. 76% of the participants favoured the tour with a cartoon over the version without, citing reasons such as its attractiveness, personalization, and visual appeal. Based on this outcome, a cartoon was incorporated into the digital prototype to provide hints and explanations during the experience.

For the learning modules, two versions were proposed: a graphical option with islands being the modules (inspired by ARCore Elements from Google) and a less interactive version with modules represented as buttons in a vertical list. 47% and 53% of the participants respectively preferred each option; hence the less interactive version was chosen due to its simpler development.

The participants were also asked to rank the functions from most necessary to least necessary on a scale of 7. The highest-ranking functions were reviewing the assignment, undoing something, receiving a hint, quitting a module, and returning home. Due to their complexity, these functions were implemented in the final prototype, except for undoing something and redoing something.

The participants selected the house icon to quit a learning module, the question mark to get a hint, and either the pile of books or the open book to check the theory of the learning material. Moreover, 23% of the participants said that the language would impact the use of the application, with English being the preferred language for 73% of them. Therefore, the application was developed in English to maximize its reach.

Finally, the survey results showed that the participants agreed that integrating multiple-choice questions will not impact negatively the user experience and may add educational value. Consequently, multiple-choice questions were incorporated into the digital prototype.

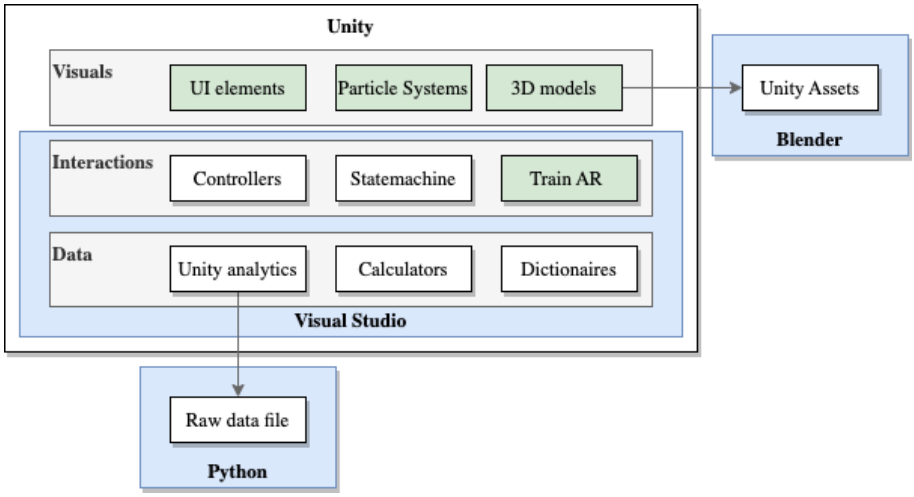


Figure 4.9: Overview of the software used in the development of the MAR Lab. The resources are categorised into visual, interaction, and data. External packages imported into the game engine are represented by green squares, while the software used in the prototype, including Unity, Blender, Visual Studio, and Python, are represented by blue squares.

4.3 Digital Prototype

The evaluation results set the starting point for developing the digital prototype. This section describes the technical implementation, the design of the digital prototype's first level, and its evaluation.

4.3.1 Technical Implementation

The technical implementation of the prototype was realised using different platforms and software. This section describes the most important resources for development and their details. Figure 4.9 illustrates the various platforms' relationships to the prototype elements. These resources can be divided into three categories: visuals, interaction, and data. Visual elements include user interface components such as graphs, notebooks, notes, menus, etc. Interaction features include controllers, Statemachine, and the TrainAR framework. Data elements include Unity analytics, raw data files, calculators, and dictionaries. The software used in the prototype includes Unity, Blender, Visual Studio, and Python. The green colour of some squares indicates that these resources were external and imported into the game engine.

Unity

The Unity game engine is a software environment and robust production tool for interactive content such as 3D animations and AR and VR experiences. One of the main advantages of using the game engine as the environment to build the application is that unlike the officially integrated development environments (Android Studio for Google's Android operating system or Xcode for macOS), which are specific to the operating system, in the Unity game engine most of the resources can be reused (e.g., prefabs, scripts, multimedia) and only the SDK specific setups have to be added or developed [54].

The MAR Lab app was developed using Unity Version 2019.4 and 2021.2. The AR Foundation [1] package was used to enable the AR capabilities within the engine. This package serves as a bridge that allows the specific features of each AR SDK (ARCore and ARKit) to be used in Unity. It is important to note that these SDKs must be installed in Unity to take advantage of the capabilities of each AR target platform.

MAR Lab was built using Unity, allowing it to be deployed on Android and iOS devices. The Android version was distributed through an APK file that could be downloaded and installed directly on the devices. A beta version of the app was also submitted to the Google Play Store for testing [2]. For iOS, the app was distributed as a beta version and was submitted to the TestFlight service for testing.

Figure 4.10 illustrates the general hierarchy of the project within the Unity game engine. The application has four main scenes: menu, logbook, tutorial, and AR scenario. Each scene has its own hierarchy, and game objects represented by arrows trigger transitions between scenes. The figure simplifies the structure by indicating with an X the levels that were added in later iterations. As mentioned in Section 4.1.1, the prototype's modularity allows for easy extension with additional levels.

Unity packages

The online asset store[3] of the Unity game engine works as an online marketplace for developers to obtain various resources, thus reducing the burden of developing every asset from scratch [54].

[1] https://docs.unity3d.com/Packages/com.unity.xr.arfoundation@5.1/manual/index.html
[2] MAR Lab 1.0: https://play.google.com/store/apps/details?id=com.thesisAR.MARLab&pli=1
[3] https://assetstore.unity.com/

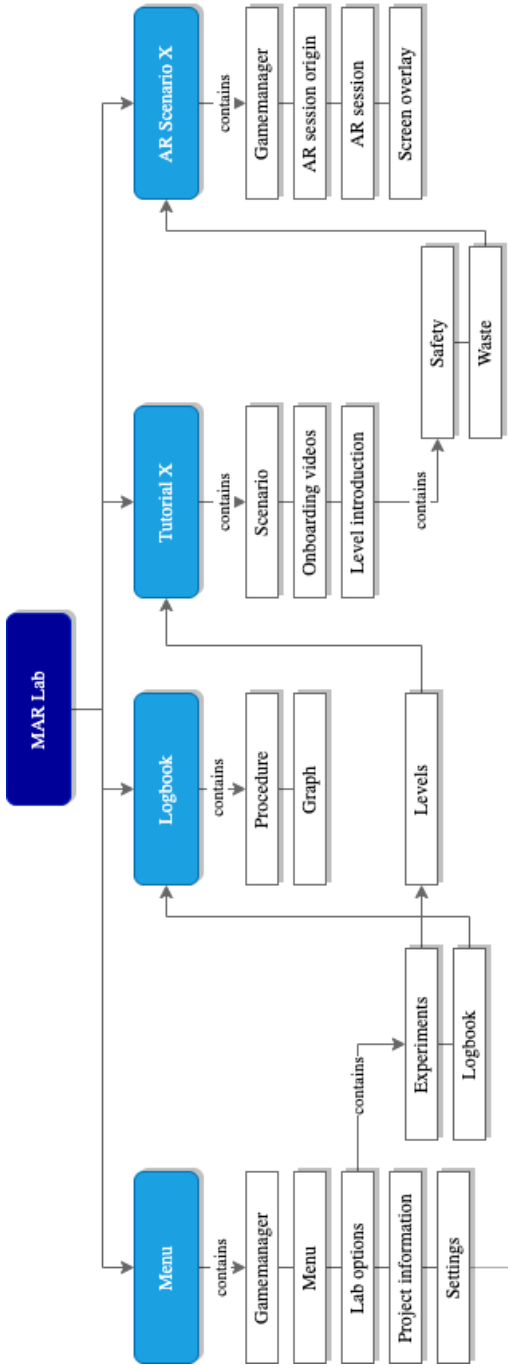


Figure 4.10: MAR Lab's Unity Hierarchy. The light blue rectangles represent the scenes that are built in MAR Lab. The white rectangles represent the game objects that belong to a certain scene. The arrows represent the transition in the application. The X in the scene tutorial and AR scenario represent the level.

The packages added to the application included Archviz Laboratory Pack (version 2.0)[4], Chemistry Laboratory Equipment (version 2.0)[5], PBR Jerrycan Free (version 1.0)[6], Liquid Volume Pro version (version 4.1)[7], Easy File Pro (version 2.6)[8], Cartoon FX Remaster Free (version 1.03)[9], Cleaning Products (version 1.0)[10]and Modern UI Pack (version 3.2) [11]

Some of the objects were modified using Blender[12], a free general-purpose graphic software that allows the creation of photorealistic models. Some modifications included the change of colour and size and adding text to some 3D models.

Other resources

Unity game engine supports different digital multimedia. Some of these are video players (AVI, MP4, and MPEG), audio resources (AIFF, WAV, MP3, and Ogg), image (JPEG, GIF, PNG, TIFF, and PSD) and text file models with various formats. Multimedia can be either generated or imported into the game engine as game objects. Additionally, the multimedia resources can be handled directly in the engine or done through scripting in C# with Unity's API for user-specific demands [54].

The multimedia imported in Unity for MAR Lab included: videos for tutorial scenes, audio with instructions, images for the menu, icons, and specific objects such as message notes.

Unity Analytics

The Unity game engine also comprises an analytic tool, so-called Unity Analytics, to track the user's actions in the application within customizable features [54]. User interactions with MAR Lab were registered through Unity Analytics via Unity's Custom Events [13]. These events allow registering custom dimensions

[4]https://assetstore.unity.com/packages/3d/environments/archviz-laboratory-pack-144317
[5]https://assetstore.unity.com/packages/3d/props/tools/chemistry-laboratory-equipment-163610
[6]https://assetstore.unity.com/packages/3d/props/pbr-jerrycan-free-80011
[7]https://assetstore.unity.com/packages/vfx/shaders/liquid-volume-pro-2-129967
[8]https://assetstore.unity.com/packages/tools/input-management/easy-file-pro-187441
[9]https://assetstore.unity.com/packages/vfx/particles/cartoon-fx-remaster-free-109565
[10]https://assetstore.unity.com/packages/3d/props/cleaning-products-154461
[11]https://assetstore.unity.com/packages/tools/gui/modern-ui-pack-201717
[12]https://www.blender.org/
[13]https://docs.unity3d.com/ScriptReference/Analytics.Analytics.CustomEvent.html

```
AnalyticsResult result = Analytics.CustomEvent("OpenLogbook",
    new Dictionary<string, object>
    {
        {"Level", StartARScene.ARlevel},
        { "id", ApplicationSettings.idCode },
        { "stateNumber", Statemachine.Instance.stateNumber}
    }
);
```

Figure 4.11: Unity's Custom Event Code. The event registers as a dictionary the level, user id, and the state in which the user performed the action "OpenLogbook"

related to an event, such as the time, level, step in the procedure, user id, etc. Figure 4.11 shows an example of these events in the MAR Lab code.

It is important to mention that the limits that Unity Analytics imposes[14] on users were extended through contact with the Unity support team. The extended limits included 500 events per hour per user. During the development of MAR Lab, an update to the Unity Analytics API was released (called Gaming Services Analytics); however, no transition was made to this new platform.

The Unity Analytics raw file with logged events can be downloaded as a JavaScript Object Notation (JSON) file format. However, this is only possible through Unity Pro accounts. For MAR Lab, this account was obtained through the university, and with it, the access to these documents.

Train AR Framework

The digital prototype of the MAR Lab app was created using the TrainAR framework [31]. This framework proposes combining interaction concepts, a didactic framework, and a unity-based authoring environment to create procedural handheld AR trainings. An early version of the authoring tool, based on ARFoundation, was used. This approach was chosen to leverage familiar handheld MAR metaphors to create procedural task-based training based solely on virtual content across differing levels of media competency.

It is important to note that some of the features of the MAR Lab app were not present in the original framework. The domain-specific extensions include using the logbook, graph, or experimental chemical hints the cartoon gives. These elements ensured that the educational need was fulfilled and, as a result, extended the use of TrainAR for chemical education.

[14] https://docs.unity3d.com/Manual/UnityAnalyticsEventLimits.html



Figure 4.12: TrainAR Interact Action: To interact with an object, the user needs to (a) move the static crosshair displayed in the centre of the smartphone screen, (b) get near the object, and (c& d) then press the "interact" button to elicit the interaction. Adapted from [31]

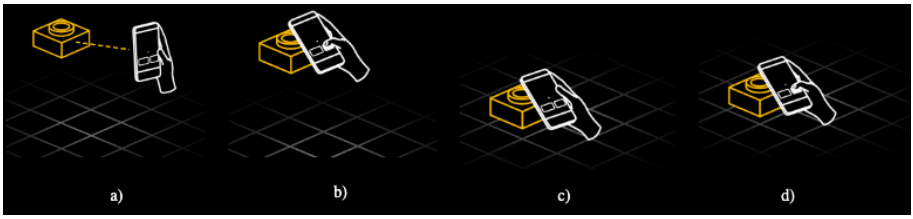


Figure 4.13: TrainAR Grab Action: To grab an object, the user needs to (a) move the static crosshair displayed in the centre of the smartphone screen, (b) get close to the object and press the "grab" button, (c) move the object to the desired location and then, (d) release the object by pressing the "release" button. Adapted from [31]

The TrainAR framework enables users to interact with virtual objects on the screen using buttons and a crosshair. To familiarize users with these interactions, instructions are provided at the beginning of the learning module through animations and textual explanations.

To select an object, the user aims the crosshair at the desired object, which is

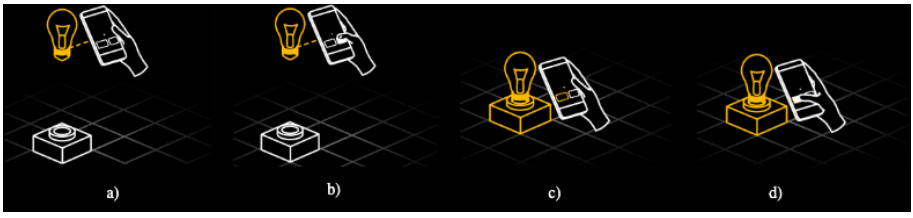


Figure 4.14: TrainAR Combine Action: To combine two objects, the user needs to (a) select the first object using the static crosshair, (b) trigger a "grab" action using the corresponding button, (c) move the object and overlap it with a second object then, (d) press the "combine" button to elicit the interaction. Adapted from [31]

then outlined. With a selected object, the user can grab, combine, or interact with the object using the corresponding button. If the user interacts with the object (Figure 4.12), different actions may be triggered, such as a sound or a note being displayed on the screen. If the user grabs the object (Figure 4.13), it appears in a fixed position in front of the screen, and its size may be adjusted to allow the user to view the object without blocking their field of view. The user can grab an object to move it within the scenario or combine it with another object. The user can release the grabbed object using the same button. If the user combines the object (Figure 4.14), an action will be triggered by overlaying the grabbed object on a secondary object within the scenario. If the user attempts to combine or interact with an object that the application does not allow, the buttons will be inactive and greyed out to indicate that the action is not possible. The button labels are congruent with the actions to facilitate the interaction and avoid confusion for the user. For example, the word "Attach" is displayed on the button when the user needs to combine a burette with a stand.

4.3.2 Level Flow

This section describes the general level flow realized after the first digital implementation.

In line with Section 4.2.5, the experience starts with explaining the context of the application and the task to be performed by the student (Figure 4.15.1). Subsequently, an optional tutorial on how to use the application is provided (e.g., showing how to interact with the objects as a dynamic visualization split into three animations (4C/ID dynamic visualizations and segmentation principle 2.3.5) (Figure 4.15.2). Before starting the experiment, both safety and waste treatment instructions are given (Figure 4.15.3). To start the experiment, the user needs to scan a surface using the camera to register the real environment in which the virtual laboratory will be placed (Figure 4.15.4). Once the tracking is stable, the virtual laboratory is deployed into the real environment. Then, the user can start the experiment by moving, selecting, grabbing, and combining the materials (Figure 4.15.8, 4.15.9, 4.15.10, 4.15.11). At the top of the screen, instructions about steps that need to be performed are given. For example, suppose the student combines objects wrongly. In that case, the system provides real-time feedback in the form of a sound signal and an outline surrounding the lab materials (Figure 4.15.10) (4C/ID training-wheels and signalling principle 2.3.5). The mistakes are also recorded to provide feedback at the end of the experiment (Figure 4.15.15). Finally, the steps performed by the user during the experiment are recorded in the logbook (Figure 4.15.12). The experimental measurements (pH vs volume of titrant) are shown in a graph (Figure 4.15.13).



Figure 4.15: This figure shows the MAR lab level flow and specific features inside the application. 1) introduction to the chemical context, 2) instructions on how to use the application, 3) safety and waste treatment instructions, 4) scanning of surface, 5) introduction by the cartoon character, 6) example of multiple-choice questions, 7) virtual content, the crosshair for selection and interaction buttons, 8) select interaction, 9) grab interaction, 10) combine interaction, 11) titration slider, 12) logbook overlay, 13) dashboard overlay, 14) hint is given after pressing the question mark in the top of the screen, and 15) feedback overlay.

4.3.3 Methodology

The goal of the test was to explore the usability of the application and the learning outcomes and get feedback regarding the user experience. For this evaluation, a mixed-method approach was taken. The quantitative analysis was used to determine the learning effect of the platform's use and measure the application's usability. On the other hand, the research uses qualitative analysis to gain detailed insights into the participants' experience with the application.

Participants

Fifteen participants were recruited two universities in Belgium in a voluntary basis. Recruitment was conducted through email and social media due to pandemic restrictions. Before starting the study, ethical clearance was granted from the ethical committee in KU Leuven G-2021-3236-R2(MAR). All the participants in the study were science or engineering students, each with their own unique background.

Procedure

The study structure follows a pretest/posttest design (Figure 4.16). The data collection and interaction with participants were performed online. The application was made available in the Google Play Store as a beta version. In total, nine participants completed a SUS questionnaire, an invariant pretest/posttest, through a web-based survey tool (Qualtrics). In addition, four users participated in an optional feedback session on a different day. Unfortunately, six users had difficulties downloading the application either because the device did not support the AR capabilities required (i.e., the phone lacked a gyroscope) or did not have enough memory available for the app to be downloaded. Therefore, only the results from users who completed the questionnaires are reported.

Instruments

SUS questionnaire According to ISO 9241-11, usability is "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use". The System Usability Scale (SUS) is a 10-item standardized Likert scale questionnaire that assesses subjective usability based on the criteria defined by ISO 9241-11. These criteria comprise effectiveness (accuracy and completeness with which

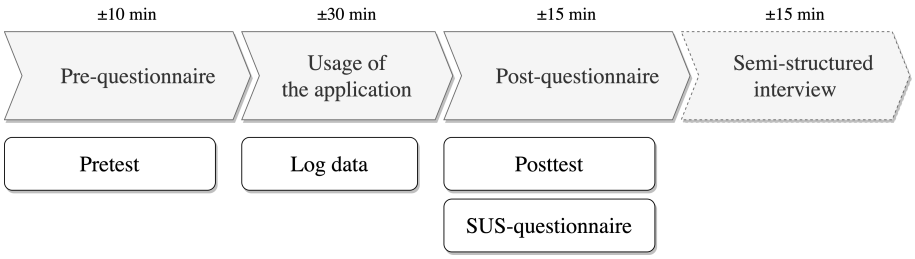


Figure 4.16: Overview of the study design and the time spent on each phase with each participant. The dotted line indicates that the phase was voluntary. The instruments appear in a rounded box below the phases.

users complete tasks), efficiency (resources used in relation to results achieved), learnability, and satisfaction (user's physical, cognitive, and emotional responses that fulfil user needs and expectations)[117]. Due to its brevity, cost-effectiveness, and broad applicability, SUS has become a widely used tool [118].

The SUS score represents the system's overall usability measure ranging from 0 to 100. The score is the result of the sum of each item's contribution. For odd-numbered questions, the contribution is the score value minus 1; for even-numbered questions, the contribution is calculated by subtracting five from the score. Finally, the overall SUS score is obtained by multiplying the sum of the scores by 2.5. In this thesis, the SUS scores are interpreted using the adjective scale proposed by Bangor et al. [118]. This scale contextualizes the numeric scores obtained by SUS into seven adjective ratings, i.e. worst imaginable, awful, poor, ok, good, excellent and best imaginable.

One of the limitations of using the System Usability Scale (SUS) is its brevity, which can lead to important details about users' experiences being overlooked. Furthermore, the subjective nature of its scales can cause discrepancies between a user's actual experience and their assessment of that experience. Additionally, using a Likert scale may not be sensitive enough to detect minor changes in usability over time. To address these limitations, this thesis takes advantage of additional methods, such as follow-up questions and objective measures of usability (e.g., task completion and error rates), to better understand the system's usability.

Learning Test An 11-item test was designed to measure participants' learning based on the learning outcomes defined in Section 4.2.3. This test was administered in English and can be found in Appendix B.2. It consists of short multiple-choice questions. The first five questions assess participants'

understanding of the theoretical concepts related to the titration experiment, such as the purpose of the titration, the properties of the chemicals used, and how they interact in the solution. Questions 6 and 7 focus on the components used in titration, such as the indicator, titrant, and analyte. Question 8 assesses recall of the steps involved in the titration experiment. Questions 9 and 10 evaluate participants' ability to interpret and use graphical elements to understand the behaviour of chemicals in solution and at specific pH values during the titration. Finally, question 11 tests participants' understanding of the proper disposal of chemicals in the laboratory. The instrument was used in this study as a pilot test.

Log data The user's interaction with the system was collected through Unity Analytics. The data collected included time spent with the application, the number of hints provided during the titration experiment, the number of interactions with the logbook, and the number of mistakes in the multiple-choice questions (Table 4.2).

4.3.4 Results

Prototype's Usability

For analysing the usability results, an open-source tool was used [119]. Figure 4.17 shows the SUS score. The mean SUS score for the first prototype MAR Lab is 72.8 (SD=14.0). According to the scale proposed by Bangor et al., [118], the result can be considered "good" and acceptable usability. These results align with the first usability studies conducted to develop the TrainAR [31] interaction framework, which also uses explicit interaction with buttons.

Learning Test

The invariant pretest/posttest was administered to participants to assess their knowledge of acid-base concepts and titration as a first pilot test. It consisted of eleven multiple choice questions (Appendix B.2), yielding a mean pre-test score of 7.89 ($SD = 2.37$) and a mean posttest score of 8.56 ($SD = 2.01$) (Figure 4.18 and Appendix B.3).

The one-tailed t-tests at an alpha risk of rejection of 0.05 revealed that the p-value ($p = 0.056$) was not statistically significant, indicating that the null hypothesis could not be rejected. As a result, it can be concluded that there is no significant difference between the mean of the pretest and the posttest

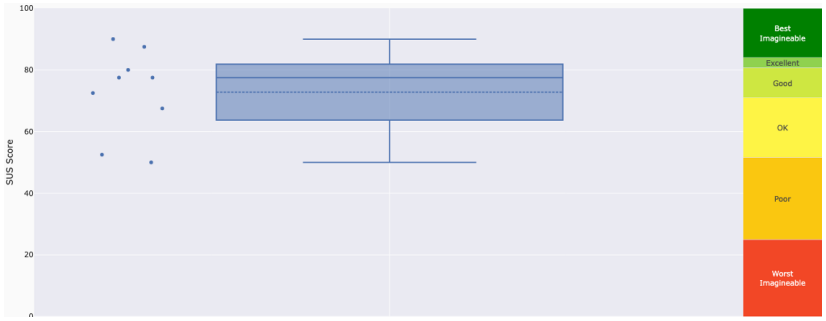


Figure 4.17: SUS score compared on different adjective scales using open-source toolkit [119].The mean score for the MAR lab 1.0 is 72.8 (SD=14); thus, it can be considered good and in the acceptable range of usability, according to [118].

Table 4.2: Results of log data recorded

Data recorded	Time spent (min)	Hints (Total hints =16)	Logbook interactions	MCQs mistakes (Total questions = 7)
Mean (N=9)	45	17	11	3
SD	19	14	6	2.7

scores. The t-test was conducted as the Shapiro-Wilk test indicated that the data were normally distributed (Appendix B.4).

Although the p-value was not statistically significant, it is important to discuss the intervention's effect size to determine if it is worthwhile investigating further. As shown in Table B.1, Cohen's d was found to be medium (0.596) [120]. Additionally, to further facilitate interpretability, the Common Language Effect Size (CLES) was calculated [121, 122]. CLES indicates that there is a 59% probability that the evaluation of posttest is higher than the evaluation of the pretest when a pair of scores is randomly selected. Overall, these results suggest that the intervention might affect knowledge acquisition and should be investigated further. However, it is important to bear in mind that the small sample size and the lack of formal validation for the instrument used for this pilot study limit the reliability of the results and further research is required to corroborate the findings.

User's Feedback

The final stage of the study comprised a feedback session with four voluntary participants. They were asked about the interaction with the application, the

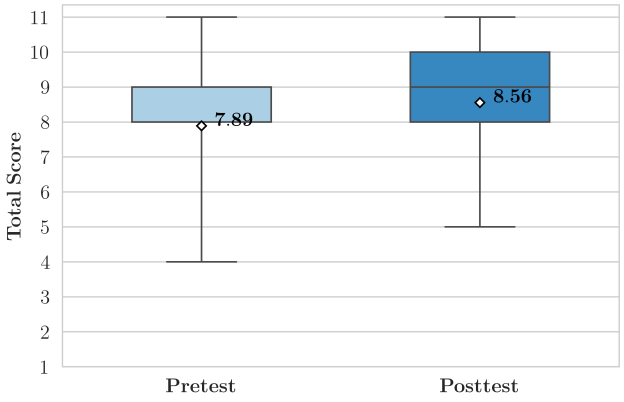


Figure 4.18: The boxplot of the pretest and posttest scores of the first KU Leuven study. The boxes represent the interquartile range (IQR), with the median of the scores indicated by the line inside the box and the mean scores indicated by the rhombus. The pretest scores have a mean score of 7.89 ($SD = 2.37$), while the posttest scores have a mean score of 8.56 ($SD = 2.01$). One-tailed t-tests found no significant differences in mean scores on the pre and posttest ($p = 0.056$).

simulation length, their opinion of the design elements (graph, MCQs, and logbook), and their general experience using the application.

A common view amongst interviewees was that interacting with virtual objects through the buttons on the screen was easy and intuitive. Although some participants expressed that they experienced some struggles with specific objects (e.g., floating objects or difficulty seeing through the burette), they were not indicated as major issues, which is also supported by the SUS score.

Interestingly, looking at the interaction with the system through the log data, it was observed that some of the steps might not have been clear to all the students. For instance, the average number of hints among users was seventeen (Table 4.2), which is higher than the total number of hints the system had (sixteen hints for sixteen steps). However, a detailed analysis revealed that during the step of "add an indicator to the analyte," students received the most hints. Therefore the interaction in this step may be simplified, and the hint rephrased to be more specific. Similarly, users expressed in the feedback session that the "cleaning" step was confusing; therefore, the instruction should be revised.

When asked about the logbook and the MCQs, the participants agreed that both elements were valuable and well-received. For example, one participant commented on the logbook: "I could find everything I needed." This comment

can be supported by the number of interactions with this element among other participants (Table 4.2). Another interviewee, when asked about the multiple-choice questions, said: "They let you think about the theory but guide you with possible answers." This comment may support that the multiple-choice questions helped users activate their knowledge or help them reflect. On the other hand, one user pointed out: "I thought I was finished, but then I realized I also have to perform the cleaning steps." Therefore, improving this element may benefit more users and help them understand the theoretical concepts that need to be conveyed by the system. For instance, the wording, the number of questions, and the timing are some aspects that can be improved.

When asked about the general functionalities, the participants agreed that the application contained many functions and, therefore, it was overwhelming at the beginning. These results were in line with previous research, which has established that AR, in some cases, can be distracting and cognitively demanding for students [21, 22]. However, the application design and additional onboarding improvement may help reduce the overload and profit from AR affordances.

In summary, these results provided important insights into what users had experienced and allowed us to rationalize what aspects benefited more users to implement them in the next iteration. Specifically, the logbook and the graph icons will be simplified by seamlessly integrating the functions as one object (a virtual notebook). This modification will allow the users to interact with the functionalities as part of the environment and avoid having extra icons on the screen.

The interaction with the multiple-choice questions will be improved by modifying the time the questions appear in the AR environment. Although the questions aim to trigger the conceptualization phase of the inquiry-based cycle and guide the conclusion phase, they may hamper the user experience by interrupting the flow of the application. Therefore, like other authors [72], the AR environment exclusively will support the investigation phase of the inquiry-based cycle, and the questions will be used at the end to guide the analysis in the conclusion phase.

Finally, apart from the technical issues to be solved (e.g., floating objects and liquids disappearing), instructions as voice-over for the cartoon character will be added to improve the clarity of some of the steps and help the users with redundant information in a different format. Regardless of the limitations of the current design, when asked about their experience with the application and its use, the qualitative feedback provided by the students showed that the technology is innovative to them, and they recognized that the application could be helpful when learning experiments like titration. Furthermore, it brings

diversity in delivering the chemical and procedural knowledge taught in the curricula. However, the results are only preliminary due to the limited sample size.

4.4 Conclusions

This chapter included the MAR lab's design, development, and preliminary test. As part of the iterative design cycle, domain-specific elements such as the logbook, graphical representations, and practical hints were identified and used to fulfil the educational need. These elements can be used as guidelines for creating new experiences using the TrainAR framework in the context of chemical education.

While preliminary, the results of this chapter suggest that although students experienced technical difficulties (unstable tracking, compatibility with devices, floating objects, etc.), the tool has acceptable usability.

Moreover, students could use the application independently, i.e., they could finish the training in a remote setup without supervision. The sample size limited this study; however, the students who tried the application were optimistic about the tool's novelty and its use for chemical experimentation.

The upcoming iteration of the development process will address the technical deficiencies that have been identified, as well as add extra content. This will be discussed further in the following chapter.

4.5 Acknowledgements

Stefanie Gantois is kindly acknowledged for her substantial contribution to the design and development of the first version of MAR Lab as part of her master thesis [116].

Chapter 5

Second Digital Prototype

Section 5.3 is partially adapted from: J. L. Domínguez Alfaro and P. Van Puyvelde, "Analysis of the Learning Experience in a Chemical AR Application Using Process Mining," in Lecture Notes in Networks and Systems, vol. 633, M. E. Auer, W. Pachatz, and T. Rüütmann, Eds. Cham: Springer International Publishing, 2023, pp. 770-777.

Following the third phase of the DIF Framework (Section 1.4), the first iteration of the MAR Lab digital prototype attempted to transform the initial idea into a working digital prototype from which improvements could be derived. After its initial deployment, user feedback indicated that usability issues hindered the learning experience. To address these issues and improve the user experience, this chapter describes the second iteration of the MAR Lab digital prototype (MAR Lab 2.0) and its evaluation. It also introduces the application of a process mining methodology to investigate the process that hampered users' usability and learning experience while using the application.

Three formative evaluations of MAR Lab are discussed. The application was tested with different populations. The results from these user tests aim to answer the following research questions:

1. How does the usability of the prototype change after the modifications?
2. Which steps in the procedure are problematic for the users in each group?
3. What is the general perception of MAR Lab among users?

4. Do participants improve their learning outcomes after using the application?

5.1 Prototype Improvements

To improve the digital prototype according to the initial evaluation observations outlined in Chapter 4, various modifications were implemented in terms of design, the context of the learning experience, and technical issues.

5.1.1 Redesign of MAR Lab

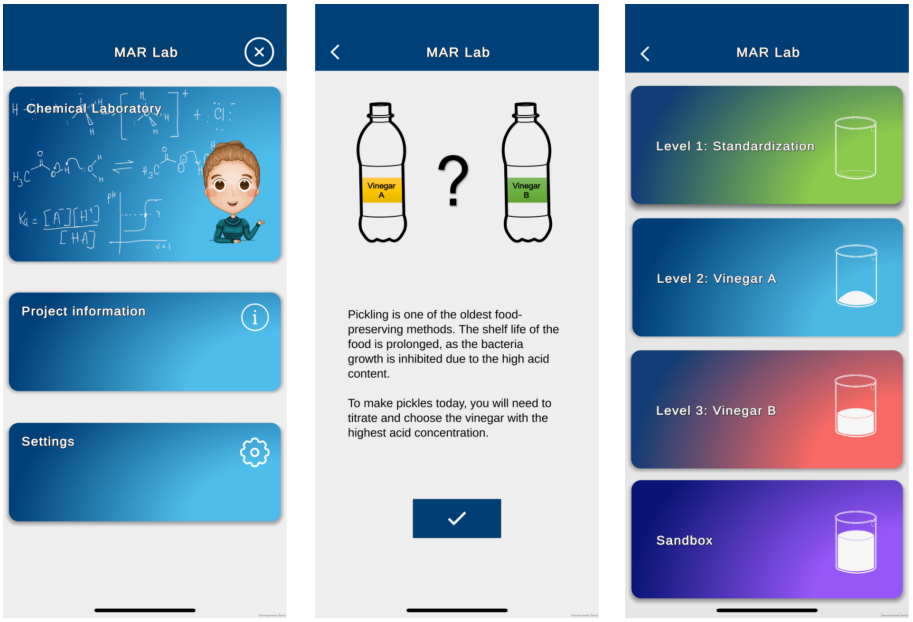
Contextualisation of the learning experience

As discussed in Section 2.3.3, according to Novak's theory [64], one of the critical components for providing meaningful learning to students is allowing them to make connections between abstract concepts learned in chemistry and their experiences in daily life. To facilitate this, the application was re-designed to use a familiar situation (eating pickles) and products (vinegar and pickles) to scaffold a learner's understanding of acid-base titration concepts. For instance, students must select the best vinegar for pickling to complete the application levels successfully. By doing this, students could potentially develop a deeper understanding of the topic, as the concept is now more relatable to their everyday experience.

The user is provided with concrete experiences built upon the same concept along the levels. The complexity of the concepts increases, to some extent, as the levels advance and user support decreases. This variation allows users to advance their learning optimally depending on their previous knowledge and experience (complex learning principles). By maintaining a low extraneous cognitive load and high germane load (through variability), we intend to foster both skill transfer and long-term retention.

The new design shifts its focus from simply completing an aim (e.g., calculating or measuring) to the experience of accomplishing a task (e.g., performing a titration experiment). This design encourages problem-solving and emphasises the role of embodied interaction (sensorimotor experience) in the process of users completing the task [123].

Figure 5.1 displays the introductory screens presented to the user. In Level 1, users are given a task to standardise a solution of titrant to use in later experiments. One purpose of this level is to familiarise students with the



a) b) c)

Figure 5.1: Introductory Screens of MAR Lab: a) the main menu, b) general problem introduction after selecting the "Chemical Laboratory" option, and c) the list of levels.

environment and its interaction. At this level, the indicator is given. In Level 2, users are tasked to titrate a sample of vinegar A. Users need to choose among two different indicators and use a pH meter to track their results. In Level 3, users are tasked to titrate a sample of vinegar B. At the end of Level 3, the user must choose between the two bottles of vinegar to prepare a can of pickles.

The procedural tasks are similar along the levels; therefore, their progress is expected to yield a better performance (measured by different metrics such as time-on-task, number of errors and hints received).

The app flow was modified to make a seamless experience and introduce the problem statement to the users in the first stage. In the first version (MAR Lab 1.0), the user started the experience with an orientation phase in 2D followed by a conceptualisation phase in the AR scenario. In this step, the user answered MCQs and subsequently started the experiment. In the second version of the prototype (MAR Lab 2.0), the introduction and conceptualisation phases are shown to the user in 2D, and the AR scenario begins with the investigation phase (Figure 5.2)

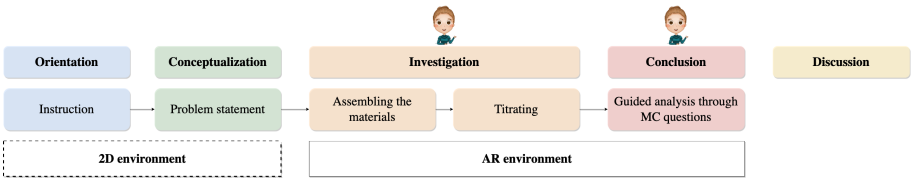


Figure 5.2: The modified flow of the MAR Lab application mapped into the inquiry-based learning cycle. Marie Curie's cartoon indicates which phase of the process she appears in and gives auditory feedback. The first two stages are shown in the 2D environment and the others in the AR environment. The two bottom squares indicate the type of implementation.

In the MAR Lab 1.0 prototype, users had to access the logbook and graph via two different icons on the screen; however, this was simplified in the MAR Lab 2.0 version. The extra icons were removed, and the logbook and graph were integrated into the scenario as one virtual object. This means that users now interact with the logbook and graph with the same actions as with any other object, and the virtual object is always accessible to them (see Figure 5.3). Furthermore, the information presented is in the same space as the other objects, allowing for simultaneous access. The system promotes interaction with the logbook at the beginning and conclusion of the experiment. Doing so allows the user to analyze errors made during the experiment and review the outcomes.

Design elements

Different design elements were improved in the second iteration based on the initial requirements and visual aesthetics [124]. First, the colours of each level were modified to match the menu, screen, and logbook (Figure 5.1.c); this change aims to provide the user with a clear indication of which level they are on. Second, the provisional cartoon character was replaced by a cartoon representing Marie Curie (Figure 5.4). Third, following the modality principle of the 4C/ID model (Section 2.3.5), the voice of Marie Curie was added to the prototype in Dutch, German, and English. The voice was created using the software Descript version 52.1.0 and the online platform Natural Readers[1].

[1] https://www.naturalreaders.com/online

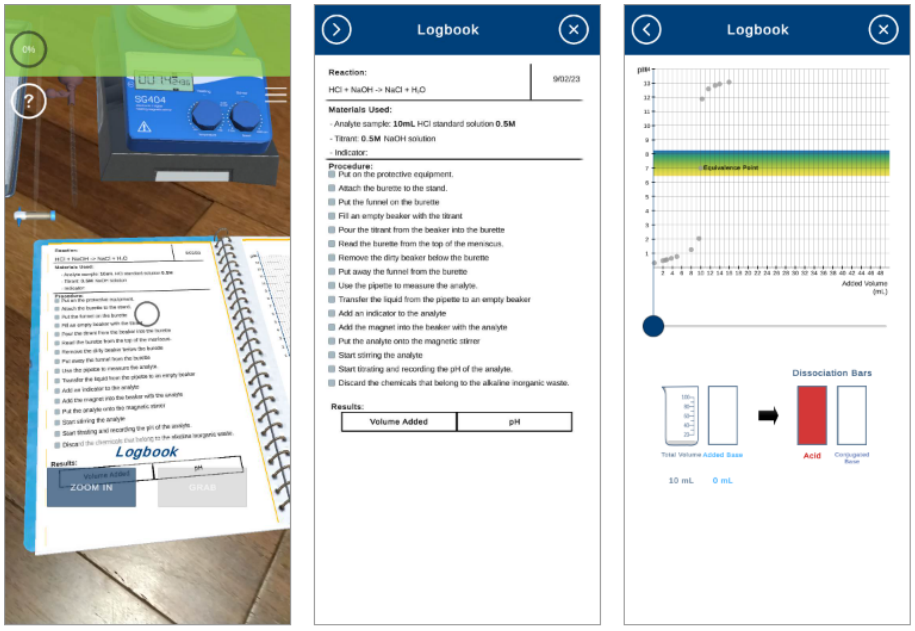


Figure 5.3: Interaction with the logbook: Figure a) depicting the integration of a logbook as a virtual object; the user can interact and zoom in to display both the logbook (b) with the instructions and the graph (c).



Figure 5.4: Cartoon character representing Marie Curie. This image is used on each level when the cartoon provides feedback during the experiment or when an important message is shown to the user.

5.2 Technical Implementation

MAR Lab 2.0 was developed using Unity Version 2021.2. The Android version was distributed through an Android Application Package (APK) file that could be downloaded and installed directly on the devices. A beta version of the app was also submitted to the Google Play Store for testing [2]. For iOS, the app was distributed as a beta version and was submitted to the TestFlight service for testing.

5.3 Process Mining

Hand in hand with developing new learning applications is the evaluation of their effectiveness. One of the most common ways to evaluate these tools is through media comparison studies [125, 126]. However, this perspective is generally limited to the outcome of the learning tests, ignoring the technology affordances and learner characteristics. Therefore, evaluation methods that consider the interaction with the system can be beneficial for designing and developing better learning tools [126]. Process mining (PM) offers a novel approach to analysing each learner's process using a digital system.

PM uses the event logs generated within a system to provide information for discovering, supervising, or improving the processes at hand [127]. In addition, some studies using process mining techniques for educational purposes have explored the interaction with online courses to enhance their quality [128], the learning behaviour in a complex problem-solving process to enhance domain modelling courses [129] or the patterns and strategies along the self-regulated learning process to find patterns among different learners [130, 131]. Overall, these studies exemplify using PM for understanding and enhancing the learning experience while using learning environments. This thesis uses PM to explore how students interact with a mobile augmented reality system.

5.4 Formative Evaluations

The formative evaluations aim to explore how the application's usability changed after the modifications, explore the learning outcomes and get feedback regarding the user experience. A mixed-method approach was taken for the three evaluations. The quantitative analysis was used to determine the learning

[2] MAR Lab 2.0: https://play.google.com/store/apps/details?id=com.CHARMING.MARLab

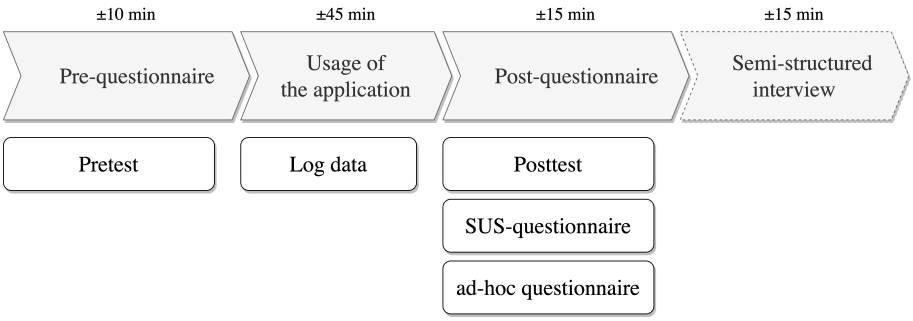


Figure 5.5: Overview of the study design and the time spent on each phase with each participant. The dotted line indicates that the phase was voluntary. The instruments appear in a rounded box below the phases.

effect of the platform's use and measure the application's usability and users' performance. On the other hand, the research uses qualitative analysis to gain detailed insights into the participants' experience with the application.

Three user evaluations were conducted in November 2021, March 2022 and April 2022, respectively. Ethical approval was obtained from the ethical committee at KU Leuven (G-2020-2532(MIN) and G-2022-4989-R2). The characteristics of each group and details of each evaluation are described in the following sections. The prototype implemented minor improvements between each evaluation, such as language translation and fixes on virtual objects.

5.5 Methodology

The evaluations followed a pretest/posttest design, which can be found in Figure 5.5. Before data collection, the participants received an explanation of the project and informed consent. On obtaining written informed consent, a pretest was administered through a web-based survey tool (Microsoft Forms). Subsequently, participants were given a mobile phone with MAR Lab installed and used the application for approx. 45 min. After this, a post-questionnaire was administered, including a posttest, the SUS questionnaire and ad-hoc open feedback questions. In the end, participants joined a voluntary semi-structured interview to share their experiences. The researcher noted the insights from these conversations. For the evaluation with the KU Leuven group (Section 5.6.3), the participants received lunch as a thank-you for their participation. However, they did not participate in an interview.

5.5.1 Instruments

Learning Test A revised version of the chemistry test was created for these evaluations to match the learning goals of the application. After that, the questions were quality checked by an educational scientist and minor adjustments were made. In total, 11 items were formulated, and the back-and-forth-translation method was used to create a questionnaire in the participants' native language.

The questionnaire was piloted in all iterations to validate the instrument. To avoid any potential influence that the improvement of the application may have had on the posttest results, the validation of the learning test was based solely on the pretest results of the first evaluation done in ACTA, resulting in a combined sample size of $n = 26$. The items were checked for quality, difficulty [132, 133] and item discrimination index ($d = 0.13 - 0.50$) to avoid possible ceiling or floor effects.

In total, three items were revised due to their marginal to poor item discrimination index ($d = 0.00 - 0.25$), while three items were either very difficult ($p < 0.41$) or very easy ($p > 0.60$). However, since the item discrimination index of those items was in an acceptable range ($d = 0.38 - 0.50$), they were kept in the final questionnaire. The final Cronbach's alpha is based on the KU Leuven test results. The pretest and posttest showed acceptable reliability, with α scores of 0.68 and 0.45, respectively, as the instrument measured distinct knowledge facets and not a coherent construct [134]. All versions of the instrument can be found in the Appendixes C.2.1, C.2.2 and C.2.3.

SUS questionnaire Similarly to the first user study, the second prototype's usability was evaluated using the SUS 10-item standardized questionnaire. The instrument was administered in German for the Merck test and in Dutch for the ACTA and KU Leuven tests.

User Feedback User feedback was collected using open-ended questions and semi-structured interviews. NVivo software (version 1.7.1) was used to analyze the responses using inductive thematic analysis. This type of analysis allows responses to be classified and the topics to be summarized in a structured way. Unlike other coding frames, the inductive thematic analysis does not require the responses to fit a pre-existing structure [135]. For the data analysis, the questions from the interview and the ad-hoc questionnaire were used to structure the data. For the KU Leuven test, only the open-ended questions were included since the users did not participate in an interview with the researchers. The questions asked during the semi-structured interviews are in Appendix C.3.

5.5.2 Data Analysis

To answer the research question *Do participants improve their learning outcomes after using the application?*, the following hypotheses were formulated:

- H_0 (null hypothesis): No significant difference exists between the mean of the pretest and the posttest.

- H_1 (alternative hypothesis): The mean of the posttest scores is significantly higher than that of the pretest.

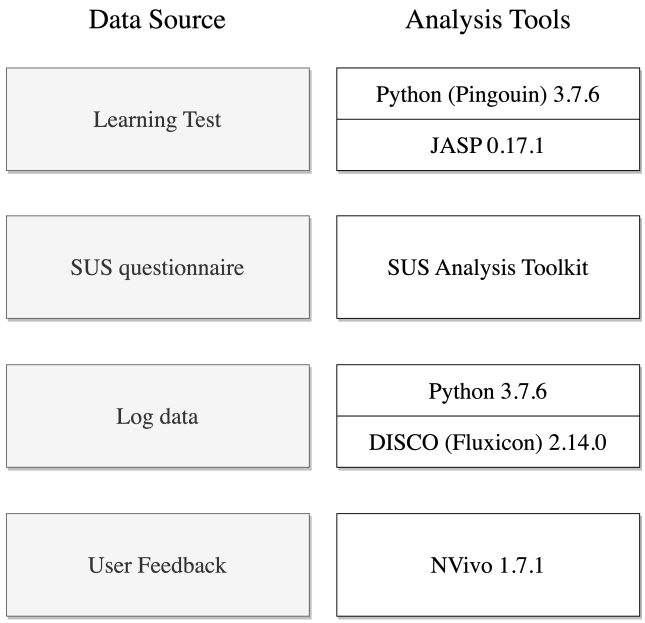


Figure 5.6: Four distinct data sources were gathered and analyzed utilizing multiple software tools. The grey squares illustrate the data sources, and the white squares indicate the software and the respective version used.

Figure 5.6 summarizes the different data sources and tools used for the analysis. Data analysis and management were conducted using Python library version 3.7.6 for descriptive statistics, the package Pingouin [136], and the software JASP 0.17.1 [137]. Depending on the data type distribution, statistical significance

was determined by either t-tests or its non-parametric equivalent. To explore the user's perception of the application, thematic analysis was conducted using the NVivo 1.7.1 software. To assess the prototype's usability, the open-source SUS Analysis Toolkit [119] was employed. Finally, a process mining method was used to explore the problematic steps for users, which is described in the following section.

Process Mining Methodology

This research used DISCO (Fluxicon) version 2.14.0 as the process mining tool to analyse the sequence of events generated by the learners' actions. The software employs the fuzzy miner algorithm to calculate the relevant activities to include in the process map [3]. The log data were pre-processed and analyzed as follows.

First, the raw file from Unity Analytics was downloaded as a JavaScript Object Notation (JSON) file format and converted to a comma-separated value (CSV). Second, the data was cleaned, the timestamp formatted, and incomplete event cases were removed. Then, the CSV file was imported to DISCO. On import, each of the file columns was assigned to a variable. The case (i.e., a process instance) was assigned to the id given to the user, the activity (i.e., a defined step in the process) was assigned to the custom event (e.g., "Done Process"), and the timestamp (i.e., the time of occurrence) was assigned to the same variable logged in Unity. Finally, the filter option inside DISCO was used to visualize the process map to include the events of interest. Moreover, the level of detail (i.e., % activities and paths %) was adjusted to the participants' most frequent activities and paths.

Each of the events presented in Table 5.1 can be selected in DISCO to compose a process map; however, the more activities are included, the more complex the process map becomes. In this chapter, three events (Table 5.1) were combined in the same process maps filtering the activities in DISCO to identify potential difficulties with the content and the system. Specifically, "Done Process," "Combine Wrong," and "Reposition Materials."

Log Data

In this version of MAR Lab, additional events were added to record different activities, such as the start or end of a procedure and fine-grained activities (i.e.,

[3] https://fluxicon.com/disco/

grabbing, placing, and combining the laboratory objects). Table 5.1 presents all the events logged by the application.

Table 5.1: Custom events registered by Unity Analytics . A custom event is triggered when the user performs an action or activity. Each event includes different parameters, which are extra information associated with the event.

Custom Events	Activity	Parameters
Combine Wrong	Combine two objects	Name of objects
Interact	Interact with an object	Correct, Wrong
Grab	Grab an object	Name object
DoneProcess	Task completed	
Open Side Menu	Open the menu	
Reposition Materials	Change the position of the virtual set-up	
Show Hint	Trigger a hint after three trials	Hint number, step

5.6 Results

5.6.1 ACTA Results

This section summarizes the results of the test performed at ACTA VZW. The analysis is done concurrently to incorporate both quantitative and qualitative data.

Participants

The evaluation took place in the plant of ACTA VZW, located in Antwerp, Belgium, in November 2021. ACTA VZW is a Belgian training centre that specialises in providing training for the chemical industry. In the CHARMING project, ACTA was one of the beneficiary institutions. Several sessions were organised in one week. The participants were recruited via ACTA's employees voluntarily. The content of the application and the questionnaires in this test were administered in English. The feedback session was done in English with the support of a native Dutch speaker.

The study involved 27 participants from two different backgrounds. Nineteen of them were students of the Apprenticeship Program at ACTA, and they are referred to as "ACTA operators" in the text later on. This group had limited experience in the laboratory. The age range of these students was between 18 and 24, consisting of 18 males and one female. Additionally, eight high school students were recruited by ACTA and invited to the company to take part in

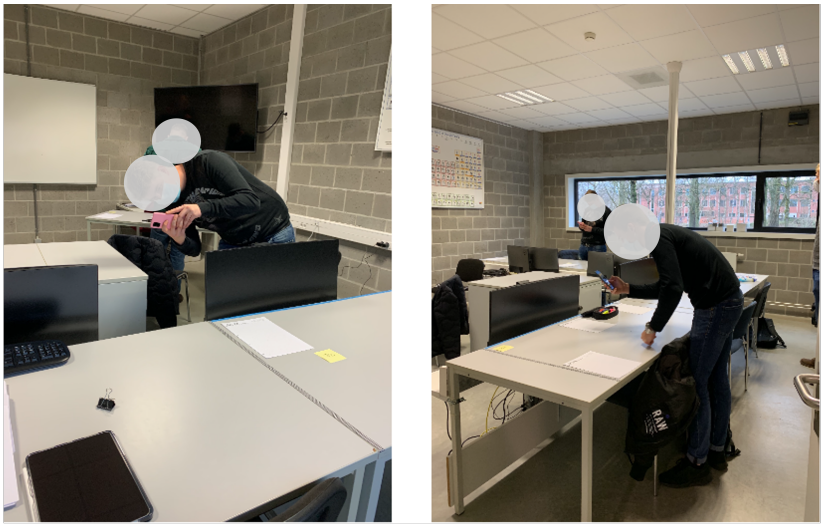


Figure 5.7: Example of multiple participants using MAR Lab during the ACTA Test.

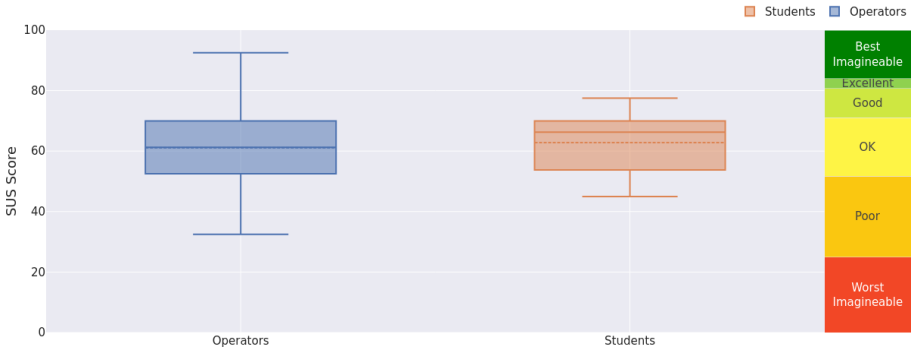
the study, with a tour included. Of this group, seven were male, one was female, four participants were under 18, and four were between 18 and 24. During the evaluation, the teacher accompanying the group noted that the participants had recently learned about titration. One participant's data were excluded from the analysis due to incompleteness; thus, the dataset consists of 26 participants.

Observations during the study

The user test was organized in four sessions with a maximum capacity of ten participants per time slot. To ensure COVID regulations, sufficient space was allocated between tables. Each participant was given a grey table, an ID number, informed consent, and a Samsung Galaxy S10/S20 Android mobile phone pre-loaded with the application under evaluation - MAR Lab 2.0. During the test, participants mainly stood up and worked independently, although they could ask for assistance from researchers or peers when needed. Seating was only used when answering the questionnaires. Figure 5.7 shows an example of multiple participants using the application.

Prototype's Usability

The average scores reported by operators and students, depicted in Figure 5.8, were 61.11 and 62.81, respectively, with standard deviations of 15.05 and 10.19. While the scores are "ok" according to Bangor et al. [118], they are below the industry benchmark score of 68 defined by Lewis et al. [138], thus classifying the prototype as "below average". A solid line represents the median value, while a dotted line represents the mean value.



Variable	SUS Score (mean)	SD	Adjective Scale	Acceptability Scale	Industry Benchmark
Operators	61.11	15.05	OK	Marginal	Below Average
Students	62.81	10.19	OK	Marginal	Below Average

Figure 5.8: System Usability Scale results contextualized on the scale proposed by Bangor et al. [118]. The mean SUS score for the operator's group is 61.11 (SD=15.05) and for the group of students is 62.81 (SD=10.19). According to the scale, the prototype has "OK" usability on the adjective scale. The figure was created using an open-source toolkit [119].

The normalized values of the System Usability Scale (SUS) questions, ranging from 0 to 10, presented in Figure 5.9, indicate how much each item contributes to the overall score. A high normalized value indicates that a question significantly impacts the score, with users recognizing and agreeing upon its importance. Conversely, a low contribution to the score implies that there is an opportunity to improve these aspects to increase their contribution and, as a result, boost the overall score.

The analysis of SUS scores revealed that the operators' group had the highest ratings for questions 4, 5, and 7 (with normalized values of 6.67, 7.08, and 6.94, respectively), which were related to system efficiency, effectiveness, and learnability. Similarly, the students' group rated questions 1, 7, and 10 as having the greatest influence on their overall score, with normalized values of 7.19, 8.44, and 7.5, respectively. These questions were linked to overall satisfaction with the system, learnability, and system efficiency.

For the operators' group, questions with the lowest impact on the SUS score were 1, 6, and 9 (normalized values of 5.42, 5.28, and 4.86, respectively), suggesting that operators generally perceived the system as inconsistent and lacked confidence while using it. For the students' group, questions 6, 4 and 8 had the weakest impact on the SUS score, with normalized values of 5.62, 4.69 and 4.38. These questions indicated that students found the application inconsistent and cumbersome to use.

In comparing the results, both groups valued learnability and the fact that they would not need support from a technical person or learn many things before using the application. At the same time, students additionally rated user satisfaction as a critical factor in their SUS score. Conversely, aspects related to consistency and perceived ease of use can be targeted to improve the overall usability score.

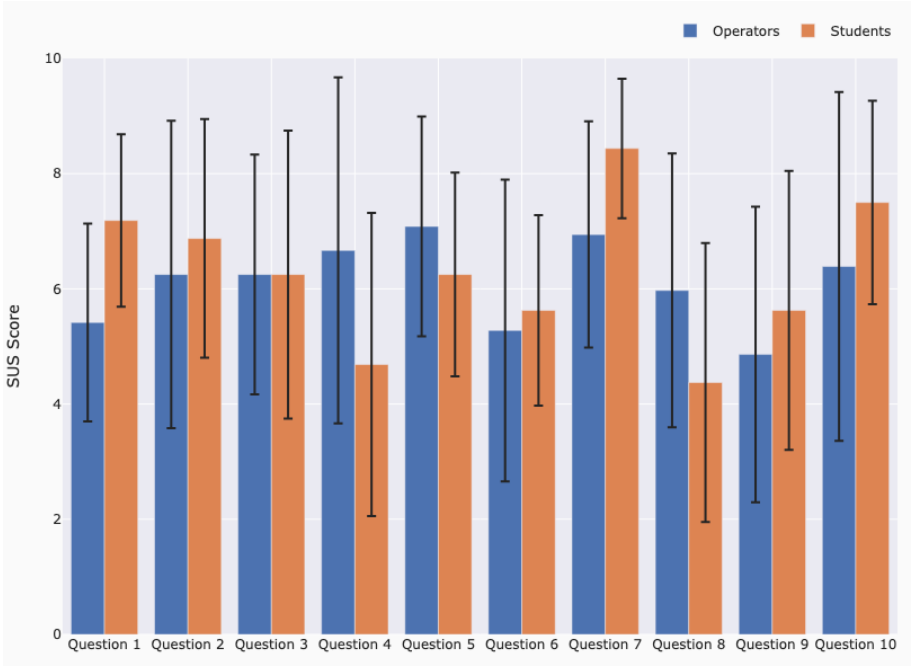


Figure 5.9: Normalized values ranging from 0 to 10 for each item of SUS and the general SUS questions for ACTA test. The horizontal line indicates the standard deviation associated with each item. The figure was created using an open-source toolkit [119]. The SUS questionnaire administered in this evaluation can be found in Appendix C.2.1.

Table 5.2: Problems reported by 26 ACTA participants

Codes	Number of mentions
Tracking	15
Selection	8
Crash	7
Materials disappearing	6
Floating objects	4
User interface	4
Overheat	3

User Feedback

During the ACTA test, the researcher conducted semi-structured individual interviews with 85% (22) of the participants in the sample, exploring their experience and use of AR. Using inductive thematic coding, the issues users experienced with MAR Lab and their remarks about design elements and user experience, positive traits, and suggestions for improvement were identified and organized into broad themes. These themes were then used to generate codes, which are summarized below (see Appendix C.1). Mentions of these themes were counted and reported if they occurred multiple times; otherwise, the individual observations were noted.

Problems The most commonly reported problems with the application are presented in Table 5.2. The thematic analysis showed that tracking was the most widely reported issue, with 15 mentions. Users described this problem in various ways, such as "A lot of times when you try to grab an object, your whole working station shoots away" and "The camera often loses where everything should be, so I have to move back up to find the three reference points". Crash and selection were the next two most common issues, with 7 and 8 mentions, respectively. These issues included "I couldn't finish because it froze at the same point each level" and "Hard to select objects close to each other, sometimes selected not the one you wanted". The next three most prominent issues were materials disappearing, floating objects and user interface issues, with 6, 4, and 4 mentions, respectively. An example of the materials disappearing was given by a user who stated, "my beaker went missing during the experiment so that I couldn't complete it", and for the floating objects, a user reported, "the equipment was sometimes floating and needed to be repositioned". User interface issues included problems related to unresponsive buttons or 2D elements that did not disappear after clicking the options.

Suggestions Users provided a variety of suggestions to improve the application, including translating it to the native language, modifying the design for sitting use, adding voice commands and reducing audio from every step, incorporating implicit interactions such as double taps, making procedural improvements like selecting the correct protective equipment or cleaning the pH probe, automatically zooming in on objects when the user gets close, and adding interactions such as closing and opening all bottles and the burette.

Learning Experience Users commented on their experience with the application. Many compared it to their real-world experience in a laboratory, with one user saying, "I remember my own experience in the laboratory." Others expressed what they learned from the application: "I learned something new, one indicator I didn't know existed", and "I had no idea what titration was. I have an idea now, but basic." Some users expressed difficulty with tasks such as waste sorting, putting the funnel in the burette, and choosing the indicator. Others noted limitations, such as difficulty following the experiment in English or learning in a real environment. There were also opposing opinions, such as "Long time ago performed a titration. I think it is good to use for other people, though; it's a good alternative to the real titration." Finally, users with no prior experience with titration expressed confusion about aspects of the experiment, such as its objective or the equivalence point, with one user saying, "I would have liked some explanation as to what the experiment was about since no chemical background."

General User Experience Users shared their overall experience with the application, highlighting the positive and negative aspects. Praise was given to the logbook, hints, colour change, realism, feedback sounds, names of the objects, and ease of use. For instance, one user remarked that the hints were "really nice" and that the sound it made when a wrong move was made was "helpful". Another praised the realistic colour shift at the equivalence point, noting that it "doesn't just change colour, but it goes in shifts; it's like the real experiment". On the other hand, users identified issues such as specific terms, repetitive audio, tracking problems, language and materials disappearing, which caused frustration and hindered the user experience. For example, one user found the setup moving "a little annoying", and another noted that "flying objects occurred often, which interrupted the learning". These comments were in line with the problems previously identified.

AR Experience In total, eight users reported having prior experience with AR, with the most frequent application being Pokémon Go. One user expressed

astonishment at the level of detail that AR can provide, saying, "I only knew Pokémon Go and only sliding in the screen, but I didn't know that AR could be so detailed." The other participants reported MAR Lab as the first AR app they had ever tried.

Learning Test

Due to the exploratory nature of this evaluation, these results should be interpreted with caution. The instrument underwent modifications, with each item assessed for quality, difficulty [132, 133], and item discrimination index ($d = 0.13-0.50$). Three items were revised due to suboptimal item discrimination ($d = 0.00 - 0.25$), while three items were either too difficult ($p < 0.41$) or too easy ($p > 0.60$). Nevertheless, for completeness, the learning test results are presented in this section.

The results of the operators' group showed a mean pretest score of 5.94 ($SD = 1.43$) and a mean posttest score of 7.39 ($SD = 1.85$) (Figure 5.10). The Wilcoxon Signed Rank Test with an alpha risk of rejection of 0.05 provided sufficient evidence ($W = 0$, $p < 0.001$, $r = -1.0$) to reject the null hypothesis and suggest an improvement on the posttest scores of the operators. The non-parametric test was used as the data did not follow a normal distribution (Appendix C.3). The effect size expressed in CLES indicates a 71% chance that the evaluation of the posttest is higher than the pretest evaluation when a pair of scores is randomly selected (Table C.2).

The students' group had a mean pretest score of 8.0 ($SD = 2.2$) and a mean posttest score of 7.12 ($SD = 1.64$) (Figure 5.11). A one-tailed t-test at a significance level of $p < 0.05$ was used as the data followed a normal distribution (Appendix C.6); however, the result ($t(7) = 1.313$, $p = 0.885$, $d = 0.46$) indicates that the null hypothesis could not be rejected; thus the mean of the pretest and the posttest are not significantly different. Moreover, the effect size expressed in CLES indicated that it is only 37.5% probable that the evaluation of the posttest is higher than the evaluation of the pretest when a pair of scores is randomly selected [121, 122].

The results of these tests are subject to limitations. The instrument was validated using the results of this test, meaning that the reliability and accuracy of the instrument itself limit the results. Additionally, both groups had relatively small sample sizes, limiting the test's power. Despite this, the test was still relevant and valuable to improve the instrument and the application as a whole.

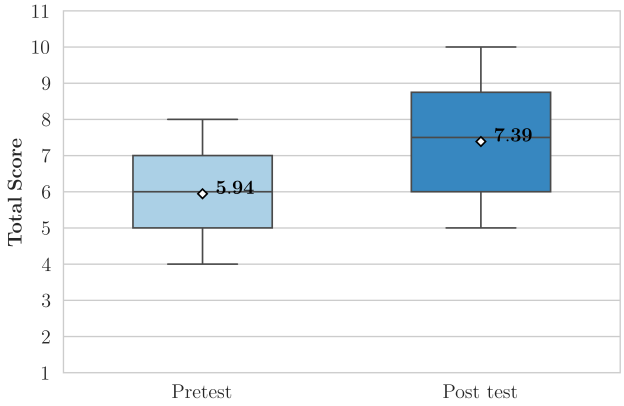


Figure 5.10: Boxplot showing the pretest and posttest scores of the ACTA operators group. The boxes represent the interquartile range (IQR), with the median of the scores indicated by the line inside the box and the mean scores indicated by the rhombus. The pretest had a mean score of 5.94 ($SD = 1.43$), while the posttest had a mean score of 7.39 ($SD = 1.85$). The Wilcoxon Signed Rank Test results showed a significant difference between the two tests ($W = 0$, $p < 0.001$, $r = -1.0$).

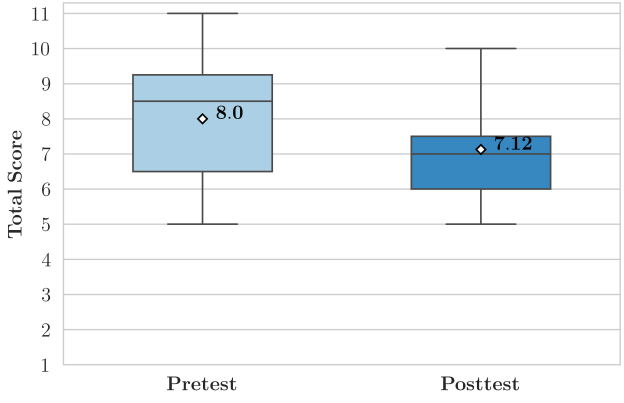


Figure 5.11: Boxplot showing the pretest and posttest scores of the ACTA students group. The boxes represent the interquartile range (IQR), with the median of the scores indicated by the line inside the box and the mean scores indicated by the rhombus. The pretest had a mean score of 8.0 ($SD = 2.2$), and the posttest had a mean score of 7.12 ($SD = 1.64$). The one-tailed t-test showed no significant difference between the two tests ($t(7) = 1.313$, $p = 0.885$, $d = 0.46$).

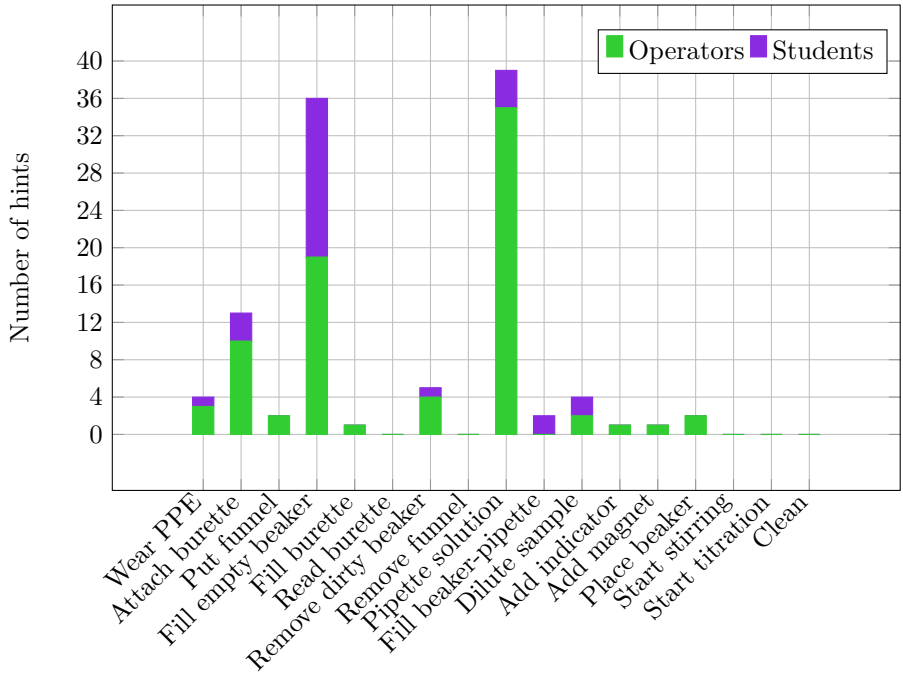


Figure 5.12: Number of hints per step given to ACTA Operators and Students in Level 1. A hint is triggered after three unsuccessful trials. The graph shows the steps in the sequential order observed in the titration process of the app.

Log Data and Process Maps

Analysis of the data logged in the ACTA test uncovered potential issues and bugs when using the system. Notably, only Level 1 could be registered in this test due to a restriction on the number of events that can be logged with Unity Analytics. However, this limitation was increased for subsequent tests.

To better understand where users struggled, a custom event, *Show Hint*, was analysed in detail. The number of hints shown to the user per step for each group (students and operators) was compared, as shown in Figure 5.12. The results indicated that users in both groups received more hints on steps "attach burette", "fill empty beaker", and "pipette solution".

Process maps were created to explore the process that each group takes when using MAR Lab (Figures 5.13 and 5.14). A combination of 100% activities and 35% paths was selected to maximise visibility and ensure the map's simplicity. Three custom events, *Done Process* (blue boxes), *Combine Wrong* (red box),

and *Reposition materials* (green box), were included in the maps. For each event, two metrics were presented; the number inside the box indicates the case frequency and is supported by the thickness of each black arrow. The number from one arrow to the next shows the mean duration between the two processes and the case frequency.

Done Process The custom event *Done Process* was investigated to explore the order of tasks performed by the users. This event is recorded when the user completes a step in the procedure. By examining the process maps, it is possible to identify which stages of the process users perform actions in a different order from what the procedure instructions indicate. This can be identified by the diverging arrows originating from the same event. For example, Figure 5.13 shows that 15 participants completed the step "remove dirty beaker" after "read burette". In contrast, the others followed a different path and completed the "read burette" step after "remove the funnel". Similarly, for the ACTA student group, six participants followed the step "remove dirty beaker" after the event, "read burette," and one person chose to "remove funnel" first (Figure 5.13). On the other hand, in the operator's process map, two events were missing ("place beaker" and "clean base"). This was due to all participants in this group quitting the level before completing the last step of the procedure, likely due to a problem with the system (bug). Additionally, only one participant in the student group completed the procedure.

Combine Wrong The *Combine Wrong* custom event identifies after which steps the users make errors when combining objects that are not allowed. For example, as can be seen from Figure 5.13, users from the ACTA operators group made the most wrong combinations before completing the steps "attach burette" (10 cases) and "fill empty beaker" (16 cases). Similarly, for the student group, the events "attach burette" (5 cases) and "fill empty beaker" (6 cases) preceded an incorrect combination. These observations align with the Figure 5.12 that indicates at which step the users received more hints.

Reposition Materials The *Reposition Materials* event is registered when the user clicks the option on the side menu. This event is integrated into the system to allow users to move laboratory objects to a different position when the virtual objects are transferred to another place or the system loses track of the environment. This was deemed an important event to track based on previous results [139], which showed that users reported having problems with the environment "floating" and needing to reposition the setup. Process maps indicated that the *Reposition Materials* event was primarily triggered before

or after "wearing PPE" (at the start of the level) for both groups. Moreover, for ACTA operators, this event was also triggered several times after the "put funnel" step, "fill empty beaker", and "pipette solution" steps.

Summary

The ACTA test included evaluating the application with two distinct groups: operators and high school students. Both groups tested the three levels of the application (MAR Lab 2.0) and provided feedback afterwards. This evaluation only includes log data from Level 1 due to technical issues with logging the data for all three levels.

The results from the ACTA test suggest that users experienced various issues with MAR Lab, such as tracking, crash, selection, materials disappearing, floating objects, user interface issues, and language. Many users provided suggestions for improvement, such as translating the language into Dutch and modifying some steps to make them clearer. Users praised the logbook, hints, colour change during the titration, and feedback sounds. Those with prior experience with AR expressed surprise at the level of detail that AR can provide.

From the log data, the starting step ("wear PPE") was problematic for both groups. Users received hints during this step and triggered the event *Reposition Materials*. These problems can be attributed to the users' unfamiliarity with AR, such as difficulties in tracking the environment in the first place. Additionally, language barriers may have been a factor, as the onboarding instructions were not in the users' native language, which may have made it more difficult to understand. Furthermore, tracking was the most commonly reported issue by users, suggesting that the problem may have been due to the application's technical issues while tracking the environment, which can be affected by lighting conditions, fast movements, and device sensors [46, 45].

The "attach burette", "pipette solution", and "fill the empty beaker" steps were identified as problematic during the process maps and Figure 5.12. The object characteristics could explain the difficulty with the step "attaching the burette". In this case, the user needs to precisely target the narrow object (i.e., burette) towards the stand, which may require some accuracy. This corroborates with users mentioning difficulties in selecting some objects. The "pipetting the solution" problem can be attributed to the user having to remove the container's lid before interacting with it, which was not explicitly stated in the scenario and could have caused confusion. Moreover, the beakers were hardcoded into the application, so the user had to figure out which container was meant for the analyte and which was for the titrant. This would not be the case in a real

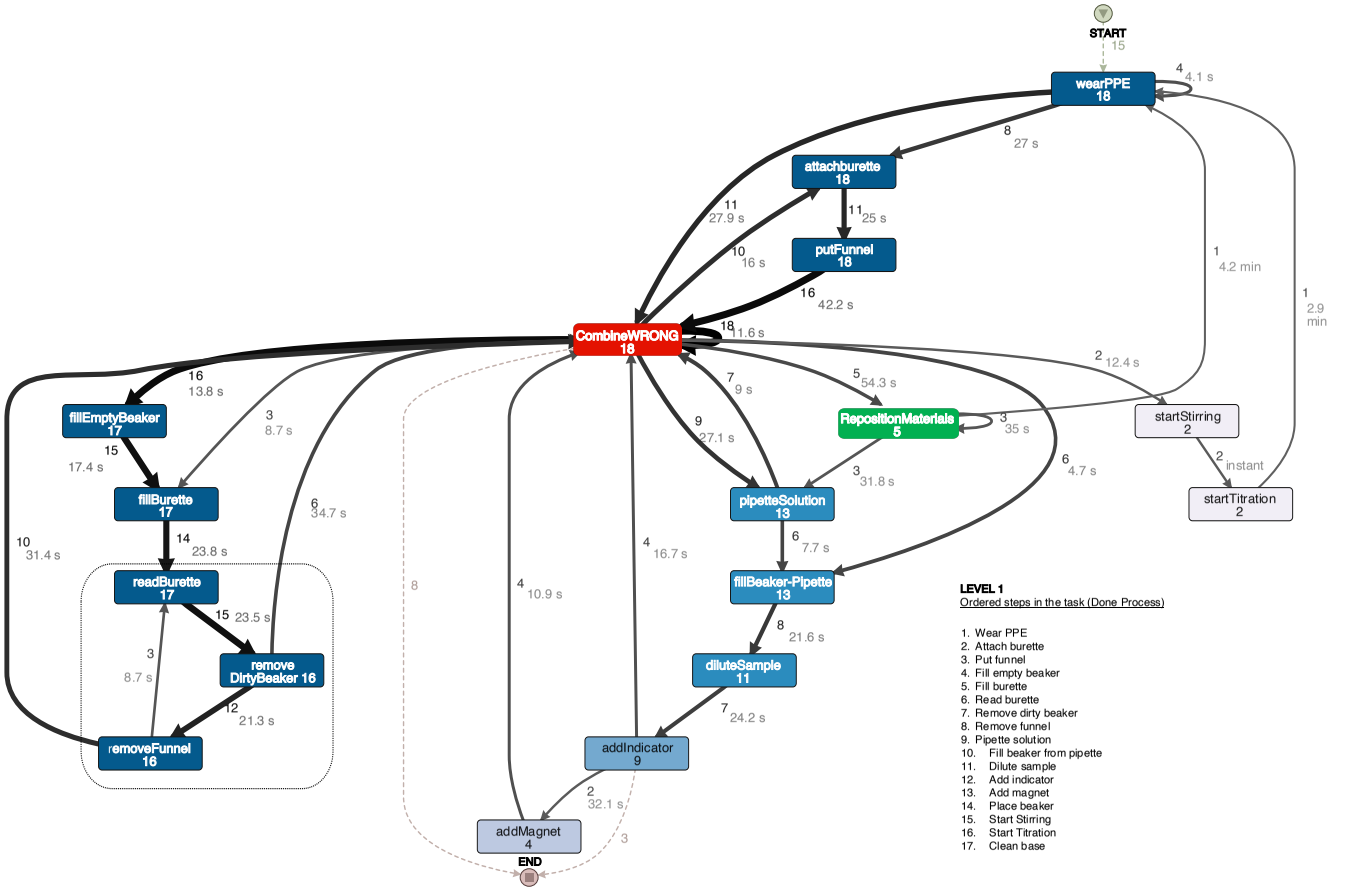


Figure 5.13: Operators' process maps showing (according to DISCO metrics) 100% activities and 35% of paths. The events logged include *Reposition Materials* (green box), *Combine Wrong* (red box), and *Done Process* (blue boxes). The divergent paths are highlighted with a box. Frequencies of events and duration are indicated by the numbers beside the arrows.

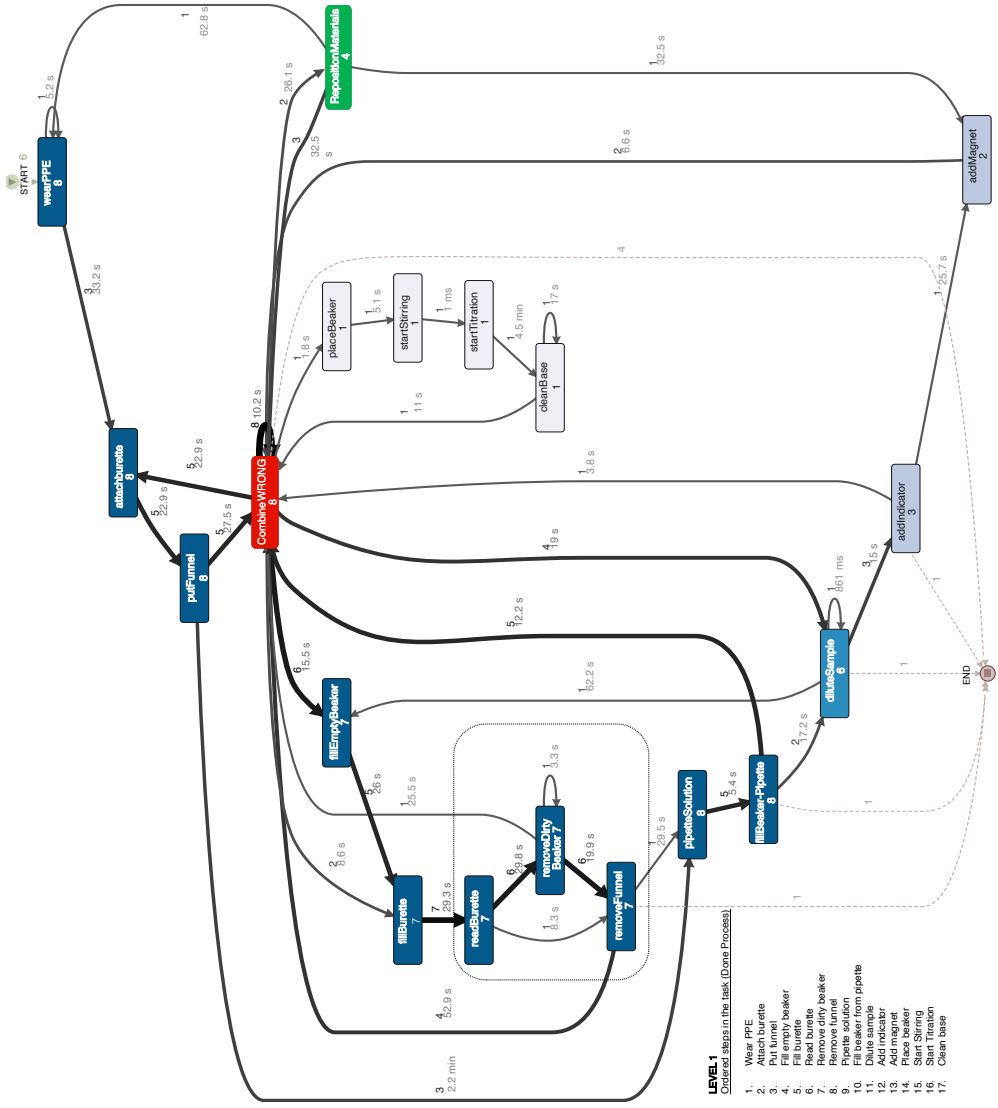


Figure 5.14: Students' process maps showing (according to DISCO metrics) 100% activities and 35% of paths. The events logged include *Reposition Materials* (green box), *Combine Wrong* (red box), and *Done Process* (blue boxes). The divergent paths are highlighted with a box. Frequencies of events and duration are indicated by the numbers beside the arrows.

laboratory, as the user could choose whatever container for whatever action. It is desirable to make the necessary actions more explicit and make the container selection more flexible to improve these steps.

Another interesting finding is that some participants followed different paths in certain steps. Notably, users from both groups usually chose a different sequence of steps after "reading burette". This suggests that although the system had given instructions, the participants still made their own decisions during the process.

As the learning test was used as a pilot to validate further findings, its results should be interpreted cautiously and viewed as merely indicative. The learning test results showed a statistically significant increase in learning among the operators, with a moderate effect size. This could be attributed to the operators having limited prior experience with titration. On the other hand, the results for the student group were not statistically significant, which could be due to their higher level of prior knowledge of titration, a small sample size, and both groups taking the test in English rather than their native language.

Implications of these results include improving user experience with MAR Lab for the next iteration by considering user needs and preferences. This includes translating instructions into users' native languages and making sure that the tracking and selection of objects are properly functioning. Additionally, it is important to make the selection of objects and containers more flexible, so users can take the initiative to complete tasks in the way that makes the most sense to them.

5.6.2 MERCK Results

Following an ACTA evaluation, an enhanced version of MAR Lab was created (MAR Lab 2.0.1). The changes included adding a German language version, resolving the event limit issue in Unity Analytics, and introducing the ability to open and close some objects (i.e., titrant flask and analyte bottle).

Participants

The evaluation took place in the plant of Merck KGaA, located in Darmstadt, Germany, in March 2021 for three days. Merck KGaA is a multinational chemical company, and in the CHARMING project, it was one of the beneficiary institutions. The participants were recruited via their daily trainer voluntarily. The content of the application and the questionnaires were administered in German. In addition, the feedback session was done in English with the support

of a native German speaker. In total, eleven students and three trainers participated in Merck's Apprenticeship Program. By the time of the evaluation, the students had completed their chemical laboratory training, where they learned about titration.

To ensure accuracy, data from seven participants were analyzed for learning and process maps. Four participants were excluded due to missing log data or incomplete questionnaires. The seven participants, consisting of three females and four males, were between the ages of 18 and 24. The trainers' data were not analysed as they did not complete the levels and learning test, but their interview feedback was still reported.

Observations during the study

The user test was organized in four different sessions with a maximum capacity of four participants per timeslot. To ensure COVID regulations, sufficient space was allocated between tables. Each participant was given a white table with a cross-shaped tracking marker, an ID number, an informed consent form, and blank papers. Participants stood up and worked independently during the test, occasionally asking the researchers questions. Seating was used primarily for answering the questionnaires. Throughout the session, a native German speaker was present if the participants needed assistance. Figure 5.15 shows an example of multiple participants using the application while adhering to social distancing guidelines.

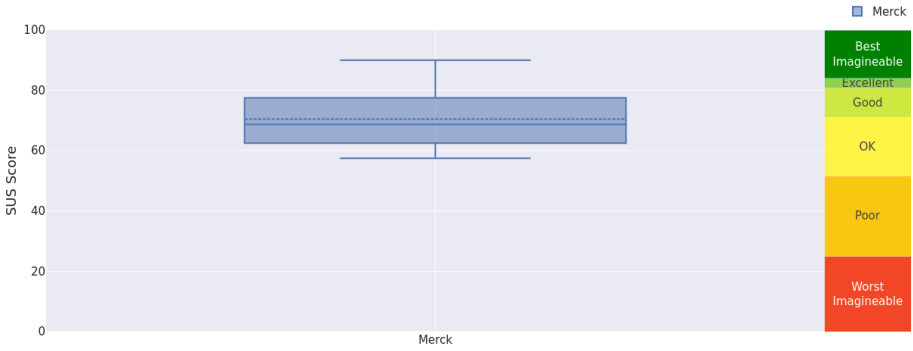
Prototype's Usability

The SUS questionnaire yielded an average score of 70.5 with a standard deviation of 9.6 for Merck participants. According to Bangor et al., [118], it is considered "ok" and indicative of marginal usability. As shown in Figure 5.16, the median of this score is represented by a solid line, and the mean value is denoted by a dotted line. Compared to the industry benchmark defined by Lewis et al. [138], the prototype is classified as "above average".

The results of the SUS score analysis in Figure 5.17 revealed that questions 2, 7, and 10 had the greatest influence on the Merck participants' overall score, with normalized values of 7.5, 8.75, and 8.25, respectively. These questions focused on the system's perceived complexity, learnability, and efficiency. Questions 1 and 6, meanwhile, had the lowest contribution to the overall score, at 5.25. These questions relate to user satisfaction and system inconsistencies.



Figure 5.15: Multiple participants using MAR Lab during the Merck Test at individual tables equipped with a cross-shaped tracking marker.



Variable	SUS Score (mean)	SD	Adjective Scale	Acceptability Scale	Industry Benchmark
Merck	70.5	9.6	OK	Marginal	Above Average

Figure 5.16: System Usability Scale results contextualized on the scale proposed by Bangor et al. [118]. The mean SUS score is 70.5 (SD=9.6). According to the scale, the prototype has "OK" usability on the adjective scale. The figure was created using an open-source toolkit [119].

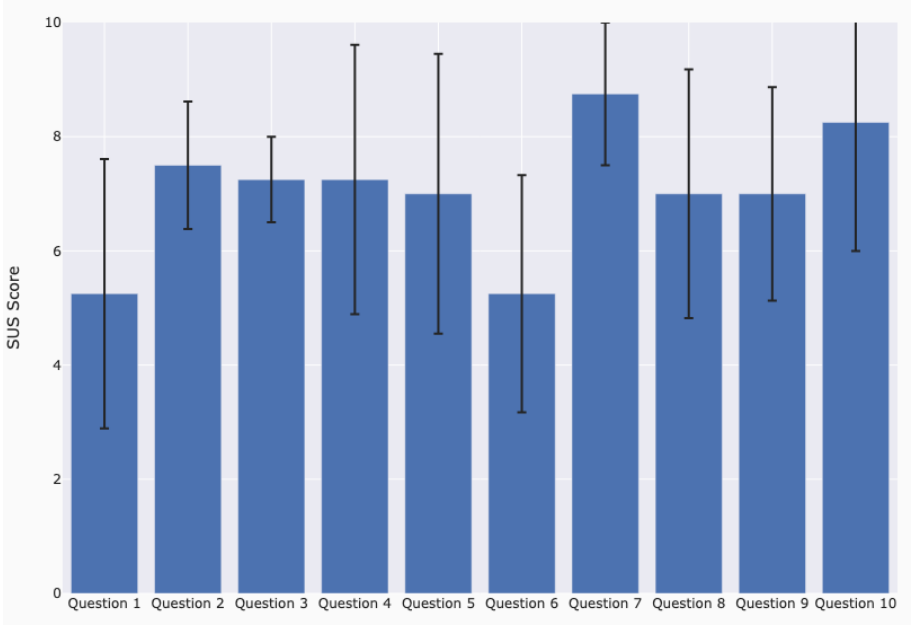


Figure 5.17: Normalized values ranging from 0 to 10 for each item of SUS and the general SUS questions for the Merck test. The horizontal line indicates the standard deviation associated with each item. The figure was created using an open-source toolkit [119]. The SUS questionnaire administered in this evaluation can be found in Appendix C.2.2.

User Feedback

To collect insightful details on the user experience of the participants at Merck, a group semi-structured interview with the students of the different timeslots of the test was performed. During the interview, a native German speaker was present to assist the researcher. Through inductive thematic coding, issues with the MAR Lab were revealed, as well as positive attributes and improvement ideas.

Problems The results of the common issues experienced by Merck's users are presented in Table 5.3. The most frequently reported issues were materials disappearing (6 mentions) and floating objects (5 mentions), as evidenced by user comments such as "items got lost", "eyedropper not found again", and "some objects either float or stick on the air". Other reported problems included user interface (3 mentions), with the responsive buttons being the most mentioned by users, as well as specific materials (4 mentions), with comments such as

Table 5.3: Problems reported by seven Merck participants and three trainers

Codes	Number of mentions
Materials disappearing	6
Floating objects	5
Problem with specific materials	4
User interface	3
Crash	2
Tracking	2
Translation	2
Overheat	1

"burette and pipette had weird behaviour" and "the burette did not work on the first attempt". Additionally, crash (2 mentions), with users expressing, "app hung up, and the attempt had to be restarted", tracking (2 mentions), with a user saying "the workstation often 'slipped' and had to be repositioned", translation (2 mentions), with mentions related to translation mistakes leading to confusion, as one user expressed, "in some places, the translation was not quite correct, which led to some confusion", and overheat (1 mention), with a user expressing "the phone got really hot" were also reported.

Suggestions Participants also made various suggestions for improvement, like adding new experiments such as organic synthesis of a compound, distillation, filtration, or measurement with a spectrometer; simplifying the procedure by removing the "diluting the sample" step; reducing the instructions for subsequent levels to allow for more freedom. One user also suggested modifying the layout of the materials as "don't have enough space to put everything on the desk", and two users suggested improving the procedure by including pH meter calibration and pipetting rules.

Learning Experience Participants expressed both positive and negative feedback regarding the learning experience. On the positive side, participants shared how MAR Lab can benefit students. One user suggested that the experience is "good as an introduction to getting familiar with the equipment." The trainers noted that the app could be helpful for students who lack access to laboratory equipment to understand the general function of titration and the order of steps. Additionally, other participants said the application could be helpful for students to remember the workflow; for example, one user expressed, "nice reminder in the laboratory to use the googles".

On the other hand, participants also provided negative points where the experiment was unclear or inaccurate. For instance, one user mentioned, "the colour change of the phenolphthalein was very abrupt. The disappearance of the colour near the equivalence point was not present." Another user suggested making the reflective intention of the quiz explicit, as it was not initially clear. Finally, they proposed that it would be better to rely on one element (either a probe or an indicator) to identify the pH, as is standard in the laboratory.

General User Experience Participants shared their overall experience with the application; they commented that the app provides a realistic experience with colour changes, hints and intuitive controls, making it easier to understand the procedure. Users praised the sorting waste step, logbook, the fact that they moved physically and remained busy, wearing PPE and the function of "repositioning materials". On the other hand, they disliked problematic materials such as burettes, pipettes, and the floating set-up. Overall, the application was perceived by users as intuitive, fun to use, and simple to show the titration process by providing different examples and levels.

AR Experience Only two participants reported having previously experienced AR applications, with one of them citing the IKEA app as an example.

Learning Test

The mean pre-test score was 8.86 ($SD = 1.46$) and the mean posttest score was 8.29 ($SD = 0.95$) (Figure 5.18). The non-parametric Wilcoxon Signed Rank Test was performed, with a significance level of $p < 0.05$, but did not sufficiently reject the null hypothesis ($W = 21$, $p = 0.912$, $r = 0.5$). Therefore, an increase in posttest scores is rejected for this evaluation. The effect size (CLES in Table C.8) revealed that it is only 31.6% likely that the posttest evaluation is higher than the pre-test evaluation when a pair of scores is randomly selected. The results imply that the platform did not significantly affect learning, and the potential for improvement may have been limited (i.e., a ceiling effect) by the high pre-test score and the small sample size.

Log Data and Process Maps

The analysis of the log data and process maps was processed per level for the Merck Test. The in-depth analysis of the *Show Hint* custom event was performed. Comparing the number of hints received by users on each step and

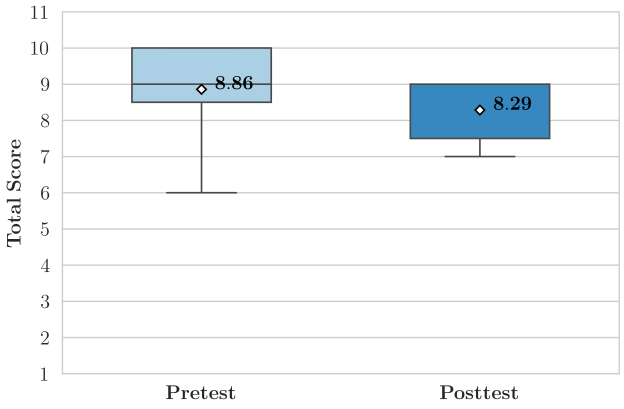


Figure 5.18: Boxplot showing the pretest and posttest scores of the Merck group. The boxes represent the interquartile range (IQR), with the line inside each box depicting the median scores and the rhombus indicating the mean scores. The pretest and posttest scores had respective means of 8.86 ($SD = 1.46$) and 8.29 ($SD = 0.95$). The Wilcoxon Signed Rank Test results showed no significant difference between the two tests ($W = 21$, $p = 0.912$, $r = 0.5$).

level (as shown in Figure 5.19) suggested that users received more hints in Level 1 and on steps "wear PPE", "fill empty beaker", and "place beaker". In Level 3, users did not receive hints. It is important to note that in Levels 2 and 3, the hints appear until after 5 or 7 unsuccessful trials.

Done Process The custom event *Done Process* was investigated to explore the order of tasks performed by the users. From Figure 5.20, it can be observed that in Level 1, 3 users followed the instructions (i.e., after "fill burette" they performed the step "read burette") while the other four users took a different path. In Levels 2 and 3 (Figures 5.21 and 5.22), a similar divergence was observed. Participants followed a different route after the step "fill burette".

Combine Wrong The custom event *Combine Wrong* is triggered when the user attempts to combine incompatible objects. In Merck, participants in Level 1 (Figure 5.20) were most likely to make incorrect combinations before the steps "wear PPE" (5 cases), "fill empty beaker" (7 cases) and "place beaker" (6 cases). In Level 2 (Figure 5.21), wrong combinations were most common before the steps "attach burette" (4 cases) and "fill empty beaker" (5 cases). Finally, in Level 3 (Figure 5.22), participants had an equal number of incorrect combinations (3 cases) before the steps "fill empty beaker" and "place beaker".

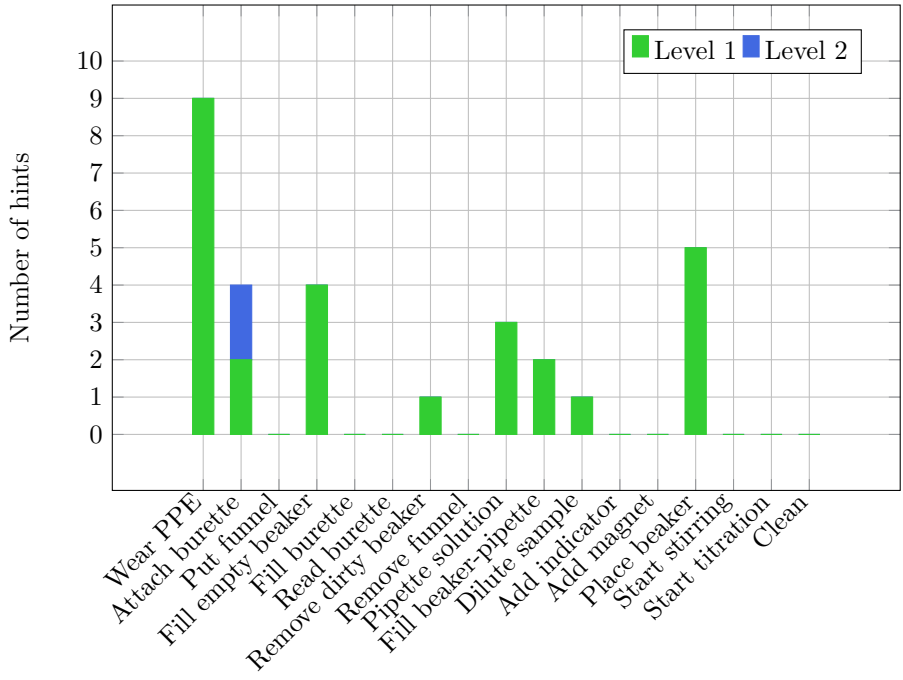


Figure 5.19: Number of hints per step given to Merck participants per level. A hint is triggered after three unsuccessful trials. The graph shows the steps in the sequential order observed in the titration process of the app.

Reposition Materials The process maps of Merck indicated that the *Reposition Materials* event was triggered twice at Level 1, after the incorrect combination of the "put funnel" step. Similarly, the reposition event was triggered once at Level 2, immediately following the "remove funnel" step. In addition, Level 3 initiated the repositioning event before the "wear PPE" step.

Summary

During the Merck evaluation, MAR Lab 2.0.1 was tested after minor modifications were made to the prototype. Seven participants interacted with all three levels of the application, which had been translated into German, and their data were analyzed.

Merck's participants reported different issues with the application, with the most frequently reported issues being materials disappearing, floating objects and unusual behaviour of the pipette and burette. Interestingly, these objects did

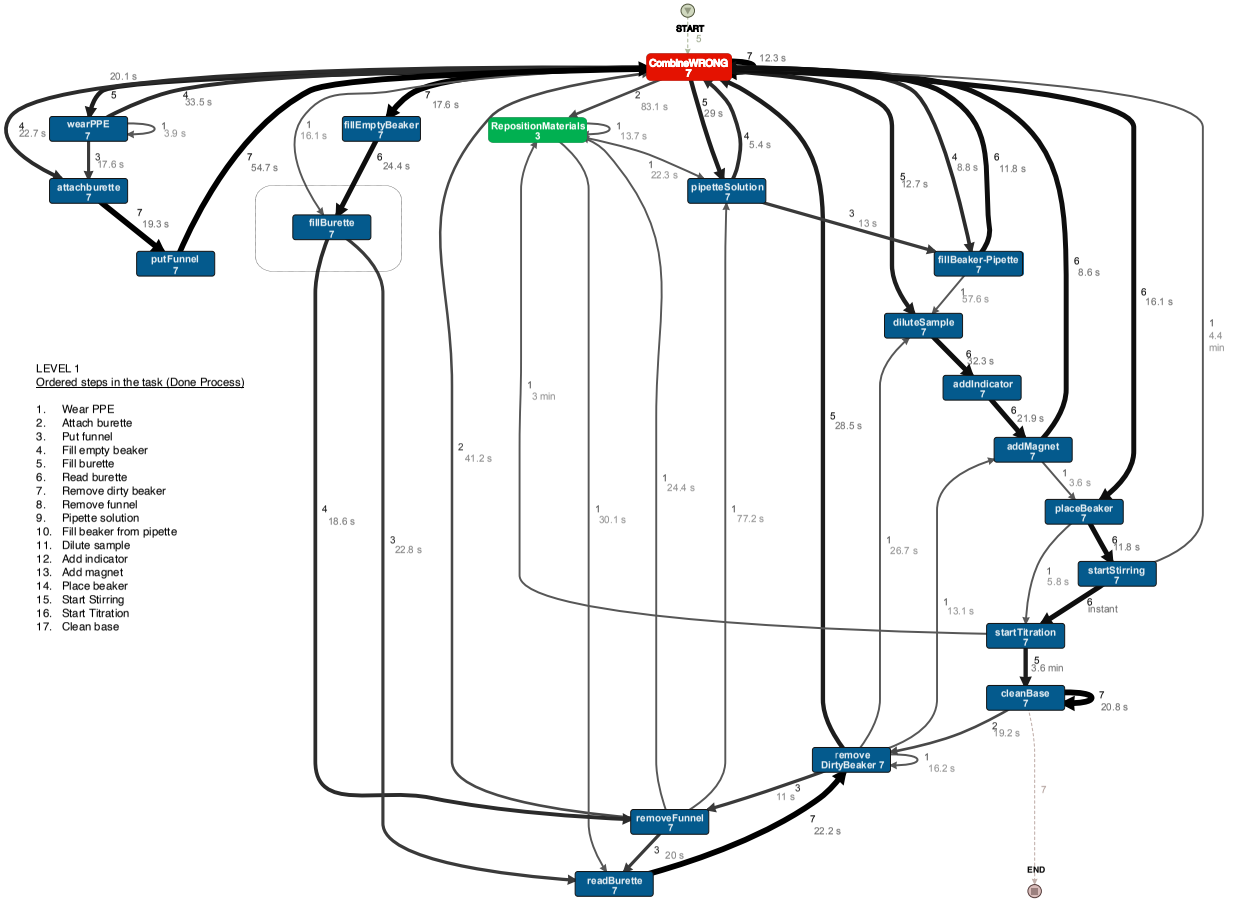


Figure 5.20: Merck process map of Level 1 showing (according to DISCO metrics) 100% activities and 35% of paths. The logged events include *Reposition Materials* (green box), *Combine Wrong* (red box) and *Done Process* (blue boxes). The box in the "fill burette" step highlights the divergent paths from the procedure. The numbers next to the arrows indicate the frequency and duration of events.

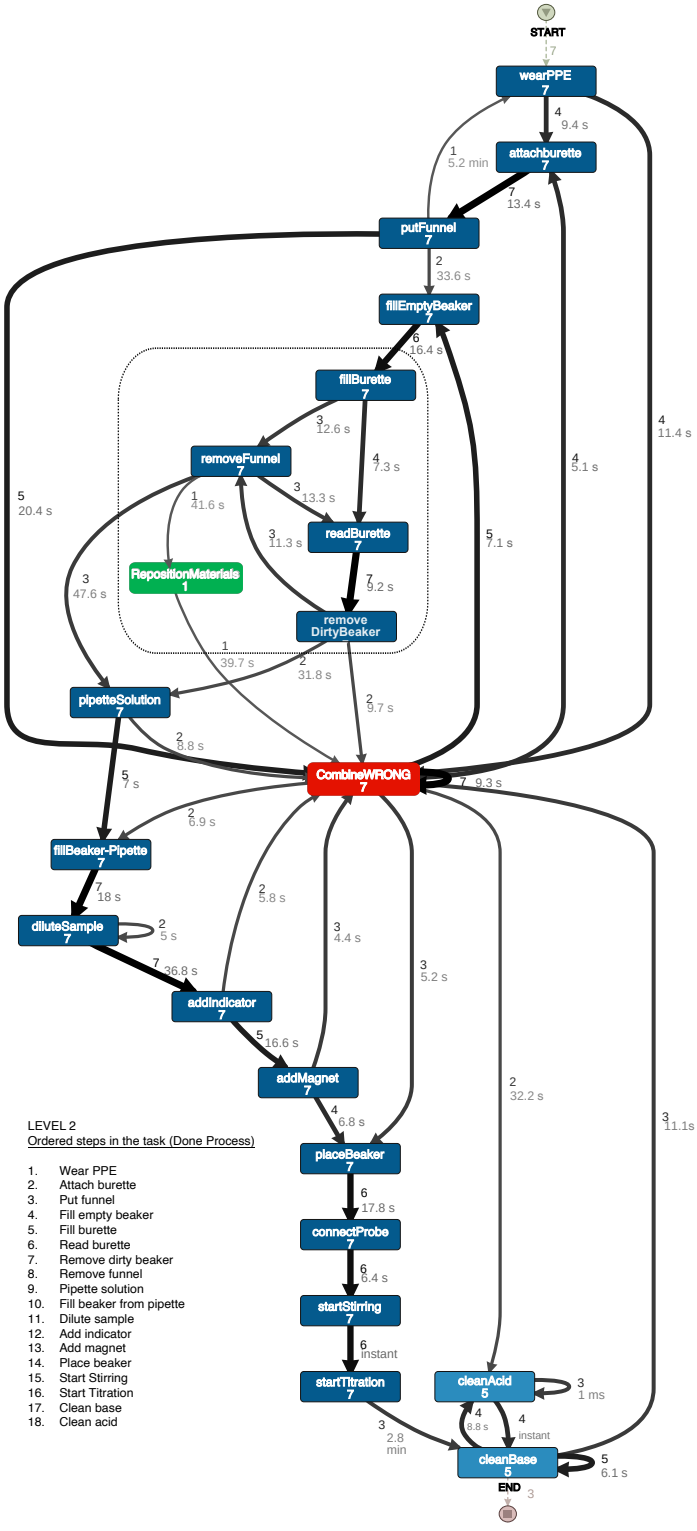


Figure 5.21: Merck process map of Level 2 showing (according to DISCO metrics) 100% activities and 35% of paths. The logged events include *Reposition Materials* (green box), *Combine Wrong* (red box) and *Done Process* (blue boxes). The box enclosing several events highlights the divergent paths from the procedure. The numbers next to the arrows indicate the frequency and duration of events.

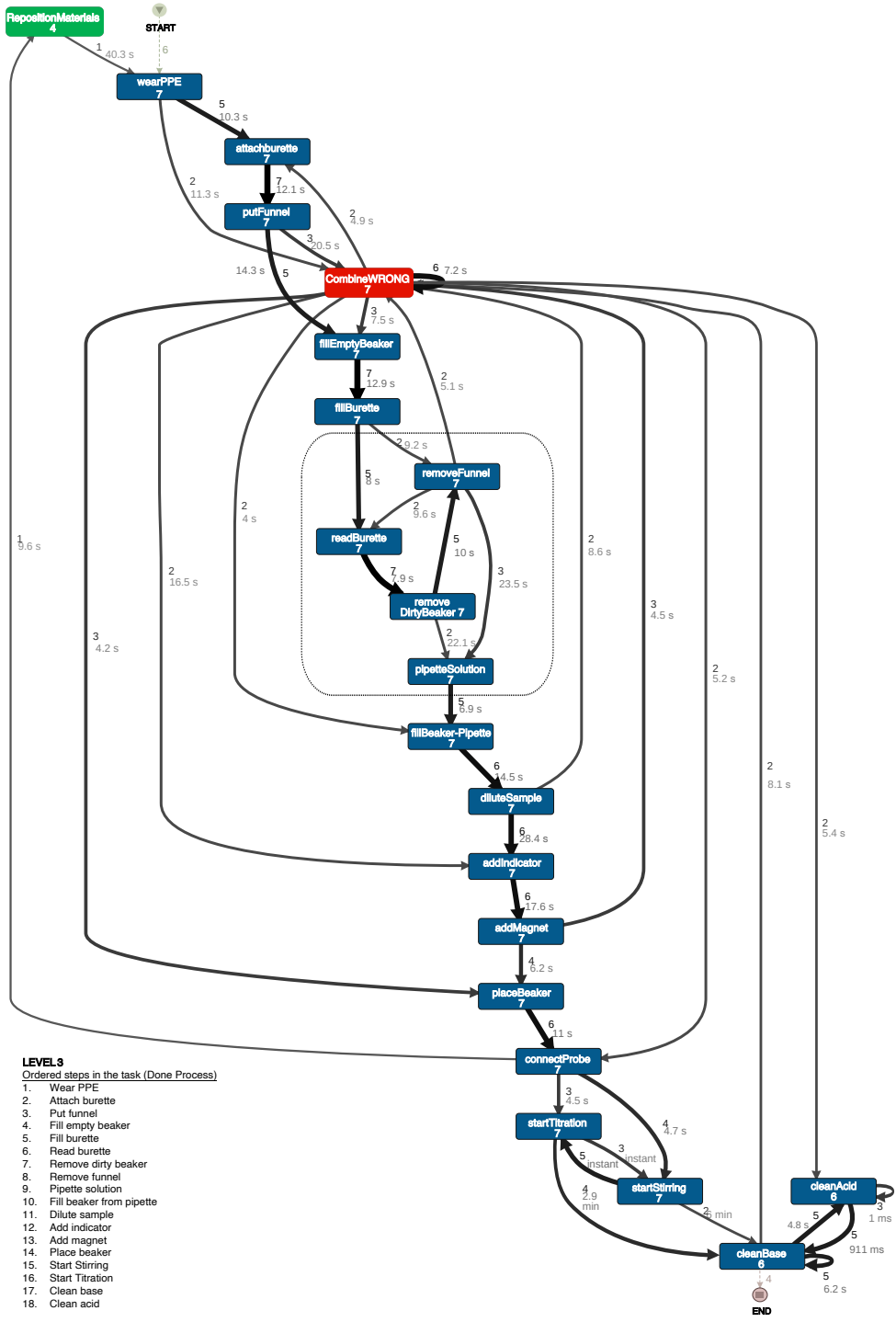


Figure 5.22: Merck process map of Level 3 showing (according to DISCO metrics) 100% activities and 35% of paths. The logged events include *Reposition Materials* (green box), *Combine Wrong* (red box) and *Done Process* (blue boxes). The box enclosing various events highlights the divergent paths from the procedure. The numbers next to the arrows indicate the frequency and duration of events.

not appear to have caused the most errors during the procedure. Participants praised the app for providing a realistic experience with hints, controls, and helpful features such as the sorting waste step, logbook, and the ability to move physically. The application was perceived as a useful tool for students to learn the general function of titration and the order of steps.

From the log data, it was clear that users had difficulty starting the first level. Participants received hints before the step "wear PPE" and had to use the reposition function to set up the environment. This observation is in line with the results from the ACTA test, where participants also struggled at the beginning of the level. Further, problematic steps experienced by Merck users were "fill empty beaker" and "place beaker". These steps were found in the process maps to trigger the event *Combine Wrong*. The issue with "fill empty beaker" was likely because users could only choose one beaker for this action (the interaction with the beakers was hardcoded). Finally, the step "place beaker" could have been difficult for users because it was unclear where it needed to be placed to complete the process. To address this issue, solutions could include enlarging the area where users must place the object or making it clearer.

The process maps demonstrated that Merck participants followed divergent paths after completing the "fill burette" step on all levels. This observation indicated that some participants, despite the instructions provided on each level, made their own decisions, likely due to their familiarity with the laboratory environment, enabling them to modify the steps according to their experience.

Merck participants reported being proficient in titration before testing MAR Lab, and their learning gains were not statistically significant, thus confirming their initial experience. In light of this, they provided suggestions on how to improve the experimental procedure. For instance, they suggested simplifying steps such as "diluting the sample" and modifying the indicator colour change to reflect better what happens in a real experiment. Furthermore, participants proposed additional instructions, such as properly closing a burette or pipetting a solution. They also suggested potential new experiments where they saw value in using a similar AR approach, such as organic synthesis, distillation, and filtration.

The results from the Merck evaluation suggest that while the MAR Lab application is useful in helping students understand the general function of titration and the order of steps, some areas could be improved. In particular, the app could be updated to make interacting with objects easier, fix problematic objects, and provide clearer instructions. Additionally, Merck participants suggested simplifying some steps, improving how the indicator changes colour and providing additional instructions on best practices in the laboratory.

5.6.3 KU Leuven Results

The MAR Lab app tested in the KU Leuven study (MAR Lab 2.0.2) underwent minor modifications from its previous version (MAR Lab 2.0.1). The colour-changing feature during titration was adjusted to make the transition smoother. Additionally, a Dutch translation was included, and a lab coat was added as a game object to the lab environment, making the safety step more realistic.

Participants

The third evaluation was done at KU Leuven in April 2022. The participants in this evaluation were recruited from the third-year bachelor program of Engineering Science at KU Leuven. The content of the application and the questionnaires were administered in Dutch. In total, 41 students were recruited via email by the lecturer of the analytical chemistry course (Analytische scheikunde B-KUL-H00Q9a).

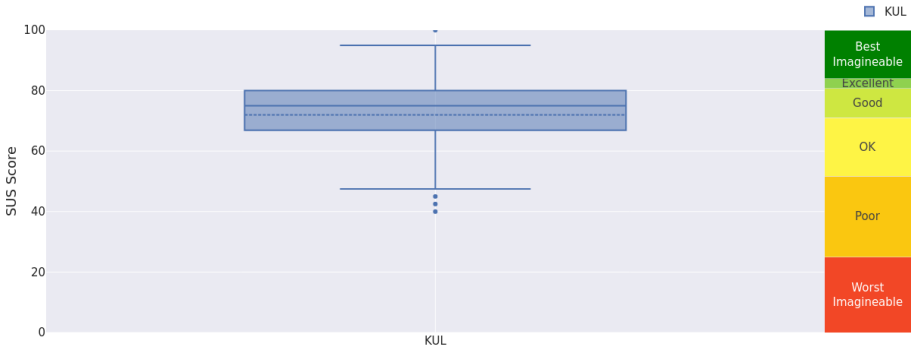
In this evaluation, data from 29 participants were examined. However, the data of 12 students were not included due to incomplete questionnaires or unavailable log data. The participants were 20-21 years old and consisted of 23 males and 6 females. At the time of the evaluation, this group had learned about titration experiment principles in their previous theoretical courses.

Observations during the study

The user test was organized in three different sessions with an average of 15 participants per timeslot. To ensure social distancing and comply with COVID regulations, sufficient space was allocated between each table provided to the participants, which included a white table, an ID number, an informed consent form, and blank papers. The participants mainly worked independently while standing, occasionally asking the researchers questions. Occasionally, participants were seated while responding to the questionnaires. The participants were instructed to download the application onto their phones. The beta version of the app was accessible through Google Play Store and TestFlight. In case a participant didn't have a phone or preferred not to use their personal device, the researcher offered a mobile phone with MAR Lab already installed. The available phones were Samsung Galaxy S10 and S20 or iPhone 8.

Prototype's Usability

The results of the SUS questionnaire indicated that the participants of KU Leuven reported an average score of 71.98, with a standard deviation of 15.21 - which is considered 'good' according to Bangor et al. [118]. This score indicates marginal usability, as evidenced by Figure 5.23. Compared to the industry benchmark defined by Lewis et al. [138], the prototype is classified as "above average".



Variable	SUS Score (mean)	SD	Adjective Scale	Acceptability Scale	Industry Benchmark
KU Leuven	71.98	15.21	Good	Marginal	Above Average

Figure 5.23: System Usability Scale results contextualized on the scale proposed by Bangor et al. [118]. The mean SUS score is 71.98 (SD=15.21). According to the scale, the prototype has "good" usability on the adjective scale. The figure was created using an open-source toolkit [119].

The results of the SUS score analysis (Figure 5.24) demonstrate that questions 4, 6, and 10 had the most impact on the overall score reported by KU Leuven students. The normalized values for these questions were 8.71, 7.93, and 8.1, respectively. These addressed the effectiveness and learnability of the system. The lowest contribution to the overall score came from questions 3 and 9, with a normalized value of 6.12. These were related to the ease of use and the user's confidence while using the system.

User Feedback

During the KU Leuven test, user feedback was collected through the open-ended questions of the post-questionnaire. Analysis of user experiences with MAR Lab through inductive thematic coding revealed positive traits and improvement areas.

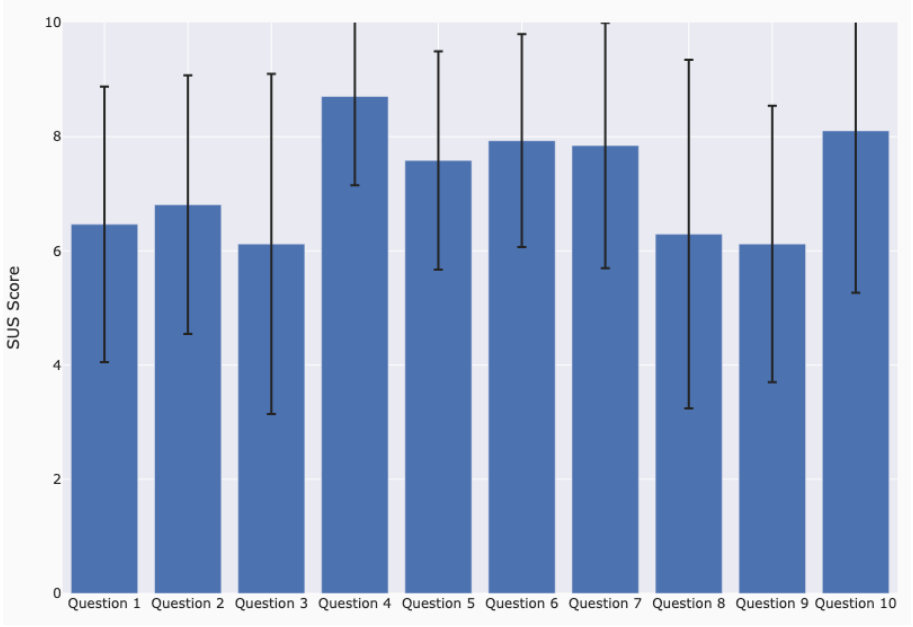


Figure 5.24: Normalized values ranging from 0 to 10 for each item of SUS and the general SUS questions for the KU Leuven test. The horizontal line indicates the standard deviation associated with each item. The figure was created using an open-source toolkit [119]. The SUS questionnaire administered in this evaluation can be found in Appendix C.2.3.

Problems Table 5.4 shows the reported problems from KU Leuven participants. The most common issue was materials disappearing, with 17 mentions. Related to this, several users reported "losing the pipette or burette," as one user expressed, "in the middle of the level, my pipette disappeared." Tracking, user interface, crash, and selection were all reported 5, 5, 4 and 4 times, respectively. Examples of the tracking issue included comments such as "recognized table was difficult" and "sometimes the set-up moved on its own and had to be repositioned". Problems with the user interface were related to the buttons, the logbook's scrolling function, and the progress bar, which showed incorrect progress. In this study, crashes and selection problems were identified by comments like "my smartphone was struggling. App quite heavy maybe?" and a selection issue such as "sometimes you have to go way too close before objects are selectable, causing your table to sprint". Other issues, such as translation and problems with specific materials, were mentioned 3 and 2 times, respectively. Users identified translation errors and problematic materials, such as the pipette and the stirrer.

Table 5.4: Problems reported by 29 KU Leuven participants

Codes	Number of mentions
Materials disappearing	17
Floating objects	5
Tracking	5
User interface	5
Crash	4
Selection	4
Translation	3
Problem with specific materials	2

Suggestions Participants provided suggestions for improvement, such as including the option to return to the main menu without losing progress and turning off the sound. Furthermore, they suggested developing a tablet version "to give more space to work with". One user suggested that a version without augmented reality would be beneficial, saying, "maybe without a camera, because that makes the app heavy and offers little added value. Just place the set-up in a fictitious lab, where you can also indicate everything."

Learning Experience Regarding the learning experience, participants of KU Leuven expressed that playing with the application was challenging at the beginning for example, one user expressed, "there is quite an entry curve. After that, it goes a lot smoother". Moreover, participants mentioned some unclear steps, such as sorting waste or removing the bottle cap.

General User Experience Users expressed both positive and negative sentiments regarding the MAR Lab app. On the positive side, users expressed their appreciation for the application, with one user stating, "it was very convenient to use". Users noted that the app could benefit students if some of its flaws were addressed, such as issues with tracking and inconsistency. For instance, one user commented, "if you have to unscrew a cap from a case, why not from the bin?" while another suggested that "if the stability can be improved so that the setup, once set up, is fixed and doesn't move or float away, then the app is very good."

AR Experience Finally, 4 participants reported using AR applications, including the Ikea app, Pokémon Go, and a demo of The Pocket. The remaining participants reported never having used an AR application.

Learning Test

The mean pretest score was 8.48($SD = 0.99$), and the mean posttest score was 8.97 ($SD = 0.94$), as depicted in Figure 5.25. A Wilcoxon Signed Rank Test at a significance level of $p < 0.05$ indicated that there was enough evidence to reject the null hypothesis ($W = 36$, $p = 0.026$, $r = -0.529$), suggesting an improvement in the posttest scores of the participants. Since the data did not follow a normal distribution (Appendix C.12), the non-parametric test was used. The effect size expressed in CLES revealed a 63.4% chance that the posttest evaluation would be higher than the pretest evaluation if a pair of scores were randomly selected (Table D.1).

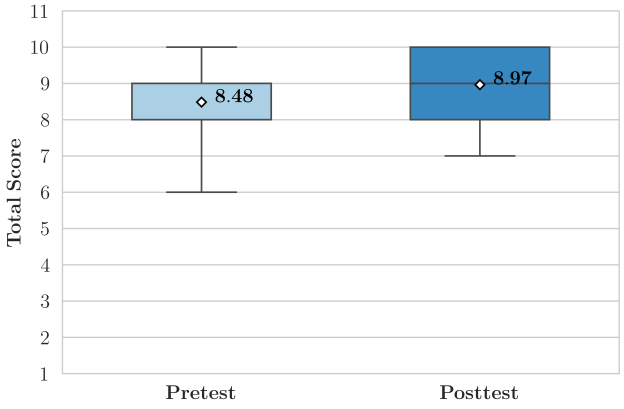


Figure 5.25: Boxplot showing the pretest and posttest scores of the KU Leuven group. The boxes represent the interquartile range (IQR), with the line inside each box depicting the median scores and the rhombus indicating the mean scores. The pretest and posttest scores had respective means of 8.48 ($SD = 0.99$) and 8.97 ($SD = 0.94$). The Wilcoxon Signed Rank Test showed significant differences between the two tests ($W = 36$, $p = 0.026$, $r = -0.529$).

Log Data and Process Maps

The log data and process maps for participants in the KU Leuven test were analyzed and treated per level. The custom event *Show Hint* was tracked to reveal which steps elicited the most hints from participants. The results, as shown in Figure 5.26, indicated that the "fill empty beaker", "remove dirty beaker", "pipette solution", and "clean" steps triggered the most hints across all levels. Notably, in Levels 2 and 3, only a few hints were given, as the hints only appeared after 5 or 7 unsuccessful trials in the app.

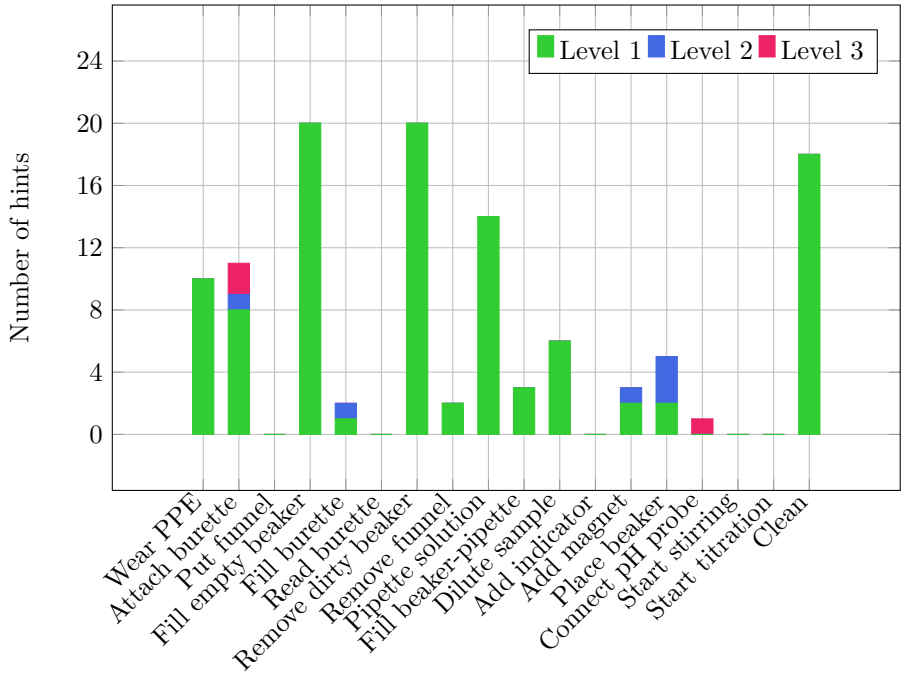


Figure 5.26: Number of hints per step given to KU Leuven students per level. A hint is triggered after three unsuccessful trials. The graph shows the steps in the sequential order observed in the titration process of the app.

Done Process Analysis of the custom event *Done Process* revealed that participants followed the procedure as instructed in Level 1 (Figure 5.27). However, for Levels 2 and 3 (Figures 5.28 and 5.29), the process maps showed that users varied in the order of tasks performed after the "remove dirty beaker" step in both levels.

Combine Wrong The analysis of the custom event *Combine Wrong* uncovered problematic actions experienced by users. Common steps were found to precede wrong combinations during the experiment across all levels. These steps included "attach burette", "fill empty beaker", "place beaker", "clean", and "fill beaker from pipette". Interestingly, the number of instances these events occurred decreased from Level 1 to 3. For example, the step "attach burette" had 21 cases in Level 1, which was reduced to 11 cases in Level 3. Similar reductions occurred with the other activities, as can be seen in the process maps of Figures 5.27, 5.28 and 5.29.

Reposition Materials The custom event *Reposition Materials* was found to be triggered in KU Leuven process maps at similar steps across all levels. Specifically, the event was registered before the step "wear PPE" and after "remove funnel", and in Level 2, it was also triggered after the step "put funnel". In Level 3, it was also triggered after the step "pipette solution".

Summary

MAR Lab 2.0.2 was tested by students from KU Leuven shortly after an evaluation conducted at Merck. Data from 29 students who interacted with the Dutch version of the application across three levels were analyzed. The study's results revealed a statistically significant increase in test scores from the pretest to the posttest.

Participants from KU Leuven reported a wide range of issues with the MAR Lab app. The most frequent issue was materials disappearing, followed by tracking, user interface, crash, and selection. Problems with specific materials were also reported (pipette and stirrer). Despite some participants noting errors with the Dutch translation, the language barrier did not significantly impede the system's learnability in this test. Participants expressed both positive and negative sentiments regarding the application, noting that it could be beneficial to students if some of its flaws were addressed.

The process maps and log data uncover problematic steps experienced by the KU Leuven students. As Figure 5.26 illustrates, the steps "fill empty beaker", "remove dirty beaker", and "clean" were the steps where the participants received the most hints. This corresponds to the problematic combinations triggered (*Combine Wrong*), suggesting that students faced difficulties with these steps, in addition to "attach burette". The issues may have been caused by system-related problems, unclear instructions or unfamiliar concepts. For example, attaching the burette at all levels could be difficult due to the small area they have to aim at, whereas, during the cleaning step, the student needs to identify whether the substance to be discarded is acidic or basic to discard it properly.

Another interesting observation came about from the analysis of the *Reposition Materials* custom event. Participants used this function at the start of the level or when putting or removing the funnel in the titration set-up. A possible explanation for this result could be that users lose track of the environment when they move the phone up to interact with the funnel. This action may cause the application difficulty locating the original tracking position. It is also important to consider that, unlike the other experiments, the participants' tracking problems during this evaluation could have been related to several different factors, such as the lighting conditions of the room, the limited time

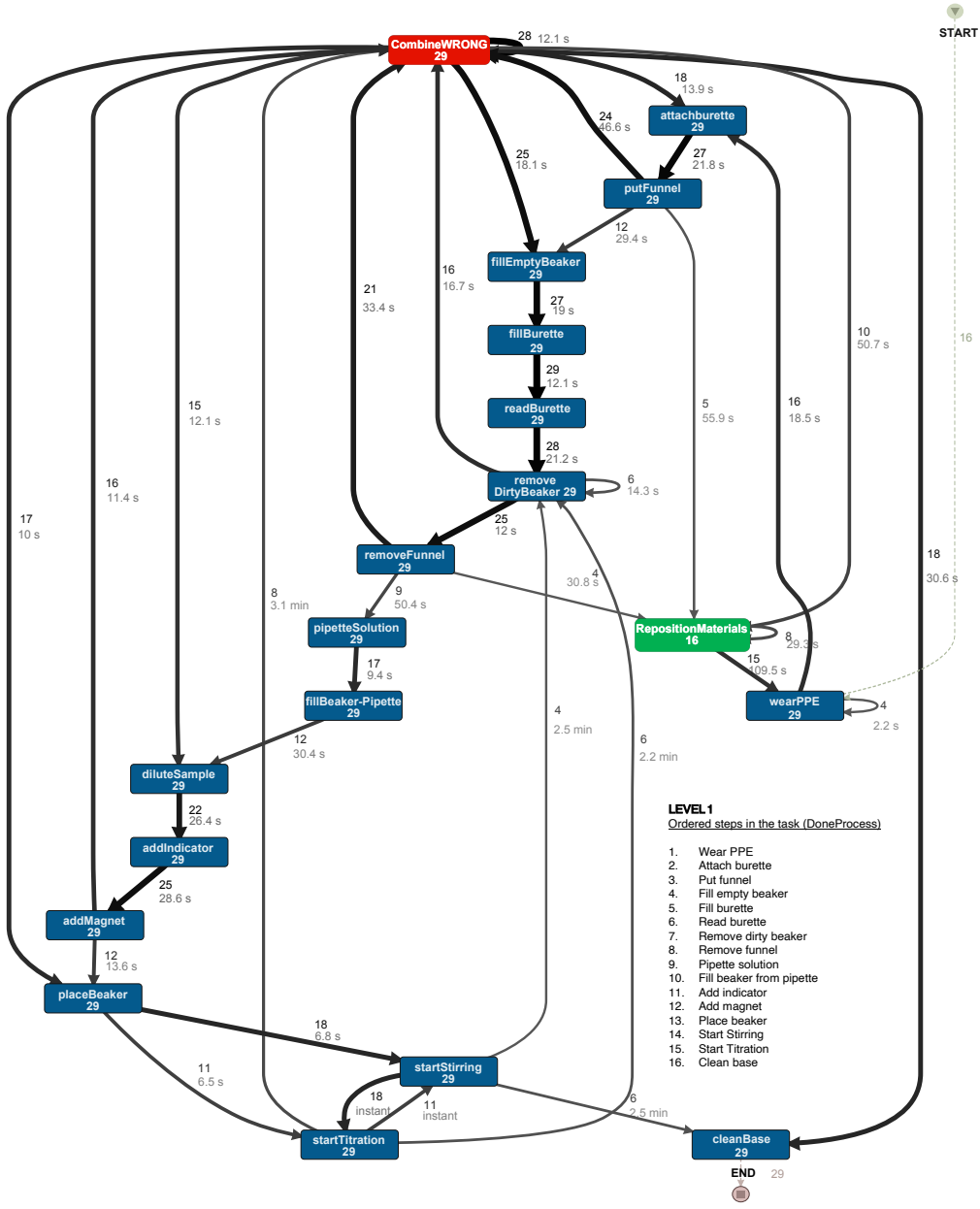


Figure 5.27: KU Leuven process map of Level 1 (as recorded by DISCO metrics) displays 100% of activities and 35% of paths. Each box features different events, with the arrows connecting them indicating the frequency and duration. The events are in distinct boxes- *Reposition Materials* (in green), *Combine Wrong* (in red) and *Done Process* (in blue).

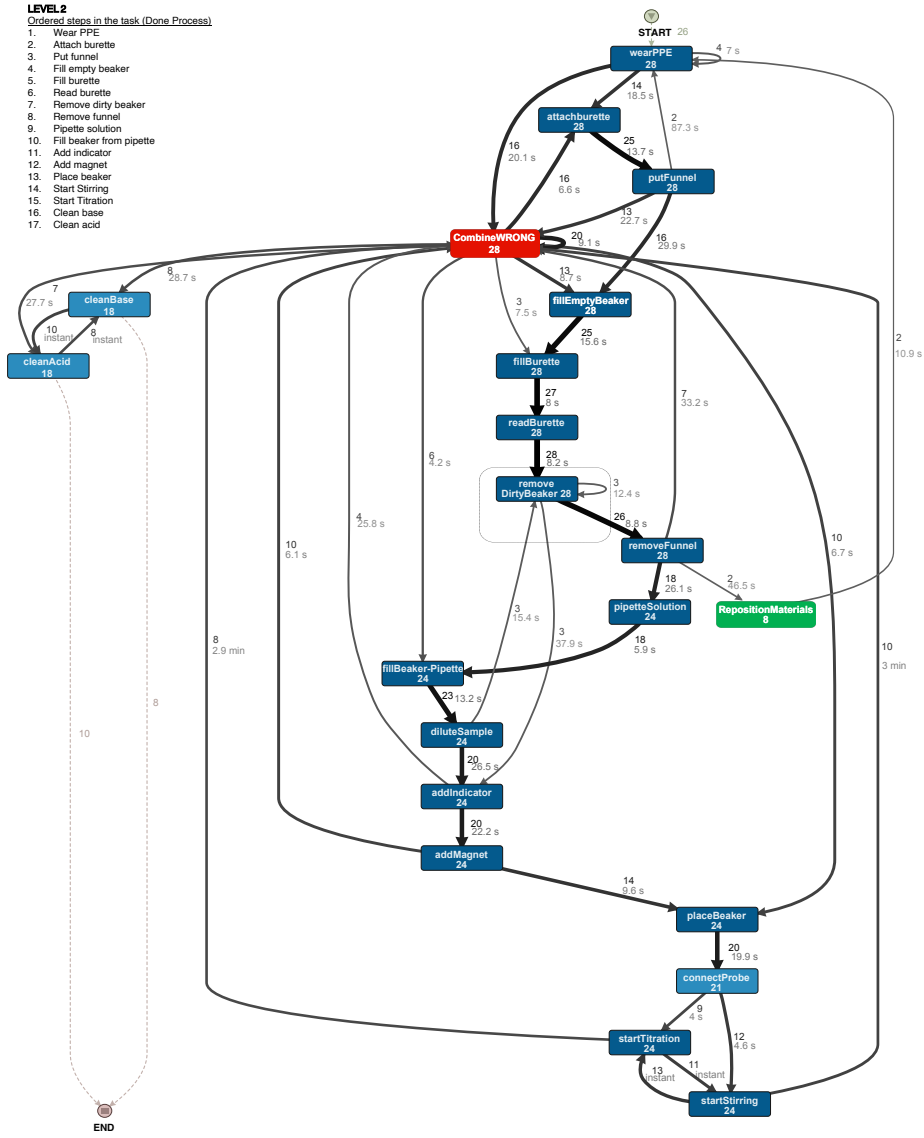


Figure 5.28: KU Leuven process map of Level 2 (as recorded by DISCO metrics) displays 100% of activities and 35% of paths. Each box features different events, with the arrows connecting them indicating the frequency and duration. The events are in distinct boxes- *Reposition Materials* (in green), *Combine Wrong* (in red) and *Done Process* (in blue). The box in the "remove dirty beaker" step highlights the divergent paths from the procedure.

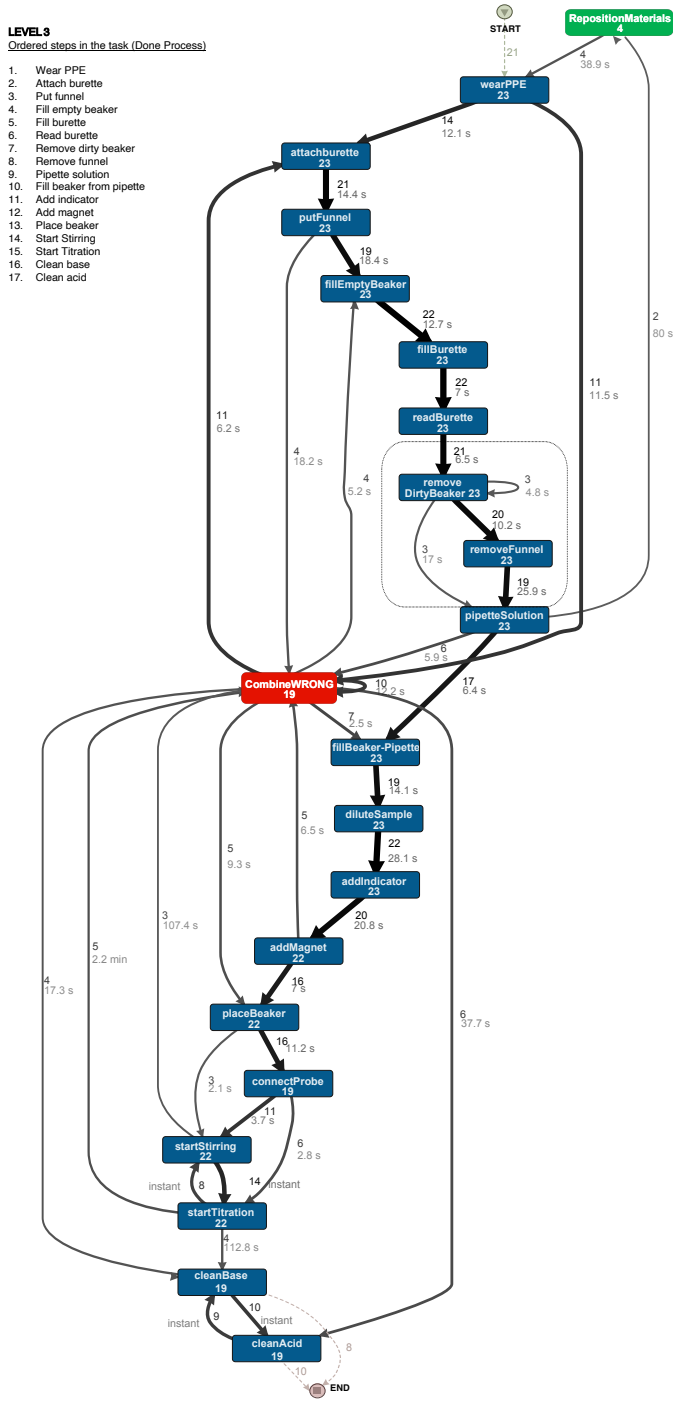


Figure 5.29: KU Leuven process map of Level 3 (as recorded by DISCO metrics) displays 100% of activities and 35% of paths. Each box features the logged events, with the arrows connecting them indicating the frequency and duration. The events are in distinct boxes- *Reposition Materials* (in green), *Combine Wrong* (in red) and *Done Process* (in blue). The box surrounding various steps emphasizes the divergent paths from the procedure.

constraint that may have caused them to play in hurry due to their upcoming classes, the capabilities of their phones or simply the unfamiliarity with AR.

5.7 Discussion

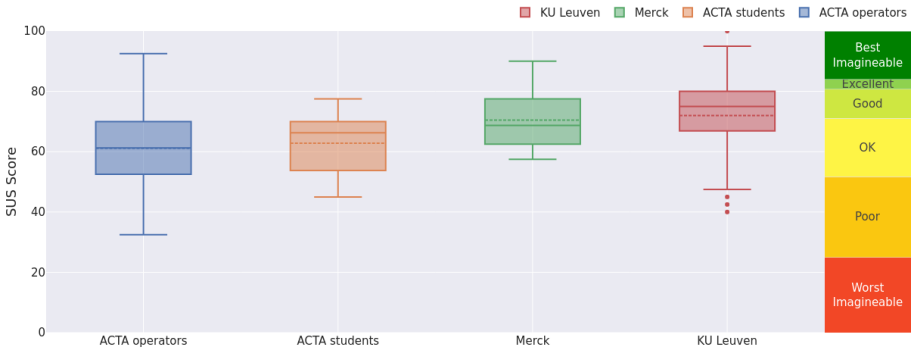
This chapter aimed to investigate how the prototype's usability changes after modifications, identify which steps in the procedure are problematic for the users in each group, and explore whether students could improve their learning on the titration experiment after the intervention. Furthermore, this chapter sought to understand the general perception of MAR Lab among users. Based on the DIF framework presented in Chapter 1, the formative evaluations described in this chapter would lead to an improved prototype in terms of usability, learning and user perception. Negative results could guide on features not to pursue in the next iteration.

The mixed-method approach for the three evaluations enabled the observations to reinforce each other's weaknesses. The quantitative analysis assessed the learning effect of the platform's use and measured the application's usability. Meanwhile, qualitative analysis was employed to obtain in-depth insights into the participants' experience with the application.

The overall SUS score improved on each iteration, culminating in 71.98 (SD=15.21) in the KU Leuven test (MAR Lab 2.0.2), which can be considered "good" according to the adjective scale [118]. As shown in Figure 5.30, these results show that the prototype improved its overall subjective usability between each iteration. However, future development should focus on improving ease of use and eliminating inconsistencies and bugs, which have been observed through user interviews and open-ended questions.

The ACTA, Merck, and KU Leuven test results indicate that the MAR Lab application has strengths and weaknesses. The ACTA test revealed issues with tracking, crashing, and selection, and users provided suggestions for improvement, such as translating the app language. Meanwhile, the Merck group reported the most frequent issues being materials disappearing and floating objects, yet praised the sorting waste step, logbook, safety procedure, and the function of "repositioning of the materials". Lastly, the KU Leuven test revealed that the most common issue was materials disappearing, followed by tracking and user interface, with users suggesting features to make progress and convenience more accessible.

Overall, users in all three test experience problems with stable tracking and handling the virtual objects (i.e., floating or disappearing). Details of the



Variable	App Version	SUS Score (mean)	Adjective Scale	Acceptability Scale	Industry Benchmark
Operators	2.0	61.11 ± 15.05	OK	Marginal	Below Average
Students	2.0	62.81 ± 10.19	OK	Marginal	Below Average
Merck	2.0.1	70.5 ± 9.6	OK	Marginal	Above Average
KU Leuven	2.0.2	71.98 ± 15.21	Good	Marginal	Above Average

Figure 5.30: System Usability Scale results of all tests. The mean SUS score of each experiment is shown in the table. According to the scale, the final iteration of the prototype (KU Leuven test) has "good" usability on the adjective scale. The figure was created using an open-source toolkit [119].

process maps showed that these problems are evident during specific parts of the procedure as shown by the custom event *Reposition Materials*. Observations of this event showed that users struggled during actions which required them to move the phone farther away from their point of view, such as when they needed to put the funnel into the burette. Additionally, unsuccessful tracking or errors were noted at the start of Level 1, which may have been due to user unfamiliarity with the AR environment, as reported by most users, or the language used in the onboarding may not have been clear. This difficulty was observed to improve in subsequent levels, likely as users learned how to complete the initial tracking steps.

On the other hand, difficulties with handling virtual objects can be explained by several factors. Firstly, users experienced issues with particular objects (e.g. pipette and burette), which were found to be problematic due to a programmed setting in the game engine (i.e., each item had multiple colliders, causing physical behaviour issues). Secondly, users had difficulties selecting and grabbing or combining objects with secondary objects (e.g. the step "place beaker"). This might be due to a lack of space in the virtual environment, resulting in unwanted collisions between virtual elements. This could be an important factor for future implementations of the TrainAR framework, wherein practitioners should consider more space to spread virtual elements or scale the objects. Lastly, another possible explanation, supported by prior research [140],

could be the difficulty of aiming a crosshair at a specific object, which requires understanding the spatial relationship between the two-dimensional projections and the real environment.

Another important finding is the indication of different paths followed by participants. For instance, the process maps showed that participants on the KU Leuven test differed in the procedure after the step "remove dirty beaker" in both levels. Interestingly, these paths are similar to those observed in other tests, with the only difference starting one step before or after. This observation could be attributed to the fact that users at these steps are not paying sufficient attention to the instructions and prioritize their actions differently, possibly depending on the "closest" object or their prior knowledge of the experiment. Another potential explanation is that an additional cognitive load of reading, remembering, or listening to instructions while performing the necessary actions may lead participants to "choose" an easier or more intuitive path.

The formative evaluations of the participants' learning outcomes using MAR Lab showed mixed results. The operators' group of the ACTA test demonstrated a statistically significant learning gain, whereas the students' group did not show a statistically significant improvement in learning. This discrepancy can be attributed to the group's different backgrounds; operators reported not having limited knowledge or experience in the laboratory, whereas the high school students reported having already studied the experiment. It is important to remember that this result should only be considered directive, as the test used in this evaluation was being piloted with this group.

Similarly, the results of the Merck test were not statistically significant; however, this group reported that even though they were aware of the experiment, remembering and recalling the procedure was beneficial and could help future students. On the other hand, the KU Leuven test revealed a statistically significant improvement in learning. While the limited sample size and potential ceiling effect of prior knowledge of the participants in titration might have influenced the results, the positive outcomes suggest that MAR Lab may be a helpful tool to improve knowledge acquisition in titration, albeit to a minimal degree. Nonetheless, it is essential to bear in mind that the external validity of these results is limited. Further research with a larger sample is required to confirm the efficacy of the application in different contexts and populations.

The results of the evaluations have prompted several modifications for the next iteration. These changes aim to address some of the main issues experienced by users, such as incorrect translations, lack of realism when selecting beakers, and lack of consistency in certain actions (e.g. opening and closing bottles). Additionally, new features will be implemented to enhance the user experience, such as the ability to opt for hearing the agent's voice or muting it, a review

menu to review results after completing a level, and a new level to allow for less-restricted experimentation. Further, the procedure can be simplified by removing unnecessary steps, such as diluting the sample and improved by adding more signalization of important steps, such as "closing the burette". It is expected that these modifications will help to improve the overall usability of MAR Lab and, ultimately, provide students with a better user learning experience.

From a general perspective, the findings of this chapter suggested that the user experience of AR might vary among different learners due to a range of factors, such as pre-existing knowledge, technology familiarity, motivation etc. Chapter 3 showed that while users generally found educational AR applications to be novel and useful, they also experienced technical difficulties and a lack of interactivity. Therefore, taken together, the technical issues but also that individual needs and preferences must be taken into account for the creation of effective AR applications and an improved overall user experience.

5.8 Limitations

The research presented in this chapter is subject to some important limitations. First, the small sample sizes on each evaluation make the findings of this research limited. Though a smaller sample size is common in usability studies [141] and may be helpful to provide an initial insight into how the technology is used, caution must be taken when interpreting the results from the learning tests.

Second, the diversity of the participants' backgrounds could be a source of bias, even after steps were taken to minimize the differences during recruitment by opening the invitation voluntarily and sharing the information through different channels. This must be considered when interpreting the results, as the groups were formed by the majority of males or people that were inherently interested in the project.

Third, the MAR Lab app was specifically tailored for titration and chemical engineering students. As such, any conclusions drawn about augmented reality, in general, may be limited to this specific topic. However, since the application was tested with different groups, this somewhat overcomes the bias.

Finally, the MAR Lab app is a prototype and as such several technical problems and novelty for users may have been present. Therefore, more research that supports similar implementations is needed to make generalisable results about the technology. Nevertheless, this research already adds to other efforts, such

as the HebAR app [31], which uses the TrainAR framework and was created in the context of midwifery training.

5.9 Conclusions

In conclusion, the results from the SUS questionnaire showed a positive overall improvement in the perceived usability of the application. Despite this, there were issues with stable tracking and handling virtual objects, and both were evident during specific parts of the procedure, such as when needing to move the phone farther away from the point of view. These issues, combined with participants taking multiple paths through the application, emphasize the need for participants to be given clear objectives and guidelines for interacting in the AR environment.

Evaluations of the learning outcomes revealed mixed results, with some participants showing a statistically significant increase in learning and others not. However, further research is necessary to provide a general conclusion due to the study's limitations. This chapter has presented research that can be used to inform future iterations of the MAR Lab. With modifications to address existing issues and the implementation of new features, it is anticipated that the MAR Lab application can be improved to provide a more effective learning experience. Additionally, it was noted that individual needs and preferences should be taken into consideration when designing effective AR applications in order to yield an enhanced user experience.

5.10 Acknowledgements

Wout De Schepper is kindly acknowledged for his contribution with regard to the development of the process mining methodology to analyze MAR Lab interactions and perform user studies as part of his master thesis [142]. Lucero Elizalde is also acknowledged for her contribution to designing the cartoon character of Marie Curie used in MAR Lab.

Chapter 6

Final Prototype Evaluation

This chapter is adapted from the manuscript submitted: J. L. Domínguez Alfaro, M. Arztmann, J. Jeuring and P. Van Puyvelde, "AR for Science Education: Students' Behaviour Patterns and the Relationship between Cognitive Load, Knowledge Acquisition and Performance," British Journal of Educational Technology.

The design and implementation of multimedia learning environments with augmented reality offer different opportunities for users to learn through its capabilities. However, the implementation and adoption of this technology are still relatively novel for practitioners and users.

One of the potential challenges that augmented reality presents to users is the additional cognitive load stemming from unfamiliar interactions. This could be particularly demanding for science education learners, as the material's complexity could also be overwhelming. To better understand the issues associated with incorporating AR into multimedia learning environments, it is essential to explore the interactions users have with AR systems, which could provide valuable insights into creating better user experiences and ultimately help learners to better engage with the learning material.

This chapter describes the final evaluation of the digital prototype of MAR Lab 2.1 with novice chemical engineering students who do not have access to a chemical laboratory in their course. The evaluation focused on the usability and participants' feedback. Additionally, the study explored how students interact with MAR Lab and assessed the cognitive load and in-app performance. Overall, the results of this user test aim to answer the following research questions:

- **RQ1**: How does the usability of the prototype change after the last modifications?
- **RQ2**: What is the general perception of MAR Lab among users?
- **RQ3**: Which steps in the procedure are problematic for the users?
- **RQ4**: How does student performance change as they progress from one level to the next?
- **RQ5**: How do cognitive load, performance and knowledge acquisition interact and influence each other while using the MAR Lab?

It is important to acknowledge that **RQ5** was formulated a posteriori, meaning that exploring the interaction between cognitive load, performance and knowledge acquisition was added as part of the prototype evaluation later in the study. The purpose of this question was to investigate how these various factors are interconnected, which is not a common approach in this field. Still, we rationalized that it was valuable exploring this interaction given the emergent and complex process that arises when using an AR application for learning and the contradictory findings reported in the literature regarding the cognitive load generated by AR systems.

6.1 AR & Cognitive Load

Despite the hype surrounding AR in education, there is still much to be understood regarding how students interact with AR applications and the cognitive load imposed upon learners. Studies on the cognitive load of AR systems have shown contradictory findings [143]. While some authors have suggested that AR systems can reduce or maintain low cognitive load [144, 145, 146, 24], others have found that AR can lead to high cognitive load, making it a potentially distracting experience for learners [21, 22, 147, 49]. As Buchner et al. [143] argued, the impact of AR on cognitive load remains inconsistent, as much of the research has focused on comparing media rather than investigating how and when AR impacts learning.

Aligned with Cognitive Load Theory, some mechanisms of how AR influences learning have been described in terms of their impact on cognitive load (CL). For instance, the temporal and spatial split attention principle [143, 66] suggests merging digital and physical objects within the same environment can positively influence CL and performance [147]. Additionally, AR environments often present learners with verbal and pictorial information concurrently, requiring

simultaneous processing of both forms of information. However, this processing is limited by the capacity of the working memory [65]. Therefore, the design of the AR environment could have either a positive or negative impact on CL [148]. Moreover, embodied learning can reduce CL and facilitate meaningful processing by transferring information processing to other modalities such as physical embodiment or "enactment" [143, 74, 76]. Ultimately, the affordances of AR and its implementation within a given scenario may have a different impact on both learning and CL.

6.2 Learning & Performance

Previous research has shown that AR increases the overall learning outcomes of students [23, 27, 149]. Moreover, one of the advantages of using technology in education is the availability of log data, which provides an objective performance measure [150, 151].

Utilizing log data to gain insight into students' learning, such as evaluating certain behaviour patterns [152], allows for implementing performance measures, like time-on-task, for assessing learning. Previous studies have demonstrated that time-on-task in traditional learning environments influences learning outcomes [153]. Consequently, a similar result is expected in an AR environment. However, since implementations of AR often result in a rather complex learning environment, using AR can lead to a high cognitive load, which can result in students spending an extended amount of time on the task due to difficulties understanding the environment and thus potentially leading to lower learning outcomes.

Time-on-task has also been used as an objective proxy to measure cognitive load. Research has shown that the difference in learners' time on task can indicate different levels of cognitive load multimedia impose [154]. In an AR environment, due to the potential unfamiliar interactions for students, this variable is expected to increase with cognitive load. Nevertheless, the increase in time-on-task may not always be a negative outcome. If systems enable learners to pause and reorient themselves by reducing the cognitive load, extended time spent on activities can be beneficial to learning and performance [155]. It is important to note that since time-on-task is an indirect measure of cognitive load, other factors like attention, motivation, emotion, or stress may also play a role in increasing time-on-task [154, 156].

6.3 The Proposed Model

This research explores the interaction between cognitive load, performance, and knowledge acquisition to understand how an application that follows instructional design principles impacts these factors.

MAR Lab is hypothesized to enhance learning, and the students' perceived cognitive load is expected to influence their knowledge acquisition. It is further anticipated that the more time spent on the application, the higher the students' learning outcomes; however, the increased cognitive load could also lead to more time spent as difficulties with the environment become apparent. This aligns with the Cognitive Affective Model of Immersive Learning (CAMIL) proposed by Makranksky and Petersen [157], who, similarly to others [126], suggest focusing on the instructional method rather than the technology. The expected relationships are visualized in Figure 6.1.

In this chapter, time-on-task is operationalised as the amount of time each student spends engaging with a single level of AR environment. Cognitive load is assessed using the NASA-TLX instrument, while knowledge acquisition is established through a custom-learning test.

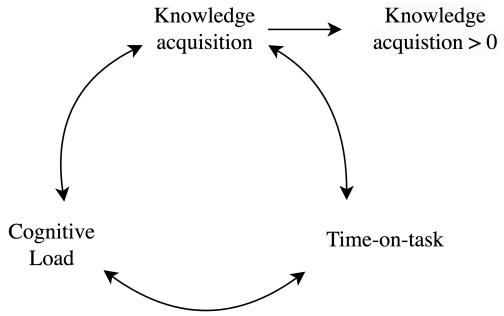


Figure 6.1: The proposed model for investigating the relationship between students' in-app performance (measured as time-on-task), knowledge acquisition (measured by the difference in scores of pre and posttest), and perceived cognitive load (measured using the NASA R-TLX).

In the literature, several reported methods and instruments are employed to measure cognitive load, including dual-task methodology and physiological and subjective measures. Of these, subjective measures are the most common and rely on the self-reports of the user for measurement [158]. The NASA Task Load Index (TLX) is a multidimensional instrument that is widely used in cognitive load studies, with the adapted version of the TLX (raw TLX or R-

TLX) without the weighting step proving to be equally effective and practical to use [159, 160, 161]. For instance, Lampen et al., [162] used R-TLX to compare cognitive load during manual assembly tasks and found that AR-supported instruction methods mitigate cognitive load compared to other instruction methods. Similarly, Alves et al., [163] used the instrument to compare the cognitive load during assembly processes, comparing different types of AR (spatial and mobile AR). The results showed a lower mental workload regarding the spatial AR type. This chapter uses R-TLX to measure students' cognitive load using MAR Lab.

6.4 Methodology

A mixed-method approach was taken for this evaluation. First, quantitative analysis was used to determine the learning effect of the platform use, users' cognitive load, system usability, and users' performance. On the other hand, the research uses qualitative analysis to gain detailed insights into the participants' experience with the application. The final user evaluation was conducted in November 2022. Ethical approval was obtained from the ethical committee at KU Leuven G-2020-2532(AMD).

Participants

All participants were students of the General and Technical Chemistry (Algemene en Technische Scheikunde B-KUL-H01A8A) course at KU Leuven. The participants were recruited via email invitation on a voluntary basis. The participants were between 17 and 25 years of age and consisted of 19 males and three females. This study initially involved 22 students, but only the data of 15 male participants are being reported. This is because some of the data, either the post-questionnaire or log data, was incomplete or missing and was therefore excluded from the analysis. The individuals in this group only had a basic knowledge of titration and the chemical concepts involved, which they had learned in high school.

6.4.1 Procedure

The study follows a pretest/posttest design, which can be found in Figure 6.2. Four sessions were organized in the first weeks of November 2022 to ensure that students have not studied the topic of acid-base and titration in their theoretical course. Before data collection, the participants received an explanation of the

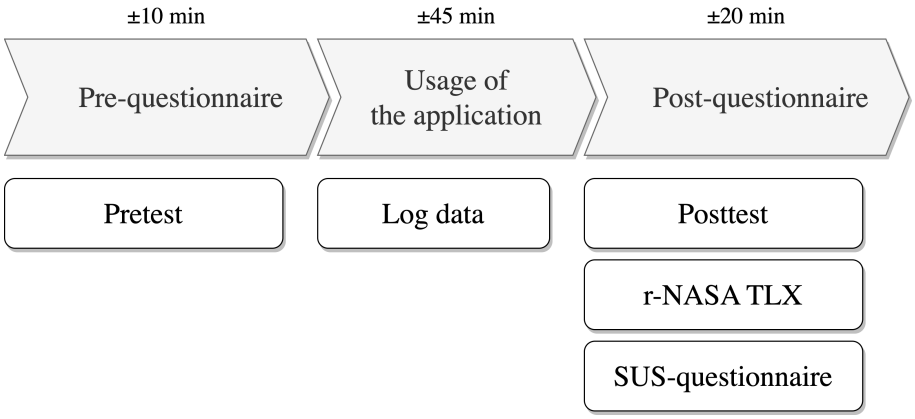


Figure 6.2: Overview of the study design and the time spent on each phase with each participant. The instruments appear in a rounded box below the phases.

project and informed consent. On obtaining written informed consent, a pretest was administered through a web-based survey tool (Qualtrics). Subsequently, participants were asked to download the application on their phones. MAR Lab 2.1 was made available as a beta version in the Google Play Store and TestFlight. If the participant did not have a phone or prefer not to use their personal device, the researcher provided a mobile phone with MAR Lab installed (Samsung Galaxy S10 and S20 or iPhone 8). After the installation, participants were asked to use the application for approximately 45 min. After this, a post-questionnaire was administered, including a posttest, the R-TLX, the SUS questionnaire, and ad-hoc feedback questions. At the end of the evaluation, participants received lunch as a thank-you for their participation.

6.4.2 Instruments

NASA TLX The NASA Task Load Index (TLX) is an instrument that has been used in several AR research studies to measure and capture the multiple dimensions of cognitive load [164, 143]. The TLX assesses subjective dimensions of mental workload, including mental demand, physical demand, temporal demand, effort, frustration, and performance [161]. The scale of each item ranges from 0 to 100, with 0 representing very low and 100 representing very high. To reduce variability among participants, the TLX assigns scores and weights the relevance of each dimension to the context. However, it has been shown that the adapted version of the TLX (raw TLX or R-TLX) without the weighting step remains effective and is more practical to use [160, 161]. Previous

research has indicated that the R-TLX items effectively align with cognitive load theory [165, 166]. Therefore, the modified version of the NASA-TLX is employed in this study to measure the cognitive load experienced by users.

Learning Test In this evaluation, the validated Dutch version of the chemistry test described in Section 5.5.1 was used to assess knowledge acquisition after using MAR Lab. The instrument can be found in Appendix D.1.

SUS questionnaire Similarly to the formative evaluations in previous chapters, the prototype's usability was evaluated using the Dutch version of the SUS 10-item standardized questionnaire. This evaluation measured the SUS score to compare how the subjective usability improved after the previous iterations.

Log data This evaluation recorded user actions similarly to Chapter 5.5.2. MAR Lab 2.1 was configured to collect data using Unity Analytics, and three additional events were added to this evaluation. These events include *ShowHintAfterAsking*, triggered when the user requests a hint; *OpenTheory*, triggered when the user clicks to check theoretical knowledge embedded in the application; and *MuteMarie*, registered when the user mutes the audio from the cartoon character. In addition, the time spent per level and the number of errors during the assembly of the titration set-up were measured to evaluate in-app performance.

6.4.3 Data Analysis

Figure 6.3 summarizes the different data sources and tools used for the analysis. To assess the prototype's usability **(RQ1)**, the open-source SUS Analysis toolkit [119] was employed. Thematic analysis was performed using the software NVivo 1.7.1 **(RQ2)**. The sequences of events generated by user actions per level were analyzed using DISCO (Fluxicon) version 2.14.0 as the process mining tool **(RQ3)**.

To investigate whether the participants improved their performance as they progressed through the levels **(RQ4)**, two metrics were examined: the total time spent on each level and the number of errors made during the assembly of the titration equipment. The analyzed data contained only those participants that completed the three levels ($N = 13$) using the Null Hypothesis Significance Testing (NHST) approach. The Friedman Test was used to determine whether there was a significant difference in the mean of total time and errors across the levels played by the same participants (i.e., paired sample). This test was

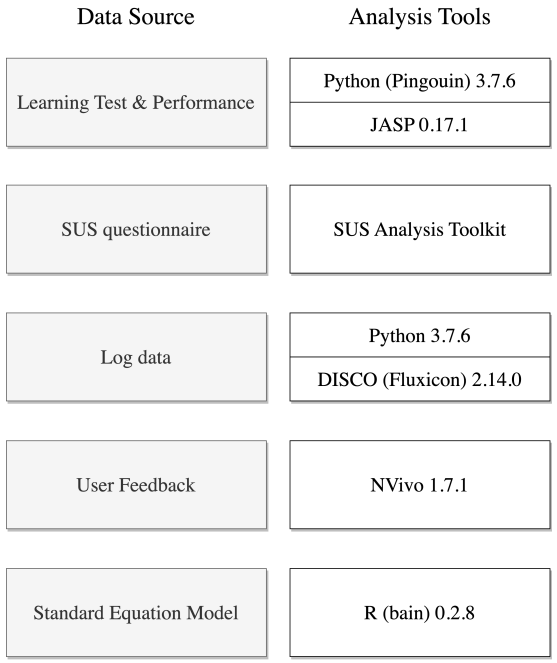


Figure 6.3: Five distinct data sources were gathered and analyzed using various software tools. The grey squares in the illustration depict the data sources, while the white squares denote the software and the respective version employed.

chosen as it allows for comparing several related samples without making any assumptions about the underlying distribution of the data [167]. Conover's Test was applied as a pairwise posthoc test to investigate further the differences between the individual levels [168, 169]. Moreover, the Bonferroni correction was utilized to reduce the chance of making a type I error when multiple comparisons were performed. Data analysis and management were conducted using Python library version 3.7.6 for descriptive statistics, the package Pingouin [136], and the software JASP 0.17.1 [137].

To answer **RQ5**, the Bayesian informative hypothesis testing approach [170] was adopted using a standard equation model (SEM) to analyze the data, with two competing hypotheses formulated based on expected relations between variables [171]. These hypotheses were formulated in terms of equality (=) and inequality (<,>) [172], and the analysis was conducted in R with the bain [173] and lavaan [174] packages version 0.2.8. Compared to classical NHST, the advantage of this approach is the ability to compare multiple hypotheses. Moreover, the support for the hypotheses of interest can be continually updated with more

data collected [170].

The hypotheses for evaluation were formulated as follows:

H_1: knowledge acquisition $\sim\sim$ cognitive load < 0 & cognitive load $\sim\sim$ time-on-task > 0 & knowledge acquisition $\sim\sim$ time-on-task > 0 & knowledge acquisition$\sim 1 > 0$

Hypothesis H_1 states that there is a negative correlation between cognitive load and knowledge acquisition and a positive correlation between knowledge acquisition and time-on-task. As cognitive load increases, more time-on-task is required to acquire knowledge, and knowledge acquisition increases as time-on-task increases.

H_2: knowledge acquisition $\sim\sim$ cognitive load < 0 & cognitive load $\sim\sim$ time-on-task $= 0$ & knowledge acquisition $\sim\sim$ time-on-task $= 0$ & knowledge acquisition$\sim 1 = 0$

The alternative hypothesis (H_2) states that cognitive load negatively correlates with knowledge acquisition, and neither cognitive load nor knowledge acquisition are correlated to time-on-task. This means that a decrease in cognitive load does not affect knowledge acquisition or time-on-task, and there is no knowledge acquisition.

In addition to the Bayesian approach, the thesis includes the application of the NHST approach for assessing knowledge acquisition for the sake of completeness. This method is employed similarly to other chapters.

6.5 Results

Observations during the study

The user test was organised in four separate sessions, with participants being assigned a unique ID number and provided with pens and blank sheets of paper. The tables were sufficiently separated to allow the students to walk around and were labelled with markers to help with tracking (Figure 6.4). At the beginning of each session, the participants were taught about the application's functionalities offered in the menu, such as a mute option and a reposition materials option. During the test, the participants primarily worked independently, occasionally asking questions to the researcher. Some participants used the paper to write their observations, and some students worked with headphones.

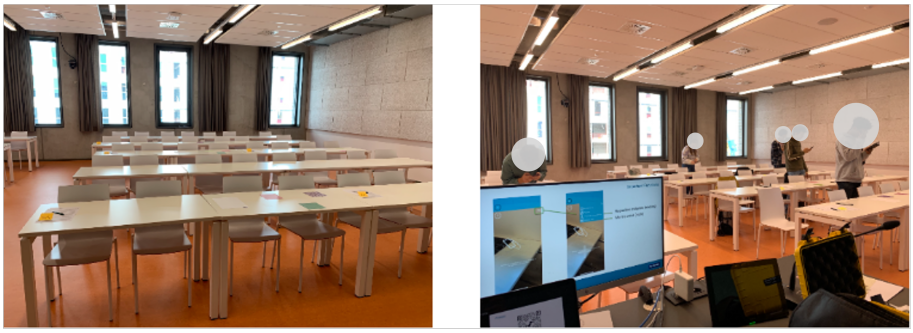
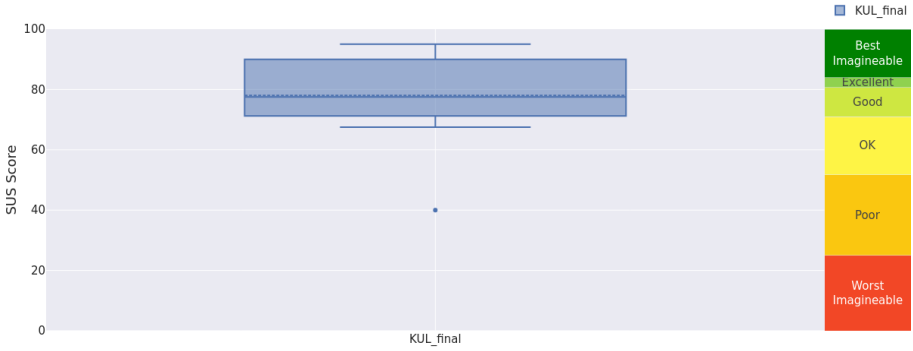


Figure 6.4: Example of multiple participants using MAR Lab during the evaluation simultaneously

Prototype's Usability

The results of the SUS questionnaire indicate an average score of 78, with a standard deviation of 13.42, which is classified as "good" according to Bangor et al. [118]. As depicted in Figure 6.5, a solid line illustrates the median value, and a dotted line represents the mean value. This score is indicative of acceptable usability.



Variable	SUS Score (mean)	SD	Adjective Scale	Acceptability Scale	Industry Benchmark
Merck	78	13.42	Good	Acceptable	Above Average

Figure 6.5: System Usability Scale results contextualized on the scale proposed by Bangor et al. [118]. The mean SUS score is 78 (SD=13.42). According to the scale, the prototype has "Good" usability on the adjective scale. The figure was created using an open-source toolkit [119].

The results of the SUS score analysis in Figure 6.6 revealed that questions 4, 5, 7 and 10 had the greatest influence on the overall score, with normalized values

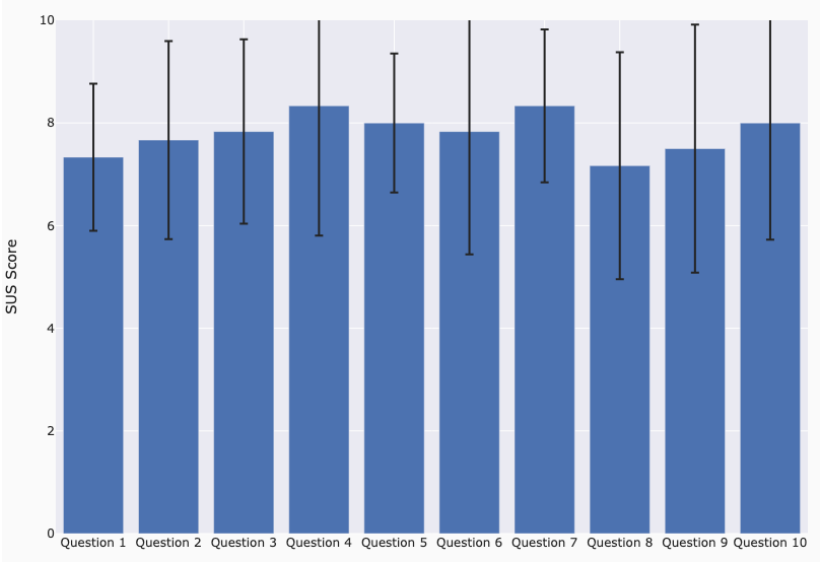


Figure 6.6: Normalized values ranging from 0 to 10 for each item of SUS and the general SUS questions for the final KU Leuven test. The horizontal line indicates the standard deviation associated with each item. The figure was created using an open-source toolkit [119]. The SUS questionnaire administered in this evaluation can be found in Appendix C.2.3.

of 8.33, 8.0, 8.33 and 8.0, respectively. These questions focused on the efficiency, integration and learnability of the system. Questions 1 and 8 had the lowest contributions to the overall score, at 7.33 and 7.17. These questions relate to user satisfaction and ease of use of the system, indicating that these aspects should be given special attention to optimizing the overall SUS score.

User Feedback

During the test, user feedback was collected through the open-ended questions of the post-questionnaire (see Appendix D.1). Analysis of user experiences with MAR Lab through inductive thematic coding revealed positive traits and areas that could be improved. The codes derived in the previous chapter were also utilized in this evaluation (see Appendix C.1), as the methodology and open questions were the same. The findings are summarized below.

Problems The thematic analysis revealed that tracking was the most commonly reported problem with the application. Users mentioned this issue six times, noting that the camera often loses its reference points or recalibrates when

they need to get close to an object. The second most common issue reported by users was translation mistakes, with a total of three mentions. Despite identifying this problem, they did not specify the type of translation errors they encountered. The least reported problems were materials disappearing (e.g. burette), selection of the stirrer, sudden crash, and floating objects.

Suggestions Participants of this study provided several suggestions to improve the user experience of MAR Lab. For example, users proposed adding features to make progress and convenience more accessible, such as the ability to skip steps if stuck and a button to acquire new items if lost. Additionally, they suggested making the magnet more visible.

Learning Experience Regarding the learning experience, some users experienced difficulty in locating certain items, such as the magnet, and it was not immediately evident which chemicals should be discarded during the cleaning process.

General User Experience Users provided general comments about their experience with the application. On the negative side, users stressed the translation problems and the tasks which hampered the experience, such as the tracking issue. On the positive side, users praised the application. They described it as good one user said, "Fantastic project and great application! Well done" (translated from Dutch: Fantastisch project en prachtige applicatie! Goed gedaan).

AR Experience Finally, 11 participants reported never using an AR application, while four reported that the most common AR application they used was Pokémon Go.

Log Data and Process Maps

The analysis of the data logged revealed the potential difficulties for students while using the system. To better understand where users were struggling, a custom event, *Show Hint*, was analysed in detail.

The results from Figure 6.7 showed that users were given more hints in certain steps for each level. In Level 1, the steps "place beaker", "wear PPE", "attach burette", and "fill empty beaker" had the most hints. For Level 2, the steps "add indicator" and "clean acid" had the highest number of hints. Lastly, in

Level 3, the step with the most hints was "clean acid". Note that the hints were automatically triggered after 3, 5, and 7 unsuccessful attempts on each level. As a result, in Levels 2 and 3, users were likely to receive fewer hints than in Level 1.

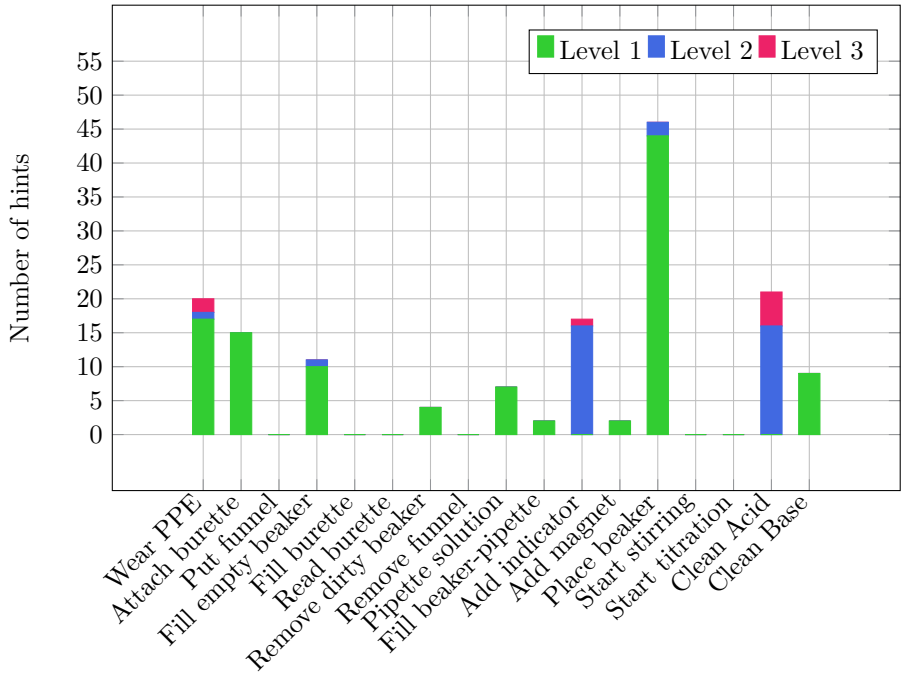


Figure 6.7: Number of hints per step given to KU Leuven students. After 3, 5, or 7 unsuccessful trials in each level, a hint is triggered automatically. The graph shows the steps in the sequential order observed in the titration process of the app.

The custom event *Done Process* was investigated to explore the order of tasks performed by the participants. This event would be recorded upon completion of a step in the procedure. Examining the process maps allowed us to determine which stages of the process participants performed actions in a different order compared to what was instructed. This could be identified by observing diverging arrows originating from the same event. For instance, in Level 1 (as shown in Figure 6.8), the participants deviated from the procedure after the steps of "put funnel", "remove funnel", and "read burette". In Level 2 (shown in Figure 6.9), the diverging arrows originate from the steps of "remove dirty beaker" and "put funnel". In contrast, in Level 3 (shown in Figure 6.10) participants only deviated after the step of "read the burette".

The *Reposition Materials* event is registered when the user clicks the option on the side menu and is integrated into the system to resolve the issue of virtual objects being transferred to another place or the system losing track of the environment. Process maps showed that this event was primarily triggered before or after the "wearing PPE" stage (at the start of the level) and after the "put funnel" step on all levels.

The *Combine Wrong* custom event identifies which steps users make errors in when attempting to combine objects that are not allowed. Figure 6.8 shows that the most wrong combinations were made in Level 1 before "place beaker", "pipette solution", "fill burette", and "fill empty beaker". Similarly, in Level 2 (Figure 6.9), the majority of wrong combinations occurred before "fill burette" and "fill empty beaker". Level 3 registered more errors, although in fewer cases than before, occurring before "clean acid" and "place beaker". This confirms what is stated in Figure 6.7, which indicates where users received more hints.

Another interesting custom event to analyze the learning process is the *Show Hint After Asking* event, triggered when the participant clicks the cartoon to receive a hint. Interestingly, this event was triggered in all levels before the "add indicator" step and "clean acid" or "clean base", possibly indicating that the participants felt unsure about either how to perform the action or their knowledge.

Participants during each level can access additional options in the side menu. One notable option is the "Open Theory", which allows them to check their theoretical knowledge. Data analysis of the process maps revealed that participants predominantly used this option before two steps; answering the final questionnaire or before completing the step of "add indicator". This could be an indication that participants feel uncertain about their prior knowledge.

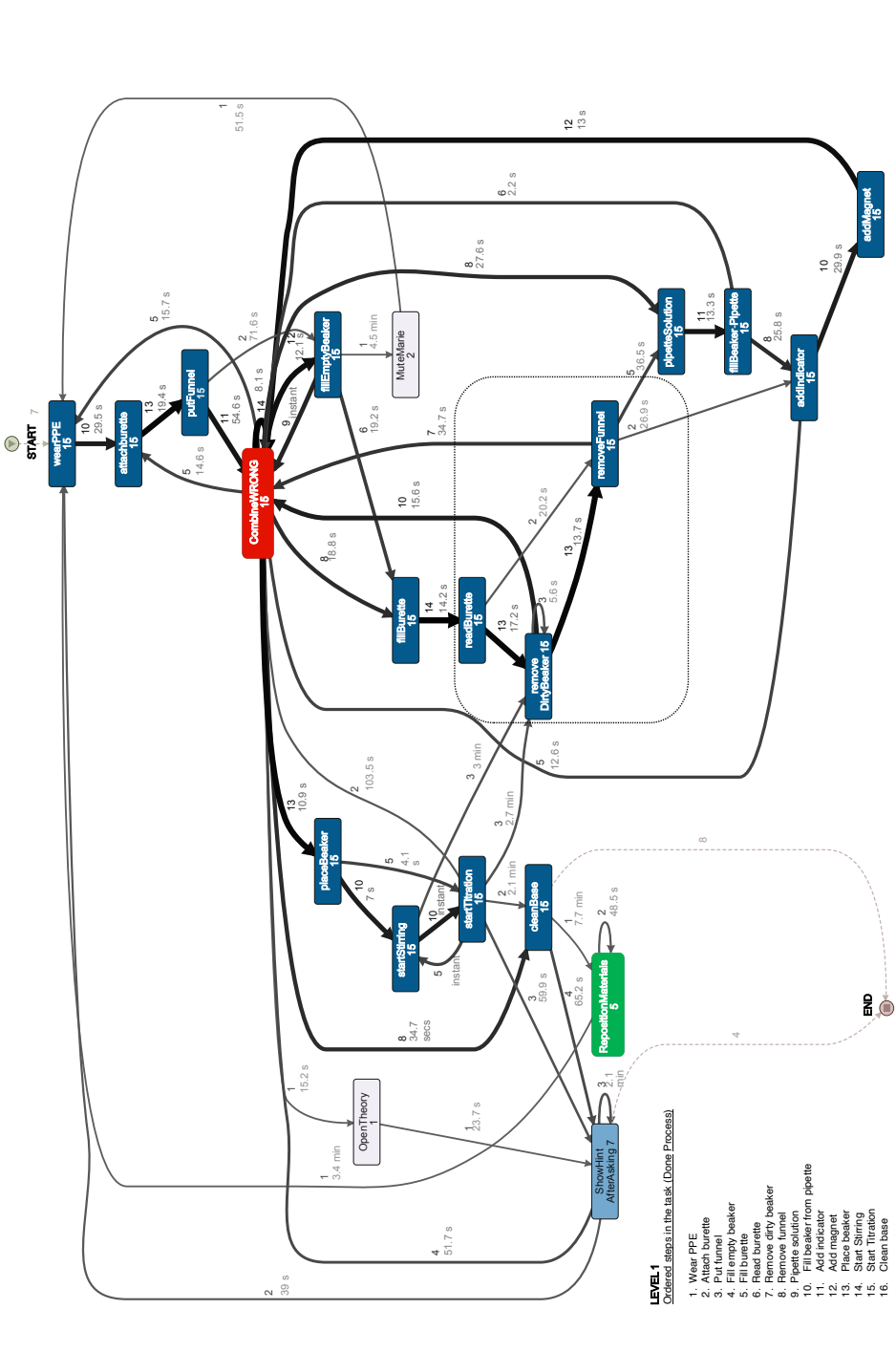


Figure 6.8: Process map of Level 1, recorded by DISCO metrics, displays 100% of activities and 35% of paths. The logged custom events are *Reposition Materials* (green box), *Combine Wrong* (red box), *Done Process* (dark blue boxes), *Open Theory and Mute Marie* (grey boxes), and *ShowHintAfterAsking* (light blue box). The frequency and duration of each event are indicated beside the arrows, while divergent paths are highlighted by enclosing different steps in boxes.

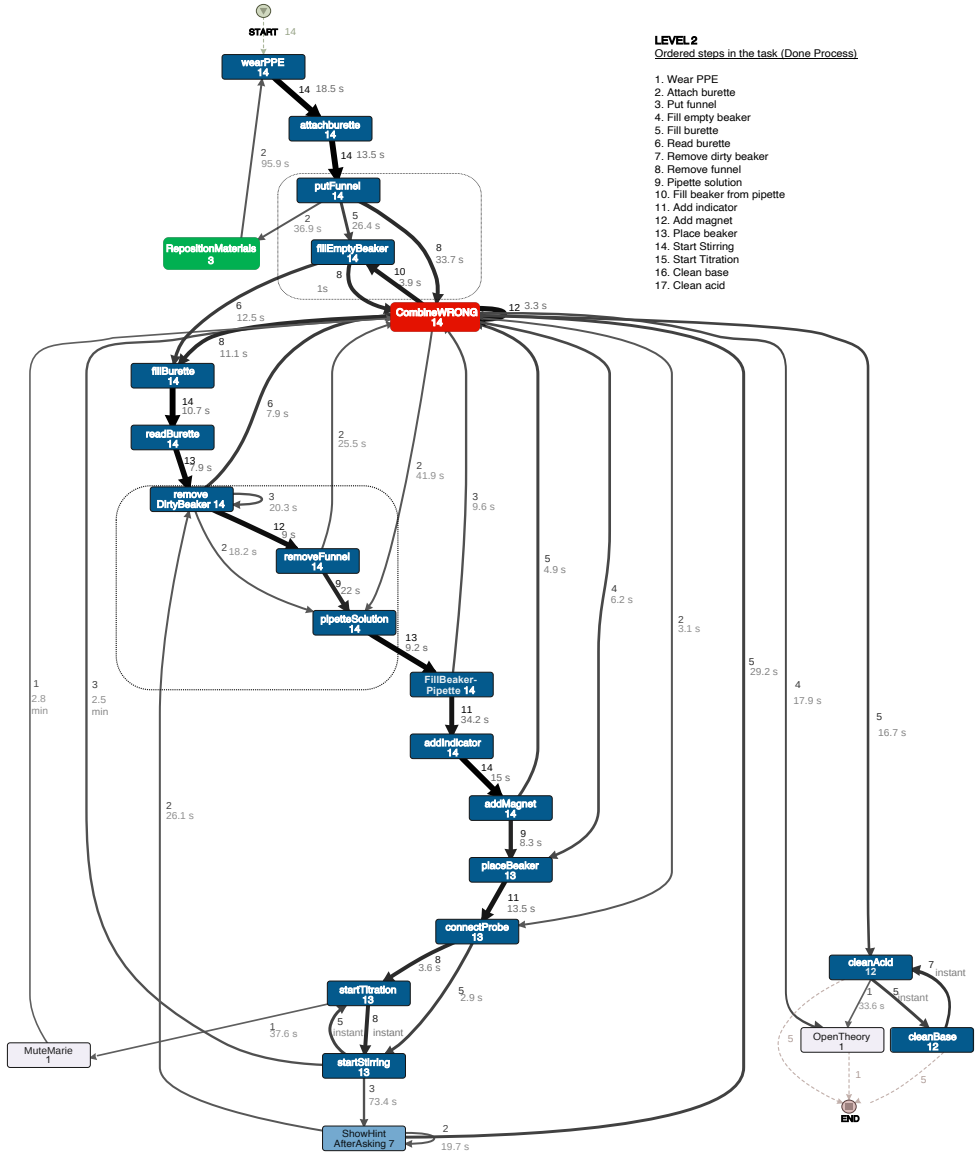


Figure 6.9: Process map of Level 2, recorded by DISCO metrics, displays 100% of activities and 35% of paths. The logged custom events are *Reposition Materials* (green box), *Combine Wrong* (red box), *Done Process* (dark blue boxes), *Open Theory and Mute Marie* (grey boxes), and *ShowHintAfterAsking* (light blue box). The frequency and duration of each event are indicated beside the arrows, while divergent paths are highlighted by enclosing different steps in boxes.

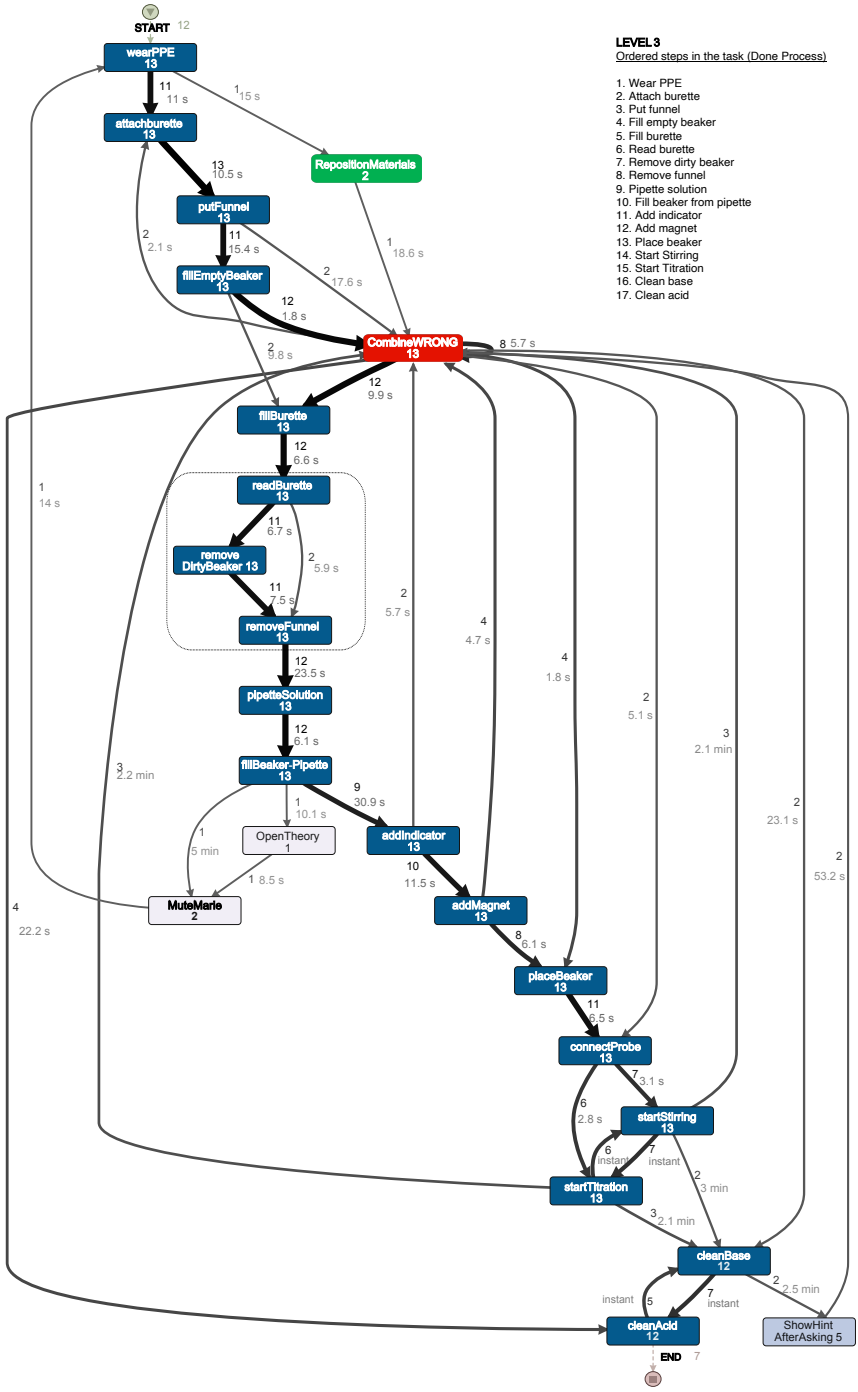


Figure 6.10: Process map of Level 3, displays 100% of activities and 35% of paths. The logged custom events are *Reposition Materials* (green box), *Combine Wrong* (red box), *Done Process* (dark blue boxes), *Open Theory and Mute Marie* (grey boxes), and *ShowHintAfterAsking* (light blue box). The frequency and duration of each event are indicated beside the arrows, while divergent paths are highlighted by enclosing different steps in boxes.

Learning Test

The learning test results were analysed to determine whether students enhanced their learning. The mean pretest score was 7.6 ($SD = 2.06$), while the mean posttest score was 8.8 ($SD = 1.47$), as seen in Figure 6.11. To investigate whether the posttest scores improved significantly, a one-tailed t-test at a significance level of $p < 0.05$ was conducted, as the data showed a normal distribution (see Table D.2). The test result ($t(14) = -2.276$, $p = 0.020$, $d = -0.59$) provided sufficient evidence to reject the null hypothesis and suggest that the mean score is 0.59 standard deviations greater than the mean of the pretest. Additionally, the effect size expressed in CLES indicates that it is 68.4% likely that the evaluation of the posttest is higher than the evaluation of the pretest when a pair of scores is randomly selected [121, 122]. Note that the evaluation did not assess learning retention. It is important to note that the results may differ if the students take the test at a later time.

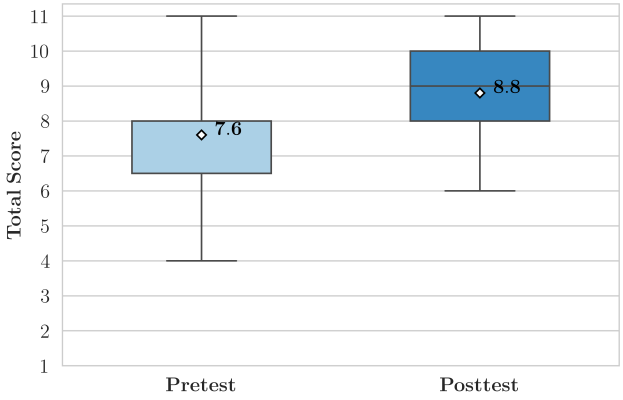


Figure 6.11: Boxplot showing the pretest and posttest scores of the KU Leuven students. The boxes represent the interquartile range (IQR), with the median of the scores indicated by the line inside the box and the mean scores indicated by the rhombus. The pretest scores had a mean of 7.6 ($SD = 2.06$), while the posttest scores had a mean of 8.8 ($SD = 1.47$). Results of the one-tailed t-test showed a significant result ($t(14) = -2.276$, $p = 0.020$, $d = -0.588$)

Performance

Two metrics were assessed to determine if the participants' performance improved throughout the levels: the total amount of time spent on each level and the number of errors made while assembling the titration equipment. It is expected that completing the same procedure will take a similar amount of time, while

with experience, the participants will reduce the assembly errors. Therefore, a decrease in these performance metrics would indicate an improved performance.

The total time spent per level was recorded to be 18.66 min ($SD = 5.29\ min$) for Level 1, 10.54 min ($SD = 3.47\ min$) for Level 2, and 9.17 min ($SD = 2.69\ min$) for Level 3 (Figure 6.12, Table D.4). The results of the Friedman test rejected the null hypothesis, indicating a statistically significant decrease across levels ($X^2F(2) = 14.308, p < 0.001, Kendall's\ W = 0.550$) (see Table D.5).

Conover's post hoc tests further revealed that participants had significantly reduced their total time from Level 1 to the next (see Table D.6); specifically, from Level 1 to Level 2 ($T-Stat = 2.942$, $p_{bonf} = 0.014$), from Level 1 to Level 3 ($T-Stat = 3.530$, $p_{bonf} = 0.005$), and from Level 2 to Level 3 ($T-Stat = 0.588$, $p_{bonf} = 0.562$) there was not a significant improvement.

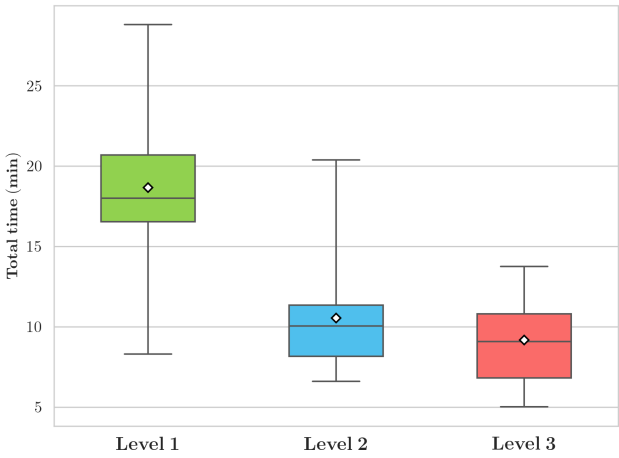


Figure 6.12: Boxplot comparing the total time spent on each level: Level 1 18.66 min ($SD = 5.29\ min$), Level 2 10.54 min ($SD = 3.47\ min$) and Level 3 9.17 min ($SD = 2.69\ min$). The Friedman test indicated sufficient evidence ($p < 0.001$) to reject the null hypothesis that there is no difference between the treatments. This indicates that the total time spent per level is not all equal, and there is a statistically significant difference between them, also confirmed by the post hoc Conover's test.

On average, participants made 26.38 ($SD = 22.10$) errors in Level 1, 13.77 ($SD = 19.24$) errors in Level 2, and 5.54 ($SD = 4.82$) errors in Level 3 (Figure 6.13, Table D.7). The assembly errors comparison resulted in a $X^2F(2) = 12.776, p = 0.002, Kendall's\ W = 0.491$, suggesting a significant decrease across levels 1 and 3 only, as confirmed by the Conover's post hoc tests: Level 1 vs

Level 2 ($T - Stat = 2.228$, $p_{bonf} = 0.107$), Level 1 vs Level 3 ($T - Stat = 3.545$, $p_{bonf} = 0.005$), and Level 2 vs Level 3 ($T - Stat = 1.317$, $p_{bonf} = 0.601$) (see Tables D.8 and D.9).

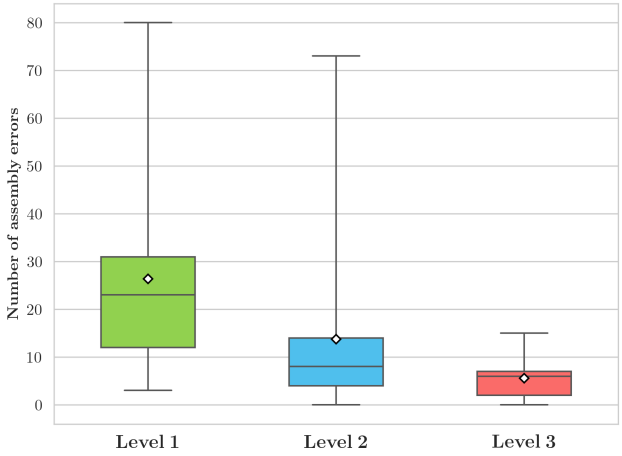


Figure 6.13: Boxplot comparing the number of assembly errors per level reveals a statistically significant difference between the treatments, with Level 1 having a mean of 26.38 ($SD = 22.10$), Level 2 13.77 ($SD = 19.24$) and Level 3 5.54 ($SD = 4.82$). The Friedman test ($p = 0.002$) provides sufficient evidence to reject the null hypothesis, indicating that the number of assembly errors per level is unequal. Furthermore, Conover's test revealed that the participants only significantly reduced their errors from Level 1 to Level 3.

Cognitive Load

The reported cognitive load, as measured by the R-TLX questionnaire, yielded an average score of 30.11 ($SD = 18.62$) (Figure 6.14). This result places the cognitive load in the first quartile (38.0) for cognitive tasks (i.e., tasks that require a mental action) and above the median (27.95) for mechanical tasks (i.e., assembly tasks) according to the reference data reported by Grier [1]. A total of 1173 reported workload scores were analyzed by Grier, and her results showed that 80% of them were between 26.08 and 68.00. Table 6.1 compares the scores reported by Grier and the average scores obtained for MAR Lab, indicating that the scores obtained in the evaluation were below or close to the reference range, suggesting that no high cognitive load was observed in the test. As a result, the cognitive load due to the learning tasks in MAR Lab can be assessed as moderate.

Table 6.1: Average R-TLX scores and cumulative data distributions of scores per task reported by Grier [1]

Task (cases)	Min	25%	50%	75%	Max
MAR Lab (15)	15.00	23.33	28.33	35.0	53.33
Cognitive tasks (31)	13.08	38.00	46.00	54.66	64.90
Mechanical tasks (22)	20.10	24.90	27.95	33.68	51.03

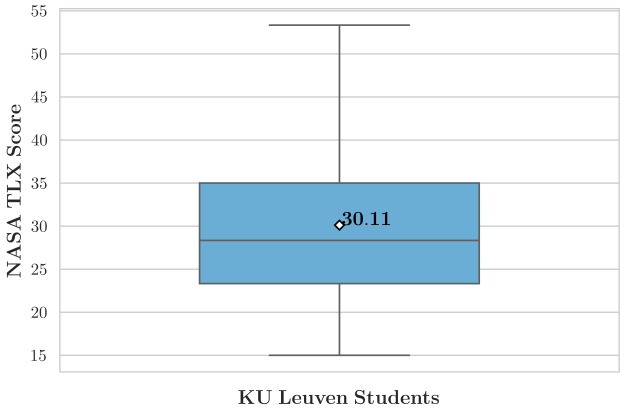


Figure 6.14: Boxplot of R-TLX average score of KU Leuven Students. The box represents the Interquartile Range (IQR) and highlights the median score (indicated by the line inside the box) and mean score (30.11 $SD = 18.62$) indicated by a rhombus. The quartiles for this distribution are 23.33, 28.33 and 35.0.

Bayesian Informative Hypothesis Testing

The dependent variables were determined as follows: knowledge acquisition was calculated from the pretest and posttest results, time-on-task was calculated as the total amount of minutes spent by each user on Level 1, and perceived cognitive load was calculated as the average of the R-TLX data.

The descriptive statistics and the covariance matrix results are displayed in Table D.10 and D.11. Looking at the confidence intervals, none of the covariances shows any significant relationship.

The posterior model probabilities (PMPs), also known as Bayesian error probabilities, indicate which hypothesis receives the most support from the data. If the proposed hypothesis receives more support than the unconstrained hypothesis (H_c), it can be confidently affirmed that the supported hypothesis is not simply the most desirable among many inadequate options [172].

Based on the Bayesian informative hypothesis testing, the results show that there was strong support for H_1 ($PMPc = 0.888$) compared to H_2 ($PMPc = 0.022$) and H_c ($PMPc = 0.090$), indicating that it is highly unlikely that other effects are at play.

These results support H_1, which suggests that cognitive load likely influences both time-on-task and knowledge acquisition: the higher the cognitive load, the less improvement in knowledge acquisition and the more time spent on the application. Moreover, the results indicate that overall knowledge acquisition increases after using the application and is influenced by time-on-task; the more time spent, the more improvement in knowledge.

The main parameter of the suggested model - knowledge acquisition - provides an effect size of $d = 0.59$ (see Table D.11), indicating that students are generally learning from the application.

6.6 Discussion

This chapter presented the final evaluation of the MAR Lab digital prototype. The evaluation included a comprehensive examination of usability, feedback from target users, and an analysis of first-year chemical engineering students' cognitive load, performance, and knowledge acquisition.

The mixed-method approach utilized for this evaluation enabled the observations to reinforce one another, highlighting weaknesses. Quantitative analysis was utilized to assess the learning effect of the platform's use and measure the application's usability and users' performance. Meanwhile, qualitative analysis was employed to gain a more in-depth understanding of the participant's experience with the application. Furthermore, Bayesian informative hypothesis testing allowed for analyzing probable relationships between the parameters studied.

The overall SUS score improved to 78.0 ($SD = 13.42$) after the last evaluation, which is considered to be "good" by the adjective scale [118]. Although this score can be improved, it still indicates acceptable usability. This is important to note, as research has shown that usability can positively impact interest, sense of control, and intended use [87, 157]. Thus, this thesis sought to achieve an acceptable usability score.

Regarding the general perception of MAR Lab, users provided direct feedback via the open-ended posttest question. Overall, most users reported that this was their first experience with an AR application and found it to be generally good. However, the participants also shared the difficulties they experienced.

The most common issue was the environment tracking, even after using the markers to help with the tracking during the evaluation (see Figure 6.4). This is consistent with earlier observations in Chapter 3, which showed that educational AR application users often face tracking difficulties. Moreover, analysis of the custom event *Reposition Materials* revealed that users mainly experienced this problem at the start of the levels or when during actions which required them to move the phone farther away from their point of view, such as when they needed to put the funnel into the burette. Interestingly, similar findings were also observed in the formative evaluations described in Chapter 5.

These observations have two main implications. Firstly, they suggest that the tracking problem is a common problem in AR experiences and thus must be a priority to solve in the design and development of future AR applications. Secondly, since tracking can be unstable for various reasons such as poor lighting conditions, fast movement, or occlusion [45], design guidelines help users find the best conditions, considering the limitations, must be part of the onboarding process of any AR application. MAR Lab's onboarding process specifies the lighting conditions and speed of tracking the environment. However, it may be the case that these instructions are not clear enough for users or that the conditions where each student performed the test may not have been optimal, thus creating the tracking issue more pronounced.

Suggestions were made to improve progress and convenience, make materials more visible and add functions to skip steps or recover lost items. However, not all options are desirable from an educational standpoint, as skipping a step in the procedure could potentially impact the quality of knowledge the students acquire. Nevertheless, some of these suggestions could help to enhance the user experience of MAR Lab without compromising the learning. In fact, the open-source version of the TrainAR framework [1] facilitates the resizing of objects, providing a reference setup.

Similarly to what was observed in Chapter 5, problematic interactions analysed through the log data and process maps indicated that users performed the most wrong interactions in the steps "place beaker", "pipette solution", "fill burette", and "fill empty beaker". Several factors can help explain why users experience difficulty handling virtual objects in these steps.

Firstly, the errors related to the "place beaker" task can be attributed to the users' difficulty in selecting, grabbing, and combining objects with secondary objects. A lack of space in the virtual environment and unwanted collisions between virtual elements may cause this. To prevent similar errors from occurring in the future when implementing AR scenarios using the TrainAR framework, it

[1] https://jblattgerste.github.io/TrainAR/manual/TrainArObjects.html

is important to ensure that there is ample space to spread out virtual elements and/or scale the objects down to the point that collisions are minimized.

Secondly, the problems with the steps "pipette solution" and "fill empty beaker" can be explained by the interaction itself. Both steps require the user to "open" a bottle or a flask in the AR scenario. Although the interaction is intended to simulate a real-life situation, the user may not be used to performing this action in a virtual setting; hence, explicit instructions are necessary to help the user complete the task. However, too many instructions may be counter-productive and reduce the authenticity of the environment. Therefore, a balance between the necessary actions for learning and other actions to increase the realism of the environment should be sought to enhance the learning experience.

Finally, the errors associated with the task of "fill burette" can be attributed to the difficulty of accurately targeting a desired object with a crosshair, which may require a good understanding of the spatial relationship between two-dimensional projections and the real world [140]. This difficulty can also be linked to the limited experience of users when interacting with AR.

The thematic analysis of the learning experience revealed several areas of uncertainty for users. These unclarities were also identified by analyzing the custom events *Show Hint After Asking* and *Open Theory*. They were distinguished from errors in that users would seek an answer within the system before taking action. In MAR Lab, process maps showed that users asked for hints before selecting an indicator and during cleaning. Furthermore, users consulted the theory before answering the final questions and adding an indicator.

This observation is interesting for several reasons. Firstly, it could suggest that users are not only undertaking procedural steps but exploring the environment, thus performing cognitive tasks simultaneously. Secondly, the steps in which these actions were seen could be enhanced for increased clarity; this could be achieved by incorporating more theory-based information or adding more steps to the cleaning process. However, there is the potential that this could increase the amount of time or cause frustration due to the additional instructions. Finally, these actions could be seen as evidence of how a user who spends more time on the application might improve their knowledge.

Process maps have revealed a potential discrepancy in how a student performs the steps. Even though the application provides instructions and hints, and the steps are identical across the three levels, some participants were observed to diverge from this linear procedure. Due to the low sample size in the evaluation, it is difficult to observe clear patterns. However, the process maps indicate that students seem to deviate in the levels when an interaction or a combination with

a non-secondary effect (such as an animation after combining two objects) takes place. It is possible that students at these steps do not pay attention to the instructions given because they are simple actions and may be prioritizing their actions based on what is closest or their experience of the experiment. Also, it is possible that the cognitive load of reading, remembering or listening to the instructions while carrying out the required tasks could lead the participants to opt for a simpler or more intuitive route. This implies the design could be improved by simplifying the steps, such as introducing the users to these simple actions first and then omitting the explicit instructions later. If the step does not need to be done in a specific order, it can be approached this way.

The reduction in total time and assembly errors were assessed to investigate how efficiently students can apply the knowledge acquired from the app to other levels. The results of users who played the three levels suggested an improvement, as demonstrated by the decrease in the total time spent on each level. These results suggest that users seemed to learn the procedure and could potentially perform better with practice. Similarly, users also decreased the number of assembly errors from level 1 to level 3. However, these results should be taken cautiously as this metric also showed a high coefficient of variation.

Regarding the interaction between cognitive load, performance, and knowledge acquisition, the Bayesian informative hypothesis testing supported hypothesis H_1. This indicated that, despite cognitive load increasing with the time spent on the application and thus negatively impacting knowledge acquisition, the model still suggested that learning could improve with more time spent on the application. Interestingly, this outcome suggests that the cognitive load users perceive during the AR experience may not necessarily be detrimental to knowledge acquisition. A possible explanation for this may come from the motivating characteristics of immersive environments, which could help mitigate possible overload effects on students [157]. This could be supported by the praised comments some participants shared about the application. In addition, the average cognitive load reported by the participants suggests that the environment itself did not impose an extreme load. This is noteworthy since most students had not used an AR application with similar interaction as MAR Lab; thus, the instructional design likely avoided imposing additional load due to the novel interactions.

The results of the Bayesian approach are consistent with other observations, indicating that students likely improve their learning when using the application. This is further supported by an individual NHST, which confirmed a significant result with a medium effect size [120], as well as an improvement in users' performance across levels. These findings align with several meta-analyses that have reported that learning effectiveness while using AR has a medium effect size [27, 26, 24, 25, 149]. It should be noted that the evaluation did not

determine whether the learning had been retained. Thus, the outcome of the test may vary if it is taken at a different time.

6.7 General Recommendations for Practitioners

Several recommendations for practitioners interested in using a similar approach to mobile augmented reality can be derived from the results of this study. These recommendations also align with some of the principles of the 4C/ID model introduced in chapter 2.

Support Learners with additional information Providing students with access to additional information, such as hints, can be especially useful when performing an action in unfamiliar environments, such as AR. This approach can support students by reducing an extraneous load of unknown interactions and supporting them with the content. This recommendation aligns with the redundancy principle, which advocates for provisional information to support learners.

Optimize the AR environment space Optimize the AR environment by taking into account the actions that the user will perform and consider the potential obstacles that students may encounter due to a lack of space or tracking issues caused by fast movement or position of the phone related to the point of view.

Balance the realism of virtual actions When designing AR scenarios, it is important to balance the essential steps needed to simulate the real world accurately and the unnecessary steps that may add burden without a meaningful learning purpose. While some actions may be required to create a realistic environment, they may also lead to frustration. To ensure students have a meaningful and educational experience with the AR environment, it is important to evaluate the learning effects of each action and determine which ones are necessary for a realistic simulation while also being beneficial to the learning process.

Support users with tracking issues and unfamiliar AR interactions One of the key challenges of AR is its unstable tracking and lack of familiarity among users. To ensure a successful AR experience, it is recommended to provide support to help users understand how to track the AR environment and identify the most effective tracking conditions. Align with the dynamic visualization principle; this can be done through, for example, a pre-training scenario demonstrating

AR-related actions. Doing so will enable users to become more comfortable with tracking and AR interactions, thus enhancing their AR experience.

6.8 Limitations

The research presented in this chapter is subject to several limitations. First, the small sample size of the study may limit the findings, and there is a probability that the observed effects may not reflect a true effect. To mitigate this, a mixed-method approach was employed, and formative evaluations were conducted in previous chapters to minimize the chances of false effects. Moreover, the Bayesian approach allows for data to be updated to support the findings and facilitate replication [170]. Replication studies with larger sample sizes or meta-analyses can further corroborate the results found in this chapter. Second, there was a lack of diversity among the participants, and all included in the data were male, which could introduce bias. It is important to consider this when interpreting the results, and further evaluations should address this factor to study diverse groups. Third, the participants were recruited voluntarily; thus, their feedback may be biased due to their interest in the study. It is essential to consider the potential bias in their responses. Fourth, although this study did not compare the cognitive load among different groups or conditions, the limitation of not considering working memory as a control variable in the model is acknowledged [159, 158]; this variable may have an effect on the measurement, potentially resulting from the load imposed by the posttest or other factors, thus influencing the overall cognitive load of the participants. Finally, the MAR Lab app was designed with a specific target group and topic in mind (i.e. titration); thus, any conclusions drawn about augmented reality or science education, in general, are limited to that particular topic and target group.

6.9 Conclusion

In conclusion, the summative evaluation of MAR Lab showed that the perceived usability of the application was improved. The prototype can be characterized as having good and acceptable usability, with students mainly experiencing problems inherent to the technology (i.e. tracking). The model that explored the interaction between performance, cognitive load and knowledge acquisition suggested that, though cognitive load increases with prolonged use of the application, negatively impacting knowledge acquisition, learning still increases with more time spent playing with MAR Lab. Moreover, the cognitive load imposed on the students was moderate, and overall, they seemed to learn from

the application. These results suggest that MAR Lab, as a tool, can serve chemical engineering students without access to a laboratory to learn about titration in an interactive and augmented manner. This chapter explores some of the multiple interactions that may be at play when studying immersive environments. Future studies with larger groups or in realistic scenarios could support the described findings and expand upon researching this complex interplay of interactions.

Chapter 7

Conclusion and Outlook

This thesis aimed to investigate **how mobile augmented reality can be used to support learning chemical practical skills**. To this end, the Design and Implementation Framework (DIF) was employed to guide the development of a functional prototype whose basic requirements were established through an exploratory study (Chapter 3). Rapid prototyping was used to create the first digital implementation of the prototype (Chapter 4), followed by subsequent iterations and formative evaluations (Chapter 5). Finally, a summative evaluation of the prototype, MAR Lab, was conducted, focusing on exploring the cognitive load experienced by students while using the application and the interaction between performance and knowledge acquisition (Chapter 6). The key findings and conclusions of the research are outlined in this chapter.

7.1 Mobile Augmented Reality Apps in Education

To define the basic requirements for the AR learning experience, an exploratory study was conducted to analyze educational mobile augmented reality (MAR) applications from the Google Play Store and synthesize user reviews into insights for improvement. This research found that most users of educational MAR apps had various problems that hindered the experience and that only a few apps were highly installed among users. These problems included technical hurdles such as device incompatibility, high power consumption, and unstable tracking. Additionally, marker-based AR was found to be the most popular type of AR in the app store, with several shortcomings, such as the need to download and print markers and general dislike from the public. Users also

requested additional content and interactivity, as most apps were merely used to visualize virtual objects. Finally, it was found that most of the apps targeted childhood science education.

7.2 First Prototype of MAR Lab

The initial implementation of MAR Lab's digital prototype (MAR Lab 1.0) yielded valuable insights for future iterations. Despite being preliminary, this evaluation revealed that students could use the application independently, regardless of their familiarity with AR technology. The students' positive feedback indicated that they valued the tool while highlighting improvement points, such as simplifying the user interface, improving the task sequence and instructions, and addressing issues with the virtual content. Just as the results from the exploratory study suggested, this evaluation uncovered certain challenges related to tracking and device incompatibility, indicating that AR methods, in general, still have general problems that are beyond the scope of this research yet are nevertheless essential to address. Lastly, the learning test was used as a pilot for further improvement.

7.3 Formative Evaluations

The second iteration of the digital prototype (MAR Lab 2.0) was improved in several ways: it was extended to include three levels that better contextualized the experience according to Novak's theory [64], design elements were refined to enhance appearance, interaction, and instructional help from the cartoon (i.e., Marie Curie), and the technical implementation of the titration calculator was improved to increase the scalability of the levels. The result of the prototype was rated as "good", with a usability score of 71.98 (SD=15.21).

Mixed results were obtained from the formative evaluation of the participants' learning outcomes using MAR Lab. The operators' group of the ACTA test showed statistically significant learning gains, while the students' group did not. This test was utilized to refine the learning test; thus, the results should be interpreted as merely indicative. The Merck test findings were not statistically significant; however, a potential ceiling effect may have hindered the results of this test. Despite this, the group reported that recalling the procedure was beneficial. The KU Leuven test revealed a statistically significant improvement in learning; however, due to the small sample sizes, these results should be interpreted with caution.

Log data and process mining also revealed that users had difficulty with stable tracking and handling virtual objects (i.e., floating or disappearing materials). They struggled with actions requiring them to move their phones farther away from their view. Despite these challenges, users expressed positive aspects of their experience with the application. They suggested that the improved application could be useful, as a tool, for students learning about titration in different settings.

Overall, the formative evaluations prompted modifications for the final prototype, which addressed issues such as incorrect translations, lack of realism and consistency in certain actions. In addition, the procedure was simplified by removing unnecessary steps and adding signalization for important steps, like "closing the burette". Furthermore, new features, such as the ability to choose between hearing or muting the agent's voice, a review menu, and a new level for unrestricted play, were implemented to enhance usability and promote a better learning experience.

7.4 Summative Evaluation

The final evaluation of the digital prototype (MAR Lab 2.1) was focused on usability, participants' feedback, and the interaction of novice chemical engineering students with the app. The evaluation also assessed the cognitive load and in-app performance experienced by students who lacked access to a chemical laboratory in their course, which was the intended audience for whom the application was designed.

The results show that the final prototype of MAR Lab was improved, ending in a "good" and acceptable SUS score of 78 ($SD = 13.42$). Tracking the AR environment was found to be the most difficult task for students; however, they independently used the application and provided positive feedback on the implementation. Analysis of the interactions with MAR Lab suggested that the students improved their performance with more levels played, and errors decreased as the experience and AR interactions became more familiar. When looking at the relationship between cognitive load, performance and knowledge acquisition, the Bayesian approach further supported the fact that although cognitive load likely increases with prolonged use, knowledge acquisition also increases with more time using the application. Moreover, the cognitive load imposed on the students was moderate, possibly due to the instructional design principles used, and overall, they seemed to learn from the application. These results suggest that MAR Lab, as a tool, can serve chemical engineering students without access to a laboratory to learn about titration in an interactive and

augmented manner. However, further replication studies should be conducted to confirm these observations due to the low number of participants.

Future research could be conducted to explore the extent to which the knowledge acquired while playing with MAR Lab can be transferred to new learning scenarios. Experimental designs that accommodate different transfer activities (i.e., to a real laboratory) could help further investigate this issue. Additionally, experiments using MAR Lab to predict laboratory performance (i.e. with regression models) could provide interesting insights to help improve the laboratory skills of science students.

MAR Lab can be analyzed from the perspective of embodied cognition. In this area of research, most studies have taken a closer look at how VR can facilitate enactment, given its immersive and realistic qualities; however, this perspective is rarely used for AR. By utilizing the taxonomy for embodied learning proposed by Johnson-Glenberg et al. [175], MAR Lab can be characterized and compared to other scenarios to investigate how embodiment impacts learning.

In this thesis, one of the limitations is the assessment of learning using traditional tests. Therefore, as appointed by Johnson-Glenberg and Megowan-Romanowicz [176], new assessment methods aligned with the activities performed in a virtual environment, such as MAR Lab, may be more suitable. For example, creating an assessment level in which students are prompted to experiment without support or to create their own experiments and setups with a given AR scenario could provide deeper insights into the learning process.

Finally, exploring the application of the tool in realistic scenarios is also interesting. For instance, interviewing students and teachers who use the tool for studying or learning about the topic for the first time, or using it in a classroom as part of the curricular activities, can bring to light both the benefits and challenges the tool still has to overcome. This can be useful for further development and implementation of similar environments in chemistry and other disciplines.

7.5 Outlook of MAR Lab

The MAR Lab app was designed as a markerless, interactive mobile augmented reality tool for training students in titration experiments. However, as suggested by participants, the application can be extended and improved in several ways.

First, additional chemical experiments can be created to train students in basic laboratory procedures, such as distillation, organic synthesis, precipitation, and filtration. Also, these experiences can include standard procedures such as

correctly using pipettes, preparing solutions etc. Because the application is designed as a complementary tool, generic experiments will be more valuable so that diverse curricular activities can be combined with it.

Second, the cartoon character (i.e., Marie Curie) could be modified to be more interactive. With the advancement in artificial intelligence (AI), this feature can be improved by providing a personalized experience to the students by being more responsive or with the ability to access the internet [177]. Also, the character can be personalized and chosen by the user, boosting learners' motivation and impacting performance [178].

Third, the complexity of the levels can be adaptive to the users' performance, providing a personalized experience and enhancing learning (individualization principle [66]). For example, if a user performs well, the application can increase the difficulty of the challenges or ask different questions; alternatively, if the user is struggling, the difficulty can be reduced to ensure that the user does not become frustrated.

Fourth, aligned with the collaborative principle of the 4C/ID model [66], collaborative tasks can be added to the application to enhance the learning process. These activities can be within the AR system or as a learning strategy that adheres to constructivist principles [179]. Specifically, in the lab context, this approach can simulate a real lab setting, allowing students to interact with other students or the instructor. For example, MAR Lab can be enhanced to enable multiple users to view and interact with the same objects simultaneously or to modify the experience so that one student experiments and the others provide support, respond to quizzes, or collect observations.

Finally, MAR Lab has been proposed in this thesis as a potential tool to support student learning and provide access to the psychomotor domain that would otherwise be lacking due to the absence of laboratory experience. Aligned with the cognitive load theory, pre-training with MAR Lab can prepare learners for new tasks and reduce cognitive overload, thus increasing their cognitive capacity to engage with new cognitive processes [126]. Meyer et al. [180] have further demonstrated that pre-training with immersive learning environments positively affects knowledge retention, transfer, and self-efficacy. Thus, pre-training with MAR Lab can potentially help students to be more efficient and attentive to chemical phenomena when conducting experiments in the laboratory.

7.6 General Insights for Developing AR Learning Tools

The development of AR tools for learning requires an interdisciplinary effort from developers, designers and educators to work together to create useful tools for students. From the research discussed in this dissertation, general insights can be derived for the related stakeholders.

For developers, the evaluation of the AR prototype has confirmed what Blattgerste et al. [31] had previously identified about interaction through buttons on the screen. Students could complete procedural tasks and interact with virtual objects using the TrainAR framework's combine, grab and interact actions with ease, as evidenced by the users' positive feedback and the acceptable usability score. However, problems related to tracking remain a major obstacle to providing a smooth AR experience. To address this, it is recommended to incorporate a support system that assists users in understanding how to track the environment and identify optimal tracking conditions. This can be achieved by introducing pre-training modules, simulations, and other methods of demonstrating AR-related actions. Additionally, leveraging log data and visualizing it through process mining tools can assist in the development process by uncovering usability problems that can help inform the development process.

For designers, the evaluation of MAR Lab emphasized the need for striking a balance between the essential steps required to simulate the real world accurately and those that may unnecessarily add burden. It also became evident that mobile device limitations, such as small screens or poor tracking performance, should be considered when designing AR scenarios. Large objects were found to be more difficult to interact with on small screens, thereby highlighting the technology's inherent limitations (tracking, hardware processing, etc.). Additionally, due to the unfamiliarity with AR interactions, support must be included in the design to facilitate a smooth experience. Finally, for a better experience with AR, user preferences should be considered. Factors such as extra audio, access to information, queries for support, preferred modes of use (standing or sitting), and language were all found to influence the user experience; thus, providing flexibility in these areas could improve the overall experience.

For educators, the evaluation of the AR prototype revealed the potential of mobile augmented reality to provide holistic learning experiences. From a learning perspective, while students may be able to learn with similar tools to MAR Lab, individual learners may have different requirements. For instance, some students may require a longer period of time to adjust to new technology, while others have no problems. Thus, educators must take this into account when employing AR for learning, accounting for any differences in familiarity

among their students. Additionally, AR can be implemented in various ways- not all of which may be necessary or applicable in every situation. For example, augmenting the student's physical environment with a simulated laboratory can benefit those without access to a real laboratory or when studying occurs in various locations. Another possible use is when AR is employed in a flexible manner- the study found that students can access the application with their own devices simultaneously in the same classroom. This can be particularly useful for educators who conduct sessions involving a larger group of students at the same time. Alternatively, this approach may be ineffective for learners who already possess certain familiarity with the content, as the content may be simplified to accommodate all users- this may, in turn, take away from the level of detail more experienced learners may seek. Thus, it is clear that AR can be utilized as a tool to enhance learning to suitable levels, provided the context of the application is kept in mind.

7.7 Future of AR in Education

Augmented reality continuously develops, with new, interactive, and creative applications becoming increasingly available as technology advances. However, several challenges remain in implementing and using AR systems, as both students and teachers are still relatively new to the technology. Recently, Perifanou et al. [181] reported that the main barriers for teachers are scarce AR applications (i.e., languages and topics), lack and cost of resources, perceived difficulty in using AR, and teacher's lack of knowledge and digital skills. To overcome these issues, studies have suggested training teachers, allocating funds for digital infrastructure and developing tools for easier development [181, 182, 183].

Additionally, although AR applications are available, teachers cannot often modify them or create content that aligns with their curricula [182, 181, 184]. To enable users without programming skills to create AR learning experiences, several authoring toolkits have emerged, such as Vuforia Studio, BlippAR, AWE, AR Media Studio, and Areeka [184]. More recently, TrainAR [1] has been released as a free, open-source tool for creating AR training, as shown in this research. Furthermore, by creating such tools, students and teachers have access to a more creative way to develop and use applications according to their needs, allowing them to boost content development.

Finally, the use of AR applications could also be improved by sharing resources on open platforms, allowing students, teachers, and other practitioners to access

[1] https://mixality.de/trainar/

and use them. This idea could be implemented similarly to the University of Colorado [2], where virtual simulations are readily available on their website.

Overall, AR holds great potential for enhancing learning, but there are still challenges that must be addressed for it to become a ubiquitous part of any learner's toolkit in the future.

7.8 Conclusion

How can mobile augmented reality be used to support learning chemical practical skills?

In this thesis, an example of a mobile augmented reality prototype was presented to demonstrate how AR technology, combined with instructional design, can support chemical engineering education.

Mobile augmented reality can be used to support the learning of chemical practical skills, providing students with an interactive experience of laboratory practice without the need to be physically present in one. Through virtual lab environments such as MAR Lab, students can assemble and work with various equipment, observe physical phenomena, and gain a deeper understanding of the processes and techniques involved in chemical experiments. Additionally, this technology can integrate information seamlessly and concurrently, providing real-time feedback to identify errors. Ultimately, engaging in these interactive experiments can help students develop their practical skills and gain a deeper understanding of the concepts.

[2] https://phet.colorado.edu/

Appendix A

Appendix of Chapter 3

A.1 Content of metadata extracted for each app

- a) Name of the app
- b) Description of the app given by the developers
- c) Number of installs, which indicates the number of users that have installed the app
- d) Average rating score, which is a value between 0 and 5 that indicates the average score given by the users
- e) The number of ratings, the number of total ratings provided by users. A rating from a user consists of the number of stars from 0 to 5 given by the user. Higher is better.
- f) The number of reviews provided by users. This number is not necessarily equal to the number of ratings
- g) Histogram showing the distribution of the ratings, i.e. the frequency of users that rated the app for each number of stars from 1 to 5
- h) Price and currency
- i) Size of the app, the number of megabytes (MB) that the app occupies in the phone's memory.
- j) Operative system (Android) version required.

- k) Information from the developer
- l) App category, the developer's chosen category, among 32 options to classify their app for ease of search.
- m) The version of the app and the date of the latest update.
- n) Uniform Resource Locator (URL) of the application for download.

A.2 ARdoc tool Architecture

The architecture of the ARdoc tool is comprised of 5 modules: 1) Parser, 2) NLP Classifier, 3) SA Classifier, 4) TA classifier, and 5) ML Classifier (Fig.A.1).

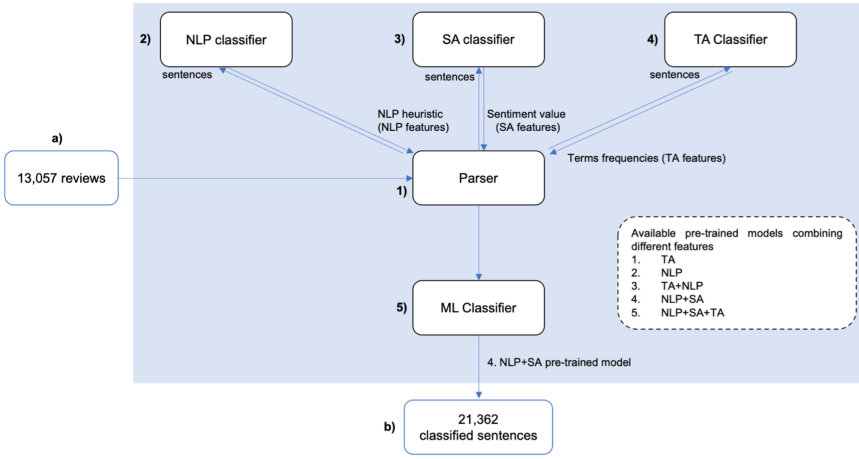


Figure A.1: The ARdoc tool architecture includes: 1) Parser module, 2) NLP Classifier, 3) SA Classifier, 4) TA Classifier, 5) ML Classifier (enclosed in the blue square). a) represents the input data that corresponds to the reviews of the 116 apps, and b) represents the output or the sentences after the classification done by the ARdoc. The dotted square lists the different pre-trained models that can be implemented with the ARdoc Tool. This study used a model that combines NLP features and SA features (NLP+SA). Figure adapted from [93]

1. The Parser module of ARdoc uses the Stanford Core NLP Library. This module prepares the reviews for classification by dividing each sentence as a token. The output of this module corresponds to individual sentences that are used as input by other modules.

2. The NLP classifier extracts the grammatical structures of the sentences by applying natural language processing (NLP) heuristics to identify recurrent linguistic structures. Panichella et al. [93] identified 246 recurrent linguistic patterns in user reviews and implemented them in the NLP classifier to detect them automatically. The output of this module is the correspondence of each sentence with an NLP pattern which is used by the ML module.

3. The SA classifier assigns a quantitative value depending on the sentiment of each sentence. ARdoc tool exploits the sentiment annotator available in the Stanford Core NLP and classifies each input sentence with a value between one and five. The output of this module corresponds to integer values for each sentence.

4. The TA classifier is responsible for processing the sentences by removing the stop words, stemming, and weighing each word for subsequent use. The output of this module is a term matrix that shows the importance (term of frequency) of the word contained in the sentence. This matrix is used as input by the ML classifier.

5. The ML classifier provides different methods for prediction based on pre-trained models using five different combinations of features extracted by the NLP, SA, and TA modules. It consists of a decision tree (the J48 algorithm) that classifies the sentence into one of the defined categories. This algorithm has previously yielded the best prediction results [95]. The classifiers were trained by Panichella et al. [93] using a set of 852 manually labelled sentences from seven popular apps.

A.3 ARdoc Categories

Table A.1: Categories defined by Panichella et al.[93] and used by ARdoc

Category	Description
Information Giving	Sentences that inform or update users or developers about an aspect related to the app
Information Seeking	Sentences related to the attempts to obtain information or help from other users or developers
Feature Request	Sentences expressing ideas, suggestions, or the need for improvement or enhancement of the app or its functionalities
Problem Discovery	Sentences describing issues with the app or unexpected behaviours
Other	Sentences that do not provide any useful feedback to developers.

A.4 Levels of education according to the International Standard Classification of Education (ISCED)

Table A.2: Levels of education according to the International Standard Classification of Education (ISCED) in its aggregated level

Aggregated level of education	ISCED 2011 Code System
Low education (Levels 0-2)	ISCED 0 = Early childhood education. ISCED 1 = Primary Education. ISCED 2 = Lower Secondary Education
Medium education (Levels 3-4)	ISCED 3 = Upper Secondary Education. ISCED 4 = Post-secondary non-Tertiary Education.
High education (Levels 5-8)	ISCED 5 = Short-cycle tertiary education. ISCED 6 = Bachelor's degree or equivalent tertiary education level.

Appendix B

Appendix of Chapter 4

B.1 Online Survey

Q1. Which operating system do you have installed on your phone?

- **a** Android
- **b** iOS
- **c** Other:

Q2. Do you have experience with AR applications?

- **a** Yes, which?
- **b** No

Q3. What is your current education level?

Q4. Does the application's language affect your motivation to use the application?

- **a** Yes
- **b** No
- **c** Sometimes

Q5. What are your language preferences?

- **a** Dutch
- **b** French

c English

Q6. The first time the user uses the application, everything is new: the structure and buttons. Would you prefer to receive a short explanation about the different functions and buttons during a first-time use?

a Yes, I would like to receive an explanation tour.

b Maybe, that depends on the application.

c No, I will learn by trying.

d Other:

Q7. We developed two different options to explain the functionalities to the user: one with cartoon and one without. Which version do you like best?



Q8. Do you want to tell us why you prefer that option?



Q9. The application will consist of several experiments focusing on different aspects of titration and with other difficulties. These levels can be sorted and presented in different ways. Choose the option that you like best.



Q10. Do you want to tell us why you prefer that option?

Q11. Choose the option that you like best.



In the application, there will be different buttons with pictograms. The functionality of these buttons should be clear. Choose the best-suited pictogram for the following functions according to you.

Q12. Do you want to tell us why you prefer that option?

Q13. "I want to quit this level and go back home."



Q14. "I want to receive a hint which is not content-oriented."

Q15. "I want background information," E.g., The difference between strong and weak acid.



Q16. "I want to receive a hint which is content-oriented."

Q17. "I want to review previously collected data."



Q18. Rank the following functions from most to least important in an educational game.

- Quit level and go back home
- Start all over
- Undo
- Redo
- Get a hint
- Review the assignment
- (Re)viewing the explanation

Q19. Choose from the following duo's the most readable option for you.



Q20. To guide the user during the learning process, multiple-choice questions will be integrated. For example, indicate if you agree or not with the following sentences.

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Integrating multiple-choice questions adds educational value to the application					
Integrating multiple-choice questions makes sure the application does not feel like a game.					
Easy experiments with much assistance would demotivate me to use the application.					
Easy multiple-choice questions would demotivate me to use the application					
Difficult experiments with limited guidance would demotivate me to use the application.					
Difficult multiple-choice questions would demotivate me to use the application.					

Q21. Do you agree with the following sentences?

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
I like to see my progress during the completion of a level.					
I like to receive a score with stars at the end of a level, like in a game.					
I like to know which effect a wrong answer to a multiple-choice question has on my progress.					
I like to receive a score every time I finish a level to try to do better next time.					

Q22. Which name would you give to the application?

Q23. Do you have any other suggestions?

B.2 Chemistry Test

The purpose of this pilot test was to evaluate the participant's comprehension of the theoretical principles, processes, and graphical representations associated with the titration experiment. The test was administered in English to assess their understanding of the subject. The test was administered as pre and posttest. In the posttest version, additional demographic questions were asked.

Q1. A titration experiment can be used to determine the _____ of base using a known concentration of acid.

a volume

b unknown concentration

c known concentration

Q2. At the equivalence point of a titration with strong base, the pH of a solution of a weak acid is _____.

a neutral

b slightly acidic

c slightly basic

Q3. At the starting point of a titration experiment, the pH of the solution _____ as the strong base is added to the solution of the weak acid.

a does not change

b quickly increases

c quickly decreases

Q4. The molecules of a weak acid such as acetic acid _____ dissociate in water.

a never

b fully

c partially

Q5. Strong base such as sodium hydroxide (NaOH)_____ when dissolved in water.

a fully dissociates

b does not dissociate

c partially dissociates

Q6. At what point will colour change occur when a suitable indicator is added to a titration?

a near the equivalence point

b at the starting point

c after the equivalence point

Q7. What is the name of the solution added slowly from a burette to a solution?

a buffer

b titrant

c analyte

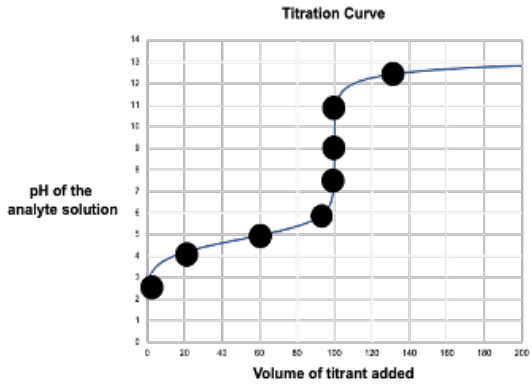
Q8. Select the correct order of events during a titration experiment set-up.

a Add indicator to the analyte, fill the burette with titrant, perform titration, record the volume used from the burette to calculate the unknown concentration.

b Add indicator to titrant, fill the burette with analyte, perform titration, record the volume used from the burette to calculate the unknown concentration.

c Add indicator to titrant, fill the burette with titrant, perform titration, record the volume used from the burette to calculate the unknown concentration.

Q9. According to the graph, at what pH does the concentration of weak acid equal that of its conjugated base?



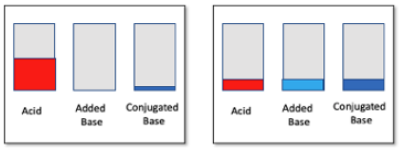
a pH=9

b pH=7.5

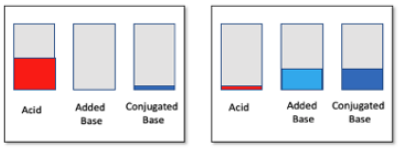
c pH=5

Q10. Which of the following pairs represents the transformation titration of a weak acid with a strong base at the equivalence point?

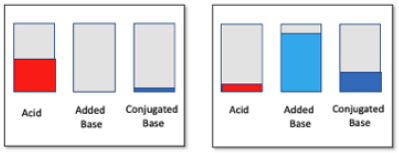
a.



b.



c.



Q11. In which container would you dispose the sodium hydroxide?

a Black container for alkaline inorganic waste

b White container acid organic liquid waste

c In the sink

B.3 Demographic Questions

Q12. How old are you?

Q13. What is your gender?

a Woman

b Man

c Prefer not to say

d Other

Q14. What is your current field of study?

Q15. Which previous chemistry courses did you follow? Ex. chemistry in high school, natural sciences in high school, Algemene Technische Scheikunde in university, etc.

B.4 Statistical Analysis

Data analysis was conducted using Python library version 3.7.6 for descriptive statistics and the package Pingouin [136] and JASP 0.17.1 [137].

To analyze if the differences between pretest and posttest scores were statistically significant, the following hypotheses were formulated:

- H_0 (null hypothesis): There is no significant difference between the mean of the pretest and the posttest.

- H_1 (alternative hypothesis): The mean of the posttest scores is significantly higher than that of the pretest.

At an alpha risk of rejection of 0.05, the one-tailed t-test showed a p-value of 0.056, which was not statistically significant. This implies that the null hypothesis could not be rejected, and it can be concluded that there is no considerable difference between the mean of the pretest and posttest scores.

The t-test was conducted as the Shapiro-Wilk test indicated that the data were normally distributed.

Table B.1: Paired Samples T-Test

								95% CI for Cohen's d		
Measure 1		Measure 2	t	df	p	Cohen's d	SE Cohen's d	Lower	Upper	CLES
Pretest	-	Posttest	−1.789	8	0.056	−0.596	0.176	-∞	0.019	0.599

Note. For all tests, the alternative hypothesis specifies that Pretest is less than Posttest.
Note. For the Student's t-test, effect size is given by Cohen's *d*.
Note. CLES was computed using the function *pg.compute_effsize* [136]

Table B.2: Test of Normality (Shapiro-Wilk)

			W	p
Pretest	-	Posttest	0.863	0.102

Note. Significant results suggest a deviation from normality.

Table B.3: Descriptives

	N	Mean	SD	SE	Coefficient of variation
Pretest	9	7.889	2.369	0.790	0.300
Posttest	9	8.556	2.007	0.669	0.235

Appendix C

Appendix of Chapter 5

C.1 Titration Code

In the first iteration of the prototype, the colour change in the titration was programmed with hard-coded calculated data of a titration of acetic acid with sodium hydroxide. However, this hampered the scalability of the levels. Therefore, a titration calculator using the following equations was created in the second iteration. The code can be found on the website [1].

The titration curve of each experiment is shown to the user in the logbook. This graph plots the pH of the analyte solution (beaker in the stirrer) as a function of the volume of added titrant (solution added from the burette). In the application, the user can monitor the reaction's progress using the pH meter values or observe the indicator's colour.

Four distinct calculations are used to calculate the pH of the solution [185], depending on the volume added from the burette, and these are:

1. The starting pH of a solution that contains only the acid or base
2. The pH values between the start and the equivalence point
3. The pH at the equivalence point
4. The pH values past the equivalence points

[1] https://alfajess92.github.io/marlab/

Strong Acid-Strong Base Titrations

In Level 1, the user performs a strong acid-strong base titration. This reaction is also called a neutralization reaction because the acid and the base react to form a neutral solution with an ionic compound or salt and water. Figure C.1 illustrates an example of a titration curve obtained from titrating 50 mL of a 0.2 M solution of a strong acid such as hydrochloric acid (HCl) with a strong base (i.e., NaOH) as the titrant. Specifically, in Level 1, a sample of HCl is titrated with a solution of NaOH. In this reaction, the acid (HCl) is neutralized by the base (NaOH), and as a result, water and NaCl are formed (C.1).

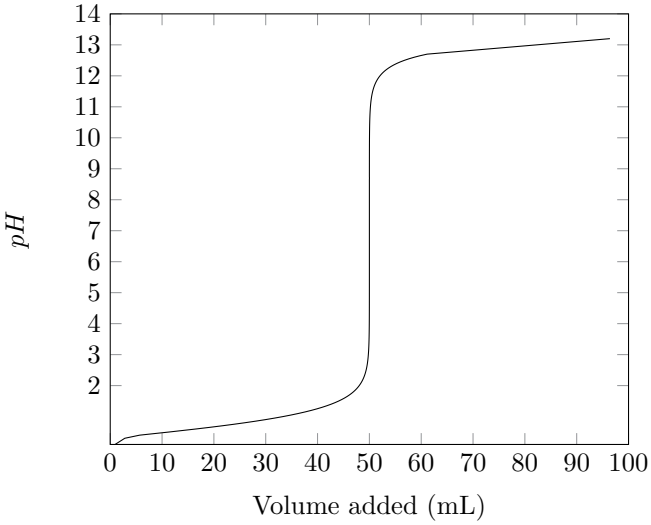


Figure C.1: Example of a titration curve obtained after titrating a sample of a strong acid with a strong base

$$\mathrm{NaOH}_{(aq)} + \mathrm{HCl}_{(aq)} \longrightarrow \mathrm{NaCl}_{(aq)} + \mathrm{H_2O}_{(l)} \tag{C.1.1}$$

Because $HCl_{(aq)}$, $NaOH_{(aq)}$ and $NaCl_{(aq)}$ are all completely dissociated, the net ionic equation for the neutralization reaction is [111]:

$$\mathrm{OH^-}_{(aq)} + \mathrm{H_3O^+}_{(aq)} \rightleftharpoons + 2\,\mathrm{H_2O}_{(l)} \tag{C.1.2}$$

At the start of the titration

In this stage of titration, the only species that contribute to the pH value is the analyte (HCl). Therefore, the initial value of the pH is calculated using the equation C.1.3

The pH of a solution is defined as the negative base-10 logarithm of the molar concentration of hydronium ions $\log[H_3O^+]$ [113].

$$pH = -\log\left[H_3O^+\right] \tag{C.1.3}$$

Where $[H_3O^+]$ denotes the concentration in mol/L or M of the hydronium ions.

Before the equivalence point

After some volume of titrant has been added to the analyte, stoichiometric calculations are needed to determine the unreacted substances that influence the pH.

Consider the charge-balanced equation:

$$[Na^+] + [H_3O^+] = [Cl^-] + [OH^-] \tag{C.1.4}$$

The following expression is obtained if one expresses the chemical species in solution using the original sample concentration of the acid C_a (expressed in mol/L or M) and the volume of the sample V_a in L. The concentration of the titrant used C_b (expressed in mol/L or M) and V_b its volume in L.

$$[Na^+] = \frac{C_b V_b}{V_a + V_b} \tag{C.1.5}$$

$$[Cl^-] = \frac{C_a V_a}{V_a + V_b} \tag{C.1.6}$$

In this region and for each added volume, the dilution factors $V_a/V_a + V_b$ and $V_b/V_a + V_b$ consider the mutual dilution of the sample and titrant. Substitution of the equations C.1.5 and C.1.6 into C.1.4 yields equation C.1.7.

$$\left[H_3O^+\right] = \frac{C_a V_a - C_b V_b}{V_a + V_b} - \left[OH^-\right] \tag{C.1.7}$$

At this point, the amount of strong base added to the solution has completely reacted with the acid to form water and a salt (neutralization reaction), resulting

in a primarily acidic solution with a negligible amount of OH^-. Therefore, the equation simplifies to:

$$[H_3O^+] = \frac{C_a V_a - C_b V_b}{V_a + V_b} \tag{C.1.8}$$

The pH of the solution in this point of the titration is calculated by substituting the value obtained in equation C.1.8 into C.1.3.

At the equivalence point

The equivalence point of the reaction is the point at which equal quantities of each reactant have been mixed, i.e. when the number of moles of the acid equals the number of moles of added base. At the equivalence point in a strong acid-strong base titration, the pH is 7.0 [113]. This is because neither the cation of a strong base (in this case, $[Na^+]$) nor the anion of a strong acid (in this case, $[Cl^-]$) can hydrolyze, so their effect on the pH is negligible. Rather, the pH value here is solely determined by the autoprotolysis of water.

After the equivalence point After the equivalence point of the titration, an excess of base is added to the solution. Substitution of the equations C.1.5 and C.1.6 into C.1.4 yields equation C.1.9.

$$[OH^-] = \frac{C_b V_b - C_a V_a}{V_a + V_b} - [H_3O^+] \tag{C.1.9}$$

Note that at this point of the reaction, one can assume that there is an excess of base, $[OH^-] >> [H_3O^+]$, therefore the equation can be simplified as follows:

$$[OH^-] = \frac{C_b V_b - C_a V_a}{V_a + V_b} \tag{C.1.10}$$

Similar to the equation C.1.3, the concentration of the hydroxide ions $[OH^-]$ can be used to calculate the pOH (C.1.11).

$$pOH = -\log [OH^-] \tag{C.1.11}$$

In order to use an expression in terms of concentrations $[OH^-]$ and $[H_3O^+]$, one can use the equilibrium equation for the autoionization of water, which symbol is K_w and its equation is:

$$K_w = [OH^-] \cdot [H_3O^+] \tag{C.1.12}$$

The ion product constant of water can then be used to calculate the pH of a solution. The pH can be calculated through its logarithmic form, which is referred to as pK_w (equation C.1.13). In pure water, the concentrations of $[OH^-]$ and $[H_3O^+]$ are equal, and it has been found that the concentrations values at $25°C$ are equal to $1.0x10^{-7}M$ for each ion. Thus, the value of $K_w = 1.0x10^{-14}$, which corresponds to a pK_w of 14, which can be substituted into equation C.1.13 to obtain the pH of the solution.

$$pH = pK_w - pOH \tag{C.1.13}$$

Weak Acid - Strong Base Titration

In Levels 2 and 3, the user needs to titrate a weak acid (acetic acid CH_3COOH) with a strong base (sodium hydroxide $NaOH$).

Figure 4.4 illustrates an example of a titration curve obtained from titrating 10 mL of a 0.5 M acetic acid solution with 0.5 M sodium hydroxide as the titrant.

Because weak acids are partially dissociated in an aqueous solution, the neutralization reaction involves the proton transfer from the weak acid to the strong base. Acetic acid reacts with aqueous $NaOH$ to produce water and sodium acetate (CH_3COONa) (C.1.14).

$$\mathrm{CH_3COOH_{(aq)} + OH^-{}_{(aq)} \rightleftharpoons CH_3COO^-_{(aq)} + H_2O_{(l)}} \tag{C.1.14}$$

The Na^+ ions do not appear in the net ionic equation because both the base NaOH and the salt CH_3COONa are entirely dissociated.

For weak acid-strong base titrations, the pH calculations at each reaction point involve the equilibrium reaction of the weak acid and its conjugated base. The acetic acid ionizes in solution according to the general equation C.1.15.

$$\mathrm{CH_3COOH + H_2O \rightleftharpoons H_3O^+ + CH_3COO^-} \tag{C.1.15}$$

The equilibrium or dissociation constant or K_a is defined as:

$$K_a = \frac{[H_3O^+][CH_3COO^-]}{[CH_3COOH]} \tag{C.1.16}$$

At the start of the titration

At the start of the titration, the solution is composed only of the weak acid (CH_3COOH). Therefore the pH of the solution is calculated using the equation C.1.16.

Let x be the amount of acetic acid that dissociates per litre. As a result, the concentration of CH_3COOH will decrease by x and both the concentration of CH_3COO^- and H_3O^+ will increase by x. When these values are substituted in C.1.16, the following expression is obtained:

$$K_a = \frac{(x)(x)}{[CH_3COOH] - x} = \frac{x^2}{[CH_3COOH] - x} \tag{C.1.17}$$

The term x can be calculated using a quadratic formula; however, because the dissociation constant K_a of the acetic acid has a minimal value, namely 1.76×10^-5, an approximation is made [113]:

$$[CH_3COOH] - x \approx [CH_3COOH] \tag{C.1.18}$$

Solving for x gives:

$$x = \sqrt{K_a \times [CH_3COOH]} \tag{C.1.19}$$

In the equation C.1.17, the term x also represents the concentration of H_3O^+; therefore, the solution of the term in equation C.1.19 can be used in equation C.1.3 to calculate the pH at the start of the titration.

Before the equivalence point

After adding the titrant, the pH of the solution is calculated using stoichiometric equations. The acetic CH_3COOH is neutralized in this step, and its conjugated base CH_3COO^- is formed. Thus in the solution, a mixture of both species is present.

Consider the mass balance C.1.20:

$$[CH_3COOH] + [CH_3COO^-] = \frac{C_a V_a}{V_a + V_b} \tag{C.1.20}$$

Considering the sum of the concentrations of the different species in the mixture as $C = [CH_3COOH] + [CH_3COO^-]$ and the concentration fraction of each specie as:

$$\alpha_{HA} = \frac{[CH_3COOH]}{[CH_3COOH] + [CH_3COO^-]} \tag{C.1.21}$$

$$\alpha_{A^-} = \frac{[CH_3COO^-]}{[CH_3COOH] + [CH_3COO^-]} \tag{C.1.22}$$

By substituting equations C.1.22 and C.1.21 into equation C.1.20, the following equations expressed in concentration and volumes are obtained:

$$[CH_3COOH] = \frac{C_a V_a \alpha_{A^-}}{V_a + V_b} \tag{C.1.23}$$

$$[CH_3COO^-] = \frac{C_a V_a \alpha_{HA}}{V_a + V_b} \tag{C.1.24}$$

The titration of a weak acid using the concentration and volumes can be expressed as follows:

$$(C_b + [H_3O^+] - [OH^-])V_b = (C_a \alpha_{HA} - [H_3O^+] + [OH^-])Va \tag{C.1.25}$$

Equation C.1.26 denotes a dimensionless parameter that represents the degree of completion of the titration. This parameter is used in the code to define the regions of the titration.

$$\varphi_{ab} = \frac{C_b V_b}{C_a V_a} \tag{C.1.26}$$

By substituting the equation C.1.25 with equation C.1.26 the following explicit expression is obtained:

$$\varphi_{ab} = \frac{\alpha_{A^-} - \frac{[H_3O^+] - [OH^-]}{C_a}}{1 + \frac{[H_3O^+] - [OH^-]}{C_b}} \tag{C.1.27}$$

Before the equivalence point, one can neglect the following terms:

$$\frac{[H_3O^+] - [OH^-]}{C_a} \quad \& \quad \frac{[H_3O^+] - [OH^-]}{C_b} \tag{C.1.28}$$

As a result, the explicit equation is reduce to $\varphi_{ab} \approx \alpha_{A^-}$ thus the concentration of hydronium ions $[H_3O^+]$ can be calculated using the following expression:

$$[H_3O^+] = \frac{K_a(1 - \varphi_{ab})}{\varphi_{ab}} \tag{C.1.29}$$

The pH of the solution in this region is calculated using equations C.1.29 and C.1.3.

At the equivalence point

At the equivalence point of the reaction between the weak acid CH_3COOH and the titrant $NaOH$, the ion CH_3COO^- is produced. This ion reacts with water to form its conjugated acid CH_3COOH and the hydroxide ion OH^-. As this reaction increases the concentration of OH^- ions in the solution, the pH rises higher than the neutral value of 7.0.

$$\mathrm{CH_3COO^- + H_2O \rightleftharpoons CH_3COOH + OH^-} \tag{C.1.30}$$

The dissociation constant or K_b is defined as:

$$K_b = \frac{[OH^-][CH_3COOH]}{[CH3COO^-]} \tag{C.1.31}$$

Similarly to the start of the titration (C.1.17), the concentration of $[OH^-]$ can be calculated using a quadratic formula or assuming that the conjugated base CH_3COO^- ionizes poorly. Therefore, one can calculate the pH of the reaction in this region using the equations C.1.32, C.1.11 and C.1.13

$$[OH^-] = \sqrt{K_b \times \left[\frac{C_a V_a}{V_a + V_b}\right]} \tag{C.1.32}$$

After the equivalence point

In this region, the concentration of $[OH^-]$ from the reaction C.1.31 is negligible compared to the $[OH^-]$ from the excess NaOH added from the burette. Therefore, the pH is determined by the concentration of $[OH^-]$ from the excess of NaOH using equations C.1.9 and C.1.11 and C.1.13.

Weak Base - Strong Acid Titration

The equations to calculate the regions of a weak base-strong acid titration were defined using a similar procedure as the previous section. This type of titration is only given as an option to the user in the extra level called "Sandbox".

C.2 Questionnaires from evaluations

C.2.1 ACTA Test

The pilot test evaluated the participant's comprehension of the theoretical principles and processes associated with the titration experiment. The chemistry test was administered in English to assess their understanding of the subject. The test was administered as pre and posttest. In the posttest version, the SUS questionnaire [117] and additional demographic questions were asked.

1. Chemistry Test

Q1. A titration experiment can be used to determine the _____ of acid using a known concentration of base.

a volume

b unknown concentration

c known concentration

Q2. Strong base such as sodium hydroxide (NaOH)_____ when dissolved in water.

a fully dissociates

b does not dissociate

c partially dissociates

Q3. The molecules of a weak acid such as acetic acid _____ dissociate in water.

a never

b fully

c partially

Q4. At the equivalence point of a titration of a strong acid with a strong base, the pH of a solution is _____.

a neutral

b slightly acidic

c slightly basic

Q5. At the equivalence point of a titration of a weak acid with a strong base, the pH of the solution is _____.

a neutral

b slightly acidic

c slightly basic

Q6. At the starting point of a titration experiment, the pH of the solution _____ as the strong base is added to the solution of the weak acid.

a does not change

b quickly increases

c quickly decreases

Q7. At what point will colour change occur when a suitable indicator is added to a titration?

a near the equivalence point

b at the starting point

c after the equivalence point

Q8. What is the name of the solution added slowly from a burette to a solution?

a buffer

b titrant

c analyte

Q9. What is the name of the solution whose concentration has to be determined?

a buffer

b titrant

c analyte

Q10. Select the correct order of events during a titration experiment set-up.

a Add indicator to analyte, fill the burette with titrant, perform titration, record the volume used from the burette to calculate the unknown concentration.

b Add indicator to titrant, fill the burette with analyte, perform titration, record the volume used from the burette to calculate the unknown concentration.

c Add indicator to titrant, fill the burette with titrant, perform titration, record the volume used from the burette to calculate the unknown concentration.

Q11. Where would you dispose a solution of sodium hydroxide?

a Black container for alkaline inorganic waste

b White container for acid inorganic waste

c In the sink

2. System Usability Scale (SUS)

Please select the answer that best expresses how you feel about each statement after using the application today.

Scale: 1 - Strongly Disagree, 2 - Somewhat Disagree, 3 - Neutral, 4 - Somewhat Agree, 5 - Strongly Agree

Q12. I think I would like to use this tool frequently.

Q13. I found the tool unnecessarily complex.

Q14. I thought the tool was easy to use.

Q15. I think that I would need the support of a technical person to be able to use this tool.

Q16. I found the various functions in this tool were well integrated.

Q17. I thought there was too much inconsistency in this tool.

Q18. I would imagine that most people would learn to use this tool very quickly.

Q19. I found the tool very cumbersome to use.

Q20. I felt very confident using the tool.

Q21. I needed to learn a lot of things before I could get going with this tool.

3. Feedback Questions

Q22. Did you experience any other problems while using the application?

Q23. Have you performed a titration in a real laboratory?

a Yes

b No

Q24. Have you tried any other Augmented Reality applications? If yes, can you name it?

a Yes

b No

c Other:

4. Demographic Questions

Q25. How old are you?

a Under 18

b 18-24

c 25-34

d 35-44

e 45-54

Q26. What is your gender?

a Woman

b Man

c Prefer not to say

C.2.2 Merck Test

After the pilot test, the questionnaire was modified and translated into German. The chemistry test was administered as a pre and posttest to assess participants' understanding of the subject. In the posttest version, the German version of SUS questionnaire [186] and additional demographic questions were asked.

1. Chemistry Test- German

Q1. Mit einem Titrationsexperiment kann _____ der Säure mit einer bekannten Konzentration der Base bestimmt werden..

a die bekannte Konzentration

b die unbekannte Konzentration

c das Volumen

Q2. Starke Basen wie Natriumhydroxid (NaOH)_____ wenn sie in Wasser gelöst werden.

a dissoziieren nicht

b dissoziieren vollständig

c dissoziieren teilweise

Q3. Eine schwache Säure wie Essigsäure (CH_3COOH) _____ wenn sie in Wasser gelöst werden.

a dissoziieren vollständig

b dissoziiert niemals

c dissoziiert teilweise

Q4. Am Äquivalenzpunkt einer Titration von einer starken Säure mit einer starken Base ist der pH-Wert der Lösung _____.

a leicht sauer

b neutral

c leicht basisch

Q5. Am Äquivalenzpunkt einer Titration von einer schwachen Säure mit einer starken Base ist der pH-Wert der Lösung _____.

a leicht sauer

b neutral

c leicht basisch

Q6. Zu Beginn eines Titrationsexperimentes _____ der pH-Wert der Lösung wenn die starke Base zur Lösung der schwachen Säure hinzugefügt wird.

a sinkt

b ändert sich nicht

c ändert sich nicht

Q7. An welchem Punkt der Titration ändert ein Indikator seine Farbe?

a in der Nähe des Äquivalenzpunktes

b am Startpunkt

c nach dem Äquivalenzpunkt

Q8. Welche der folgenden Aussagen ist richtig?

a Der Phenolphthalein-Indikator wird durch den pH-Wert der Lösung nicht beeinflusst.

b Die Ablesung der Bürettenskala sollte am unteren Rand des Meniskus erfolgen.

c Der Titrant ist die Lösung mit unbekannter Konzentration

Q9. Wie nennt man die Lösung, deren Konzentration bestimmt werden muss?

a Puffer

b Titrant

c Analyt

Q10. Wählen Sie die richtige Reihenfolge der Ereignisse beim Aufbau eines Titrationsexperiments.

a Indikator zum Analyten hinzufügen, die Bürette mit Titriermittel füllen, die Titration durchführen und das aus der Bürette verwendete Volumen notieren, um die unbekannte Konzentration zu berechnen.

b Indikator zum Titriermittel hinzufügen, die Bürette mit dem Analyten füllen, die Titration durchführen, das Volumen der Bürette zur Berechnung der unbekannten Konzentration notieren.

c Indikator zum Titriermittel hinzufügen, die Bürette mit Titriermittel füllen, die Titration durchführen, das von der Bürette verwendete Volumen zur Berechnung der unbekannten Konzentration notieren.

Q11. In welchen Behälter würden Sie die Natriumhydroxidlösung (NaOH) entsorgen?

a In den Behälter für anorganische alkalische Abfälle

b In den Behälter für anorganische saure Abfälle

c In der Spüle

2. System Usability Scale (SUS)- German

Dies ist ein Standardfragebogen, der die allgemeine Benutzerfreundlichkeit eines Systems misst. Bitte wählen Sie die Antwort aus, die am besten ausdrückt, wie Sie die einzelnen Aussagen nach der heutigen Nutzung der Anwendung empfinden.

Skala: 1 - Stimme überhaupt nicht zu, 2 - Eher nicht einverstanden, 3 - Neutral, 4 - Stimmt eher zu, 5 - Stimme voll und ganz zu

Q12. Ich denke, dass ich dieses Tool häufig verwenden möchte.

Q13. Ich fand das Tool unnötig komplex.

Q14. Ich dachte, das Tool war einfach zu bedienen.

Q15. Ich denke, dass ich die Unterstützung einer technischen Person brauche, um dieses Tool nutzen zu könen.

Q16. Ich fand, die verschiedenen Funktionen in diesem Tool waren gut integriert.

Q17. Ich dauchte, dass dieses Tool nicht konsistent genug war.

Q18. Ich würde mir vorstellen, dass die meisten Leute sehr schnell lernen würden, dieses Tool zu benutzen

Q19. Ich fand dieses Tool sehr umstämdlich zu benutzen.

Q20. Ich habe mich sehr selbstsicher gefühlt, dieses Tool zu verwerden.

Q21. Ich musste eine Menge Dinge lernen, bevor ich mit diesem Tool loslegen konnte.

3. Feedback Questions- German

Q22. Sind Sie bei der Verwendung der Anwendung auf ein Problem gestossen?

Q23. Haben Sie schon einmal eine Titration in einem echten Labor durchgeführt?

a Ja

b Nein

Q24. Haben Sie weitere Empfehlungen oder Kommentare?

Q25. Haben Sie eine andere Augmented Reality-Anwendung ausprobiert? Wenn ja, können Sie sie nennen?

a Ja

b Nein

c :

4. Demographic Questions- German

Q26. Wie alt sind Sie?

a Unter 18 Jahren

b 18-24

c 25-34

d 35-44

e 45-54

Q27. Welches ist Ihr Geschlecht?

a Frau

b Männer

c Lieber nicht sagen

C.2.3 KU Leuven Test

The questionnaires were modified and translated into Dutch from the previous version of the pilot test (ACTA). The chemistry test was administered as a pre- and posttest to assess participants' understanding of the subject. Additionally, a Dutch version of the SUS questionnaire [187] and additional demographic questions were included in the post-test.

1. Chemistry Test - Dutch

Q1. Een titratieëxperiment kan gebruikt worden om _____ van een zuur te bepalen door een gekende concentratie van een base te gebruiken.

a de gekende concentratie

b de onbekende concentratie

c het volume

Q2. Sterke basen zoals natriumhydroxide (NaOH) _____ na het oplossen in water.

a dissociëren niet

b dissociëren volledig

c dissociëren deels

Q3. Een zwak zuur zoals azijnzuur (CH_3COOH) _____ na het oplossen in water.

a dissocieert volledig

b dissocieert niet

c dissocieert deels

Q4. Op het equivalentiepunt van een titratie van een sterk zuur met sterke base, is de pH van de oplossing _____.

a licht zuur

b neutraal

c licht basisch

Q5. Op het equivalentiepunt van een titratieëxperiment van een zwak zuur met sterke base, is de pH van de oplossing _____.

a licht zuur

b neutraal

c licht basisch

Q6. Op het startpunt van een titratieëxperiment, _____ de pH van de oplossing _____ wanneer de sterke base wordt toegevoegd aan de oplossing van het zwakke zuur.

a daalt

b blijft constant

c stijgt

Q7. Op welke punt in de titratie verandert de indicator van kleur?

a rond het equivalentiepunt

b op het startpunt

c na het equivalentiepunt

Q8. Welk van de volgende stellingen is juist?

a De fenolftaleïne indicator is niet beïnvloed door de pH van de oplossing.

b Het aflezen van de buret moet gedaan worden vanaf de onderkant van de meniscus.

c Het titrant is de oplossing van de ongekende concentratie.

Q9. Wat is de naam van de oplossing waarvan de concentratie bepaald dient te worden?

a buffer

b titrant

c analiet

Q10. Kies de juiste volgorde van evenementen tijdens een titratieëxperiment opstelling.

a Voeg indicator toe aan het analiet, vul de buret met titrant, voer de titratie uit, noteer het gebruikte volume van de buret om de ongekende concentratie te berekenen.

b Voeg indicator toe aan het titrant, vul de buret met het analiet, voer de titratie uit, noteer het gebruikte volume van de buret om de ongekende concentratie te berekenen.

c Voeg indicator toe aan het titrant, vul de buret met het titrant, voer de titratie uit, noteer het gebruikte volume van de buret om de ongekende concentratie te berekenen.

Q11. In welke afvalcontainer zou je natriumhydroxide (NaOH) weggooien?

a Container voor basisch inorganisch afval

b Container voor zuur inorganisch afval

c In de gootsteen

2. System Usability Scale (SUS) - Dutch

Kies het antwoord dat het beste weergeeft wat u van elke uitspraak vindt na het gebruik van de applicatie vandaag.

Schaal: 1 - Sterk mee oneens, 2 - Enigszins oneens, 3 - Neutraal, 4 - Enigszins mee eens, 5 - Sterk mee eens

Q12. Ik denk dat ik dit systeem graag regelmatig wil gebruiken.

Q13. Ik vond het systeem onnodig complex.

Q14. Ik vond het systeem makkelijk te gebruiken.

Q15. Ik denk dat ik ondersteuning nodig heb van een technisch persoon om dit systeem te kunnen gebruiken.

Q16. Ik vond dat de verschillende functies in dit systeem erg goed geintegreerd zijn.

Q17. Ik vond dat er teveel tegenstrijdigheden in het systeem zaten.

Q18. Ik kan me voorstellen dat de meeste mensen zeer snel leren om dit systeem te gebruiken.

Q19. Ik vond het systeem erg omslachtig in gebruik.

Q20. Ik voelde me erg vertrouwd met het systeem.

Q21. Ik moest erg veel leren voordat ik aan de gang kon gaan met dit systeem.

3. Feedback Questions - Dutch

Q22. Heeft u andere problemen ondervonden tijdens het gebruik van de applicatie?

Q23. Heb je een titratie uitgevoerd in een echt laboratorium?

a Ja

b Nee

Q24. Heeft u nog andere aanbevelingen of opmerkingen?

Q25. Heb je een andere Augmented Reality toepassing geprobeerd? Zo ja, kunt u die noemen?

a Ja

b Neen

c :

4. Demographic Questions - Dutch

Q25. Hoe oud ben je?

Q26. Wat is je geslacht?

a Vrouw

b Vrouw

c Zeg ik liever niet

Q27. Wat is je huidige opleiding?

Q28. Welke vorige chemische vakken heb je gevolgd? (Bv. chemie in het middelbaar, natuurkunde in het middelbaar, Algemene Technische Scheikunde op de universiteit)

C.3 Interview Questions

Instructions for researchers interviewing participants:

- Explain to the user about the research
- Provide information that we will take notes on, and they will be used in our research
- Use the list of questions below as a guide during the interview
- The interview will be divided into questions concerning details of the app, experience with AR and future use.

C.3.1 Details of the app

- Could you describe how your experience with the application was?
- What do you think about the application?
- Do you think the virtual experiment is close to or similar to what someone will do in a real laboratory?
- How many levels did you complete?
- Do you think you learned about how to perform titration?

- Did the application help you to learn something you did not know before? (if yes, what did you learn?)
- How was your experience with the audio?
- What do you think about the hints given by Marie Curie? Did you use them (when you were stuck)?
- What do you think about the duration of each experiment?
- Did you like the design/layout?
- Were all the buttons easy to understand?
- What do you think about the quiz at the end of each level?
- Is there anything you would change?
- Was anything surprising, or did it not work as expected?
- Which was the easiest task to accomplish?
- What was the hardest task to accomplish?
- Can you mention the three aspects you liked the most?
- Can you mention the three aspects you disliked or think can be improved?
- Can you offer some advice for improving this tool? (Suggestions)

C.3.2 Experience with AR and future use

- Have you used augmented reality before?
- Suppose you can have access to the application. Would you use it?
- Do you think using AR technology to prepare you for real laboratory experiments in your training can be helpful?
- Do you have any suggestions for other laboratory experiments you would like to have an AR version of?

C.4 Themes and Codes

The results of the interviews and open-ended questions from the evaluations were analyzed and coded using inductive thematic coding. The issues users experienced with MAR Lab were identified, organized into broad themes, and used to generate codes. These codes, along with users' comments regarding design elements, user experience, positive traits, and suggestions for improvement, are presented in the following table C.1.

C.5 Statistical Analysis

This section presents the statistical analysis results from the three evaluations. The analysis was performed using Python library version 3.7.6 for descriptive statistics, and the Pingouin package [136], and JASP 0.17.1 software [137].

Table C.1: Thematic Analysis Codes

Code	Subcode	Description
General UX		Comments related to the subjective general user experience
	Usefulness	Comments or scenarios where users find the application useful
	Unclarities	Unclarities about how to use the application or the content
	Things liked	Features that users liked
	Things disliked	Features that users disliked
	Praise	Positive comments about the application and experience
Suggestions		Suggestions or improvements suggested by users
Problems		Problems experienced by users
	UI	Problems related to the user interface
	Translation	Translation problems
	Tracking	Problems experienced by users while tracking the real environment
	Selection	Problems experienced by users while selecting virtual objects
	Overheat	Overheat of the devices while using the application
	Materials disappear	Materials disappearing of objects during the experience
	Floating objects	Problems related to floating objects in the environment
	Crash	Crash or unwanted closure of the app during the experience
Learning Experience		Comments related to the user learning experience
	Hardest task	Hardest tasks according to users
Design Elements		Comments about the application design elements
	Language	Comments about the language and translation of the app
	Hints	Comments about the hints provided by the system
	Duration	Comments about the duration of the experience
	Quiz	Comments about the implementation of the quiz
	Audio	Comments about the voice-over and sound when interacting with objects
AR experience		Comments about users' experience with AR in general
Other comments		Comments that cannot be classified into any other codes

Table C.2: ACTA operators: Wilcoxon signed-rank test

Measure 1		Measure 2	W	z	df	p	RBC	SE RBC	95% CI for RBC		CLES
									Lower	Upper	
Pretest	-	Post-test	0.000	−3.180		< .001	−1.000	0.305	-∞	−1.000	0.711

Note. For all tests, the alternative hypothesis specifies that Pretest is less than Post-test.
Note. For the Wilcoxon test, the effect size is given by the matched rank biserial correlation (RBC)
Note. CLES was computed using the function *pg.compute_effsize* [136]

Table C.3: ACTA operators: Test of Normality (Shapiro-Wilk)

			W	p
Pretest	-	Post-test	0.869	0.017

Note. Significant results suggest a deviation from normality.

Table C.4: ACTA operators: Descriptives

	N	Mean	SD	SE	Coefficient of variation
Pretest	18	5.944	1.434	0.338	0.241
Post-test	18	7.389	1.852	0.436	0.251

Table C.5: ACTA students: Paired Samples T-Test

Measure 1		Measure 2	t	df	p	Cohen's d	SE Cohen's d	95% CI for Cohen's d		CLES
								Lower	Upper	
Pretest	-	Post-test	1.313	7	0.885	0.464	0.352	-∞	1.064	0.375

Note. For all tests, the alternative hypothesis specifies that Pretest is less than Post-test.
Note. For the Student's t-test, the effect size is given by Cohen's *d*.
Note. CLES was computed using the function *pg.compute_effsize* [136]

Table C.6: ACTA students: Test of Normality (Shapiro-Wilk)

			W	p
Pretest	-	Post-test	0.866	0.137

Note. Significant results suggest a deviation from normality.

Table C.7: ACTA students: Descriptives

	N	Mean	SD	SE	Coefficient of variation
Pretest	8	8.000	2.204	0.779	0.275
Post-test	8	7.125	1.642	0.581	0.230

Table C.8: Merck test: Wilcoxon signed-rank test

Measure 1		Measure 2	W	z	df	p	RBC	SE RBC	95% CI for RBC		CLES
									Lower	Upper	
Pretest	-	Post-test	21.000	1.183		0.912	0.500	0.399	-∞	0.847	0.316

Note. For all tests, the alternative hypothesis specifies that Pretest is less than Post-test.
Note. For the Wilcoxon test, the effect size is given by the matched rank biserial correlation (RBC)
Note. CLES was computed using the function *pg.compute_effsize* [136]

Table C.9: Merck test: Test of Normality (Shapiro-Wilk)

			W	p
Pretest	-	Post-test	0.794	0.036

Note. Significant results suggest a deviation from normality.

Table C.10: Merck test: Descriptives

	N	Mean	SD	SE	Coefficient of variation
Pretest	7	8.857	1.464	0.553	0.165
Post-test	7	8.286	0.951	0.360	0.115

Table C.11: KU Leuven test: Wilcoxon signed-rank test

Measure 1		Measure 2	W	z	df	p	RBC	SE RBC	95% CI for RBC		CLES
									Lower	Upper	
Pretest	-	Post-test	36.000	−1.917		0.026	−0.529	0.269	-∞	−0.134	0.634

Note. For all tests, the alternative hypothesis specifies that Pretest is less than Post-test.
Note. For the Wilcoxon test, the effect size is given by the matched rank biserial correlation (RBC)
Note. CLES was computed using the function *pg.compute_effsize* [136]

Table C.12: KU Leuven test: Test of Normality (Shapiro-Wilk)

			W	p
Pretest	-	Post-test	0.921	0.032

Note. Significant results suggest a deviation from normality.

Table C.13: KU Leuven test: Descriptives

	N	Mean	SD	SE	Coefficient of variation
Pretest	29	8.483	0.986	0.183	0.116
Post-test	29	8.966	0.944	0.175	0.105

Appendix D

Appendix of Chapter 6

D.1 Final Questionnaire

In the final evaluation, the Dutch version of the tests was used, which had previously been validated (See Appendix C.2.3). The intervention included a pretest consisting only of chemistry questions, followed by a posttest, including the R-TLX instrument [188], the SUS questionnaire, the chemistry test and demographic questions. Only the Dutch version of the R-TLX is presented here.

D.1.1 R-TLX

Kies het punt op de schaal dat het beste uw ervaring met de app weergeeft.

Schaal: 0 (Heel weinig) – 100 (Heel veel)

Q1. Hoeveel moeite kostte de taak je mentaal (verstandelijk) gezien (bijvoorbeeld: denken, beslissen, herinneren, zoeken, enz.)?

Q2. In welke mate vormde de taak een lichamelijke belasting? (scrollen, klikken, fixeren, controleren, enz.)

Q3. Hoeveel tijdsdruk ervaarde je door het tempo waarin je de taak moest volbrengen?

Q4. Hoeveel inspanning moest je leveren om deze prestatie te bereiken?

Q5. Hoe succesvol heb je de opdracht denk je volbracht?

Q6. Hoe was je frustratieniveau tijdens het doen van de opdracht? Hoe onzeker, ontmoedigd, geïrriteerd en gestrest was je, ten opzichte van zeker, ontspannen en voldaan?

D.2 Statistical Analysis

This section presents the results of the statistical analysis of the final evaluation. The data was analyzed using Python library version 3.7.6 for descriptive statistics, the package Pingouin[136] and JASP 0.17.1 [137].

T-tests were used to assess statistical significance. To investigate whether the participants improved their performance as they progressed through the levels, two metrics were examined: the total time spent on each level and the number of errors made during the assembly of the titration equipment. The Friedman Test was used to determine whether there was a significant difference in the mean of total time and errors across the levels played by the same participants (i.e. paired samples). Conover's Test was applied as a pairwise posthoc test to investigate further the differences between the individual levels. The Bonferroni correction was used to reduce the chance of making a type I error when multiple comparisons were performed [168, 169].

Moreover, the Bayesian informative hypothesis testing approach [170] with a standard equation model (SEM) was adopted to analyze the data, with two competing hypotheses formulated based on expected relations between variables [171]. The analysis was conducted in R with the bain [173] and lavaan [174] packages.

Table D.1: Final Evaluation Learning Test: Paired Samples T-test

Measure 1		Measure 2	t	df	p	Cohen's d	SE Cohen's d	95% CI for Cohen's d		CLES
								Lower	Upper	
Pretest	-	Post-test	−2.276	14	0.020	−0.59	0.314	-∞	−0.116	0.684

Note. For all tests, the alternative hypothesis specifies that Pretest is less than Posttest.
Note. For the Student's t-test, effect size is given by Cohen's *d*.
Note. CLES was computed using the function *pg.compute_effsize* [136]

Table D.2: Final Evaluation Learning Test: Test of Normality (Shapiro-Wilk)

			W	p
Pretest	-	Post-test	0.949	0.511

Table D.3: Final Evaluation Learning Test: Descriptives

	N	Mean	SD	SE	Coefficient of variation
Pretest	15	7.600	2.063	0.533	0.271
Post-test	15	8.800	1.474	0.380	0.167

Table D.4: Total time per level: Descriptives

RM Factor 1	N	Mean	SD	SE	Coefficient of variation
Level 1	13	18.665	5.294	1.468	0.284
Level 2	13	10.541	3.466	0.961	0.329
Level 3	13	9.168	2.687	0.745	0.293

Table D.5: Total time per level: Friedman Test

Factor	Chi-Squared	df	p	Kendall's W
RM Factor 1	14.308	2	$< .001$	0.550

Table D.6: Total time per level: Conover's Post Hoc Comparisons - RM Factor 1

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
Level 1	Level 2	2.942	24	37.000	22.000	0.007	0.021	0.014
	Level 3	3.530	24	37.000	19.000	0.002	0.005	0.005
Level 2	Level 3	0.588	24	22.000	19.000	0.562	1.000	0.562

Table D.7: Assembly errors per level: Descriptives

RM Factor 1	N	Mean	SD	SE	Coefficient of variation
Level 1	13	26.385	22.100	6.130	0.838
Level 2	13	13.769	19.240	5.336	1.397
Level 3	13	5.538	4.824	1.338	0.871

Table D.8: Assembly errors per level: Friedman Test

Factor	Chi-Squared	df	p	Kendall's W
RM Factor 1	12.776	2	0.002	0.491

Table D.9: Assembly errors per level: Conover's Post Hoc Comparisons - RM Factor 1

		T-Stat	df	W_i	W_j	p	p_{bonf}	p_{holm}
Level 1	Level 2	2.228	24	35.500	24.500	0.036	0.107	0.071
	Level 3	3.545	24	35.500	18.000	0.002	0.005	0.005
Level 2	Level 3	1.317	24	24.500	18.000	0.200	0.601	0.200

Table D.10: Bayesian Informative Hypothesis Testing: Descriptive statistics

	n	M	SD
Cognitive load	15	30.11	10.27
Time-on-task	15	19.07	5.01
Pretest	15	7.60	2.06
Post-test	15	8.80	1.47

Note. n = sample size, M = mean, SD = standard deviation

Table D.11: Bayesian Informative Hypothesis Testing: Covariance matrix

	Estimate	SE	ES	95% CI
Knowledge Acquisition – Cognitive load	-5.02	5.22	-.26	-15.25, 5.20
Knowledge Acquisition – Time-on-task	-0.45	2.47	-.05	-5.30, 4.39
Cognitive load – Time-on-task	22.91	13.75	.48	-4.04, 49,86
Knowledge Acquisition	1.20	0.51	0.59	0.20, 2.20

Note. The mean instead of the estimate is reported for the variable knowledge acquisition.
SE = standard error, ES = effect size, which is expressed with Pearson correlation coefficients and Cohen's d for knowledge acquisition, 95%CI = confidence interval.

Bibliography

[1] R. A. Grier, "How high is high? A meta-analysis of NASA-TLX global workload scores," *Proceedings of the Human Factors and Ergonomics Society*, vol. 2015-Janua, no. October, pp. 1727–1731, 2015.

[2] J. J. Mintzes, J. H. Wandersee, and J. D. Novak, "Meaningful learning in science: The human constructivist perspective," *Handbook of Academic Learning*, pp. 405–447, 1997.

[3] European Commission, "Skills for industry: Curriculum Guidelines 4.0," Tech. Rep. January, European Commision, Brussels, 2020.

[4] J. C. Hayes and D. J. Kraemer, "Grounded understanding of abstract concepts: The case of STEM learning," *Cognitive Research: Principles and Implications*, vol. 2, no. 1, 2017.

[5] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, "Active learning increases student performance in science, engineering, and mathematics," *Proceedings of the National Academy of Sciences*, vol. 111, pp. 8410–8415, 6 2014.

[6] N. Yannier, S. E. Hudson, K. R. Koedinger, K. Hirsh-Pasek, R. M. Golinkoff, Y. Munakata, S. Doebel, D. L. Schwartz, L. Deslauriers, L. McCarty, K. Callaghan, E. J. Theobald, S. Freeman, K. M. Cooper, and S. E. Brownell, "Active learning: "Hands-on" meets "minds-on"," *Science*, vol. 374, pp. 26–30, 10 2021.

[7] H. Y. Agustian, L. T. Finne, J. T. Jørgensen, M. I. Pedersen, F. V. Christiansen, B. Gammelgaard, and J. A. Nielsen, "Learning outcomes of university chemistry teaching in laboratories: A systematic review of empirical literature," *Review of Education*, vol. 10, pp. 1–41, 8 2022.

[8] V. N. Lunetta, A. Hofstein, and M. P. Clough, "Learning and Teaching in the School Science Laboratory : An analysis of research , theory and

practice," in *Handbook of Research on Science Education* (S. Abell and N. Lederman, eds.), ch. 15, pp. 393–441, Lawrence Erlbaum Associates Publishers, 2007.

[9] B. Sanii, "Creating Augmented Reality USDZ Files to Visualize 3D Objects on Student Phones in the Classroom," *Journal of Chemical Education*, vol. 97, pp. 253–257, 1 2020.

[10] J. M. Argüello and R. E. Dempski, "Fast, Simple, Student Generated Augmented Reality Approach for Protein Visualization in the Classroom and Home Study," *Journal of Chemical Education*, vol. 97, pp. 2327–2331, 8 2020.

[11] S. Yang, B. Mei, and X. Yue, "Mobile Augmented Reality Assisted Chemical Education: Insights from Elements 4D," *Journal of Chemical Education*, vol. 95, pp. 1060–1062, 6 2018.

[12] J. K. Aw, K. C. Boellaard, T. K. Tan, J. Yap, Y. P. Loh, B. Colasson, Ã. Blanc, Y. Lam, and F. M. Fung, "Interacting with Three-Dimensional Molecular Structures Using an Augmented Reality Mobile App," *Journal of Chemical Education*, vol. 97, pp. 3877–3881, 10 2020.

[13] K. N. Plunkett, "A Simple and Practical Method for Incorporating Augmented Reality into the Classroom and Laboratory," *Journal of Chemical Education*, vol. 96, pp. 2628–2631, 11 2019.

[14] J. A. Naese, D. McAteer, K. D. Hughes, C. Kelbon, A. Mugweru, and J. P. Grinias, "Use of Augmented Reality in the Instruction of Analytical Instrumentation Design," *Journal of Chemical Education*, vol. 96, pp. 593–596, 3 2019.

[15] J. An, L.-P. Poly, and T. A. Holme, "Usability Testing and the Development of an Augmented Reality Application for Laboratory Learning," *Journal of Chemical Education*, vol. 97, pp. 97–105, 1 2020.

[16] O. C. Agbonifo, O. A. Sarumi, and Y. M. Akinola, "A chemistry laboratory platform enhanced with virtual reality for students' adaptive learning," *Research in Learning Technology*, vol. 28, no. 1063519, pp. 1–9, 2020.

[17] B. J. Wu, S. K. Wong, and T. W. Li, "Virtual titration laboratory experiment with differentiated instruction," *Computer Animation and Virtual Worlds*, vol. 30, no. 3-4, pp. 1–11, 2019.

[18] N. Y. K. Tee, H. S. Gan, J. Li, B. H. P. Cheong, H. Y. Tan, O. W. Liew, and T. W. Ng, "Developing and Demonstrating an Augmented Reality Colorimetric Titration Tool," *Journal of Chemical Education*, vol. 95, no. 3, pp. 393–399, 2018.

[19] J. Bacca, S. Baldiris, R. Fabregat, S. Graf, and Kinshuk, "Augmented Reality Trends in Education: A Systematic Review of Research and Applications," *Journal of Educational Technology & Society*, vol. 17, pp. 133–149, 4 2014.

[20] S. Irshad and D. Rohaya Bt Awang Rambli, "User experience of mobile augmented reality: A review of studies," in *2014 3rd International Conference on User Science and Engineering (i-USEr)*, no. February 2015, pp. 125–130, IEEE, 9 2014.

[21] H.-K. Wu, S. W.-Y. Lee, H.-Y. Chang, and J.-C. Liang, "Current status, opportunities and challenges of augmented reality in education," *Computers & Education*, vol. 62, pp. 41–49, 3 2013.

[22] M. Akçayir and G. Akçayir, "Advantages and challenges associated with augmented reality for education: A systematic review of the literature," *Educational Research Review*, vol. 20, pp. 1–11, 2 2017.

[23] J. Garzón, J. Pavón, and S. Baldiris, "Systematic review and meta-analysis of augmented reality in educational settings," *Virtual Reality*, vol. 23, pp. 447–459, 12 2019.

[24] M. E. C. Santos, A. Chen, T. Taketomi, G. Yamamoto, J. Miyazaki, and H. Kato, "Augmented Reality Learning Experiences: Survey of Prototype Design and Evaluation," *IEEE Transactions on Learning Technologies*, vol. 7, pp. 38–56, 1 2014.

[25] H. Tekedere, "Examining the Effectiveness of Augmented Reality Applications in Education: A Meta-Analysis," *International Journal of Environmental & Science Education*, vol. 11, no. 16, pp. 9469–9481, 2016.

[26] M. Özdemir, C. Sahin, S. Arcagok, and M. K. Demir, "The Effect of Augmented Reality Applications in the Learning Process: A MetaAnalysis Study," *Eurasian Journal of Educational Research*, vol. 18, pp. 1–22, 4 2018.

[27] J. Garzón and J. Acevedo, "Meta-analysis of the impact of Augmented Reality on students' learning gains," *Educational Research Review*, vol. 27, pp. 244–260, 6 2019.

[28] Z. A. Yilmaz and V. Batdi, "A Meta-Analytic and Thematic Comparative Analysis of the Integration of Augmented Reality Applications into Education," *Education and Science*, vol. 41, pp. 273–289, 12 2016.

[29] R. Wojciechowski and W. Cellary, "Evaluation of learners' attitude toward learning in ARIES augmented reality environments," *Computers & Education*, vol. 68, pp. 570–585, 10 2013.

[30] M. Brown, M. McCormack, J. Reeves, D. C. Brook, S. Grajek, B. Alexander, M. Bali, S. Bulger, S. Dark, N. Engelbert, K. Gannon, A. Gauthier, D. Gibson, R. Gibson, B. Lundin, G. Veletsianos, and N. Weber, "2020 Educause Horizon Report Teaching and Learning Edition," tech. rep., EDUCAUSE, Louisville, CO, 2020.

[31] J. Blattgerste, K. Luksch, C. Lewa, and T. Pfeiffer, "TrainAR: A Scalable Interaction Concept and Didactic Framework for Procedural Trainings Using Handheld Augmented Reality," *Multimodal Technologies and Interaction*, vol. 5, p. 30, 6 2021.

[32] M. L. Stone, K. M. Kent, R. D. Roscoe, K. M. Corley, L. K. Allen, and D. S. McNamara, "The Design Implementation Framework," in *End-User Considerations in Educational Technology Design* (R. D. Roscoe, S. D. Craig, and I. Douglas, eds.), pp. 76–98, IGI Global, 2018.

[33] S. D. Tripp and B. Bichelmeyer, "Rapid prototyping: An alternative instructional design strategy," *Educational Technology Research and Development*, vol. 38, pp. 31–44, 3 1990.

[34] C.-S. Yang, D. M. Moore, and J. K. Burton, "Managing courseware production: An instructional design model with a software engineering approach," *Educational Technology Research and Development*, vol. 43, pp. 60–70, 12 1995.

[35] T. S. Jones and R. C. Richey, "Rapid prototyping methodology in action: A developmental study," *Educational Technology Research and Development*, vol. 48, pp. 63–80, 6 2000.

[36] A. Suh and J. Prophet, "The state of immersive technology research: A literature analysis," *Computers in Human Behavior*, vol. 86, pp. 77–90, 2018.

[37] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino, "Augmented reality: A class of displays on the reality-virtuality continuum," *Telemanipulator and Telepresence Technologies*, vol. 2351, no. December 2013, pp. 282–292, 1995.

[38] P. Milgram and F. Kishino, "A Taxonomy of Mixed Reality Visual Displays," *IEICE Transactions on Information Systems*, no. 12, pp. 1321–1329, 1994.

[39] P. A. Rauschnabel, R. Felix, C. Hinsch, H. Shahab, and F. Alt, "What is XR? Towards a Framework for Augmented and Virtual Reality," *Computers in Human Behavior*, vol. 133, p. 107289, 8 2022.

[40] I. E. Sutherland, "A head-mounted three dimensional display," in *Fall Joint Computer Conference*, vol. 33, (New York, New York, USA), pp. 757–764, ACM Press, 1968.

[41] T. Caudell and D. Mizell, "Augmented reality: an application of heads-up display technology to manual manufacturing processes," in *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*, pp. 659–669, IEEE, 1992.

[42] R. T. Azuma, "A Survey of Augmented Reality," *Presence: Teleoperators and Virtual Environments*, vol. 6, pp. 355–385, 8 1997.

[43] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre, "Recent advances in augmented reality," *IEEE Computer Graphics and Applications*, vol. 21, no. 6, pp. 34–47, 2001.

[44] F. C.-A. Janin, "Boeing Wire Bundles, An Augmented Reality Application," in *Understanding Virtual Reality*, pp. 513–520, Elsevier, 2003.

[45] J. Herling and W. Broll, *Handbook of Augmented Reality*. New York, NY: Springer New York, 2011.

[46] R. Doerner, W. Broll, P. Grimm, and B. Jung, eds., *Virtual and Augmented Reality (VR/AR)*. Cham: Springer International Publishing, 2022.

[47] M. Fiala, "Designing Highly Reliable Fiducial Markers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1317–1324, 7 2010.

[48] D. Jurado-Rodriguez, R. Muñoz-Salinas, S. Garrido-Jurado, and R. Medina-Carnicer, "Planar fiducial markers: a comparative study," *Virtual Reality*, 2 2023.

[49] K.-H. Cheng and C.-C. Tsai, "Affordances of Augmented Reality in Science Learning: Suggestions for Future Research," *Journal of Science Education and Technology*, vol. 22, pp. 449–462, 8 2013.

[50] M. Akçayir, G. Akçayir, H. Pektas, and M. Ocak, "Augmented reality in science laboratories: The effects of augmented reality on university students' laboratory skills and attitudes toward science laboratories," *Computers in Human Behavior*, vol. 57, pp. 334–342, 4 2016.

[51] J. R. Schmid, M. J. Ernst, and G. Thiele, "Structural Chemistry 2.0: Combining Augmented Reality and 3D Online Models," *Journal of Chemical Education*, vol. 97, pp. 4515–4519, 12 2020.

[52] M. P. Chen and B. C. Liao, "Augmented reality laboratory for high school electrochemistry course," *Proceedings - IEEE 15th International Conference on Advanced Learning Technologies: Advanced Technologies for Supporting Open Access to Formal and Informal Learning, ICALT 2015*, pp. 132–136, 2015.

[53] P. Gunawan, J. Kwan, Y. Cai, and R. Yang, "Augmented Reality Application for Chemical Engineering Unit Operations," pp. 29–43, Springer Singapore, 2021.

[54] S. Solmaz, J. L. Dominguez Alfaro, P. Santos, P. Van Puyvelde, and T. Van Gerven, "A practical development of engineering simulation-assisted educational AR environments," *Education for Chemical Engineers*, vol. 35, pp. 81–93, 4 2021.

[55] P. Nowacki and M. Woda, "Capabilities of ARCore and ARKit Platforms for AR/VR Applications," vol. 987 of *Advances in Intelligent Systems and Computing*, pp. 358–370, Cham: Springer International Publishing, 2020.

[56] C. Kivunja, "Theoretical Perspectives of How Digital Natives Learn," *International Journal of Higher Education*, vol. 3, no. 1, 2014.

[57] J. P. Byrnes, "Categorizing and combining theories of cognitive development and learning," *Educational Psychology Review*, vol. 4, no. 3, pp. 309–343, 1992.

[58] "Encyclopedia of the Sciences of Learning," in *Encyclopedia of the Sciences of Learning* (N. M. Seel, ed.), pp. 495–904, Boston, MA: Springer US, 2012.

[59] B. Akpan and T. J. Kennedy, eds., *Science Education in Theory and Practice*. Springer Texts in Education, Cham: Springer International Publishing, 2020.

[60] N. M. Seel, ed., *Encyclopedia of the Sciences of Learning*. Boston, MA: Springer New York, NY, 2012.

[61] D. A. Kolb, "The Process of Experiential Learning," in *Strategic Learning in a Knowledge Economy*, no. 1984, pp. 313–331, Elsevier, 2000.

[62] L. A. Hirschfeld, *Mapping the Mind*. London: Cambridge University Press, 4 1994.

[63] J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, "Expert and Novice Performance in Solving Physics Problems," *Science*, vol. 208, pp. 1335–1342, 6 1980.

[64] S. L. Bretz, "Novak's theory of education: Human constructivism and meaningful learning," *Journal of Chemical Education*, vol. 78, no. 8, p. 1107, 2001.

[65] J. Sweller, J. J. G. van Merriënboer, and F. Paas, "Cognitive Architecture and Instructional Design: 20 Years Later," *Educational Psychology Review*, vol. 31, pp. 261–292, 6 2019.

[66] J. J. G. van Merriënboer and L. Kester, "The Four-Component Instructional Design Model : Multimedia Principles in Environments for Complex Learning," in *The Cambridge Handbook of Multimedia Learning* (R. Mayer, ed.), pp. 71–94, Cambridge: Cambridge University Press, 8 2005.

[67] R. E. Mayer, *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press, 7 2014.

[68] P. Kirschner and J. J. G. van Merrienboer, "Ten Steps to Complex Learning: A New Approach to Instruction and Instructional Design," in *21st Century Education: A Reference Handbook*, pp. I–244–I–253, 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., 2008.

[69] E. Sung and R. E. Mayer, "When graphics improve liking but not learning from online lessons," *Computers in Human Behavior*, vol. 28, pp. 1618–1625, 9 2012.

[70] E. Van Genuchten, K. Scheiter, and A. Schüler, "Examining learning from text and pictures for different task types: Does the multimedia effect differ for conceptual, causal, and procedural tasks?," *Computers in Human Behavior*, vol. 28, no. 6, pp. 2209–2218, 2012.

[71] M. Pedaste, M. Mäeots, L. A. Siiman, T. de Jong, S. A. van Riesen, E. T. Kamp, C. C. Manoli, Z. C. Zacharia, and E. Tsourlidaki, "Phases of inquiry-based learning: Definitions and the inquiry cycle," *Educational Research Review*, vol. 14, pp. 47–61, 2 2015.

[72] M. Pedaste, G. Mitt, and T. Jürivete, "What Is the Effect of Using Mobile Augmented Reality in K12 Inquiry-Based Learning," *Education Sciences*, 2020.

[73] L. Shapiro and S. A. Stolz, "Embodied cognition and its significance for education," *Theory and Research in Education*, vol. 17, pp. 19–39, 3 2019.

[74] S. M. Weisberg and N. S. Newcombe, "Embodied cognition and STEM learning: overview of a topical collection in CR:PI," *Cognitive Research: Principles and Implications*, vol. 2, no. 1, 2017.

[75] K. R. Bujak, I. Radu, R. Catrambone, B. MacIntyre, R. Zheng, and G. Golubski, "A psychological perspective on augmented reality in the mathematics classroom," *Computers & Education*, vol. 68, pp. 536–544, 10 2013.

[76] I. Radu and A. Antle, "Embodied learning mechanics and their relationship to usability of handheld augmented reality," *2017 IEEE Virtual Reality Workshop on K-12 Embodied Learning through Virtual and Augmented Reality, KELVAR 2017*, pp. 3–7, 2017.

[77] P. E. Kourouthanassis, C. Boletsis, and G. Lekakos, "Demystifying the design of mobile augmented reality applications," *Multimedia Tools and Applications*, vol. 74, pp. 1045–1066, 2 2015.

[78] M. Mekni and A. Lemieux, "Augmented Reality : Applications, Challenges and Future Trends," in *13th International Conference on Applied Computer and Computation Sciences (ACACOS'14)*, pp. 205–214, WSEAS Press, 2014.

[79] M.-B. Ibáñez and C. Delgado-Kloos, "Augmented reality for STEM learning: A systematic review," *Computers & Education*, vol. 123, pp. 109–123, 8 2018.

[80] I. Radu, "Augmented reality in education: a meta-review and cross-media analysis," *Personal and Ubiquitous Computing*, vol. 18, pp. 1533–1543, 8 2014.

[81] A. Dirin and T. Laine, "User Experience in Mobile Augmented Reality: Emotions, Challenges, Opportunities and Best Practices," *Computers*, vol. 7, p. 33, 5 2018.

[82] H. Li, A. Gupta, J. Zhang, and N. Flor, "Who will use augmented reality? An integrated approach based on text analytics and field survey," *European Journal of Operational Research*, vol. 281, pp. 502–516, 3 2020.

[83] T. Olsson and M. Salo, "Online user survey on current mobile augmented reality applications," in *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pp. 75–84, IEEE, 10 2011.

[84] Y. Arifin, T. G. Sastria, and E. Barlian, "User Experience Metric for Augmented Reality Application: A Review," *Procedia Computer Science*, vol. 135, pp. 648–656, 2018.

[85] S. L. Schneider, "The International Standard Classification of Education 2011," in *Class and Stratification Analysis* (G. E. Birkelund, ed.), vol. 30 of *Comparative Social Research*, pp. 365–379, Emerald Group Publishing Limited, 1 2013.

[86] ISO, "ISO 9241-210: Ergonomics of human-system interaction, Part 210: Human-centred design for interactive systems." 2019.

[87] R. Salar, F. Arici, S. Caliklar, and R. M. Yilmaz, "A Model for Augmented Reality Immersion Experiences of University Students Studying in Science Education," *Journal of Science Education and Technology*, vol. 29, pp. 257–271, 4 2020.

[88] C. Iacob and R. Harrison, "Retrieving and analyzing mobile apps feature requests from online reviews," in *2013 10th Working Conference on Mining Software Repositories (MSR)*, (San Francisco, CA, USA), pp. 41–44, IEEE, 5 2013.

[89] D. Pagano and W. Maalej, "User feedback in the appstore: An empirical study," in *2013 21st IEEE International Requirements Engineering Conference (RE)*, pp. 125–134, IEEE, 7 2013.

[90] K. S. Thach, "User's perception on mental health applications: a qualitative analysis of user reviews," in *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, pp. 47–52, IEEE, 11 2018.

[91] J. Nicholas, A. S. Fogarty, K. Boydell, and H. Christensen, "The Reviews Are in: A Qualitative Content Analysis of Consumer Perspectives on Apps for Bipolar Disorder," *Journal of Medical Internet Research*, vol. 19, p. e105, 4 2017.

[92] F. Alqahtani and R. Orji, "Insights from user reviews to improve mental health apps," *Health Informatics Journal*, vol. 26, pp. 2042–2066, 9 2020.

[93] S. Panichella, A. Di Sorbo, E. Guzman, C. A. Visaggio, G. Canfora, and H. C. Gall, "ARdoc: app reviews development oriented classifier," in *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering - FSE 2016*, vol. 13-18-Nove, (New York, New York, USA), pp. 1023–1027, ACM Press, 11 2016.

[94] C. Iacob, R. Harrison, and S. Faily, "Online Reviews as First Class Artifacts in Mobile App Development," in *Mobile Computing, Applications, and Services* (G. Memmi and U. Blanke, eds.), pp. 47–53, Cham: Springer International Publishing, 2014.

[95] S. Panichella, A. Di Sorbo, E. Guzman, C. A. Visaggio, G. Canfora, and H. C. Gall, "How can i improve my app? Classifying user reviews for software maintenance and evolution," in *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, no. 2, (Bremen, Germany), pp. 281–290, IEEE, 9 2015.

[96] J. L. Dominguez Alfaro and P. Van Puyvelde, "Mobile Augmented Reality Apps in Education: Exploring the User Experience Through Large-Scale Public Reviews," in *Lecture Notes in Computer Science* (L. De Paolis, P. Arpaia, and P. Bourdot, eds.), vol. 12980, pp. 428–450, Cham: Springer International Publishing, 2021.

[97] A. Mackey and S. M. Gass, *Second language research: Methodology and design.* Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers, 2005.

[98] H. Chen, Y. Dai, H. Meng, Y. Chen, and T. Li, "Understanding the Characteristics of Mobile Augmented Reality Applications," in *2018 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, no. July, pp. 128–138, IEEE, 4 2018.

[99] X. Qiao, P. Ren, S. Dustdar, L. Liu, H. Ma, and J. Chen, "Web AR: A Promising Future for Mobile Augmented Reality - State of the Art, Challenges, and Insights," *Proceedings of the IEEE*, vol. 107, pp. 651–666, 4 2019.

[100] X. Qiao, P. Ren, G. Nan, L. Liu, S. Dustdar, and J. Chen, "Mobile web augmented reality in 5G and beyond: Challenges, opportunities, and future directions," *China Communications*, vol. 16, pp. 141–154, 9 2019.

[101] D. Harborth and S. Pape, "Investigating privacy concerns related to mobile augmented reality Apps - A vignette based online experiment," *Computers in Human Behavior*, vol. 122, p. 106833, 9 2021.

[102] J. Glassey and F. D. Magalhães, "Virtual labs- love them or hate them, they are likely to be used more in the future," *Education for Chemical Engineers*, vol. 33, pp. 76–77, 10 2020.

[103] N. d'Heureuse, F. Huici, M. Arumaithurai, M. Ahmed, K. Papagiannaki, and S. Niccolini, "What's app?," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 16, pp. 16–27, 11 2012.

[104] M. Ahasanuzzaman, S. Hassan, C.-P. Bezemer, and A. E. Hassan, "A longitudinal study of popular ad libraries in the Google Play Store," *Empirical Software Engineering*, vol. 25, pp. 824–858, 1 2020.

[105] A. Mahmood, "Identifying the influence of various factor of apps on google play apps ratings," *Journal of Data, Information and Management*, vol. 2, pp. 15–23, 3 2020.

[106] Y. Tian, M. Nagappan, D. Lo, and A. E. Hassan, "What are the characteristics of high-rated apps? A case study on free Android Applications," in *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp. 301–310, IEEE, 9 2015.

[107] P. Nowacki and M. Woda, "Capabilities of ARCore and ARKit Platforms for AR/VR Applications," in *Engineering in Dependability of Computer Systems and Networks* (W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, and J. Kacprzyk, eds.), pp. 358–370, Cham: Springer International Publishing, 2020.

[108] P. Roberto, F. Emanuele, Z. Primo, M. Adriano, L. Jelena, and P. Marina, "Design, Large-Scale Usage Testing, and Important Metrics for Augmented Reality Gaming Applications," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, pp. 1–18, 6 2019.

[109] S. Papadakis and M. Kalogiannakis, "Mobile educational applications for children: what educators and parents need to know," *International Journal of Mobile Learning and Organisation*, vol. 11, no. 3, p. 256, 2017.

[110] K. Sheppard, "High school students' understanding of titrations and related acid-base phenomena," *Chem. Educ. Res. Pract.*, vol. 7, no. 1, pp. 32–45, 2006.

[111] J. McMurry and R. C. Fay, *Chemistry*. Upper Saddle River, NJ: Pearson Education, fourth ed., 2004.

[112] D. A. McQuarrie, P. A. Rock, and E. B. Gallogly, *General Chemistry*. Mill Valley, California SE -: University Science Books, fourth edi ed., 2011.

[113] J. Brady, N. Jespersen, and A. Hyslop, *Chemistry*. John Wiley & Sons, Inc, seventh ed ed., 2014.

[114] J. N. Butler, "Calculating titration errors," *Journal of Chemical Education*, vol. 40, p. 66, 2 1963.

[115] P. W. Airasian, L. W. Anderson, K. A. Cruikshank, D. R. Krathwohl, R. E. Mayer, P. R. Pintrich, J. Raths, and M. C. Wittrock, *Taxonomy for Assessing a Revision of Bloom's Taxonomy of Educational Objectives*. New York, NY, USA: Longman, 2001.

[116] S. Gantois, "Mobile augmented reality prototype of a titration experiment for chemical engineering education (Master Thesis)," 2021.

[117] J. B. Brooke, "SUS: A 'Quick and Dirty' Usability Scale," in *Usability Evaluation In Industry*, pp. 207–212, CRC Press, 6 1996.

[118] A. Bangor, T. Staff, P. Kortum, J. Miller, and T. Staff, "Determining what individual SUS scores mean: adding an adjective rating scale," *Journal of Usability Studies*, vol. 4, no. 3, pp. 114–123, 2009.

[119] J. Blattgerste, J. Behrends, and T. Pfeiffer, "A Web-Based Analysis Toolkit for the System Usability Scale," 2022.

[120] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. New York: Routledge, 5 1988.

[121] K. O. McGraw and S. P. Wong, "A Common Language Effect Size Statistic," *Psychological Bulletin*, vol. 111, pp. 361–365, 3 1992.

[122] P. D. Ellis, *The Essential Guide to Effect Sizes*. Cambridge University Press, 7 2010.

[123] R. Lindgren, M. Tscholl, S. Wang, and E. Johnson, "Enhancing learning and engagement through embodied interaction within a mixed reality simulation," *Computers and Education*, vol. 95, pp. 174–187, 2016.

[124] J. Silvennoinen, M. Vogel, and S. Kujala, "Experiencing Visual Usability and Aesthetics in Two Mobile Application Contexts," *Journal of Usability Studies*, vol. 10, no. 1, pp. 46–62, 2014.

[125] P. Chan, T. Van Gerven, J.-L. Dubois, and K. Bernaerts, "Virtual chemical laboratories: A systematic literature review of research, technologies and instructional design," *Computers and Education Open*, vol. 2, p. 100053, 12 2021.

[126] J. Buchner and M. Kerres, "Media comparison studies dominate comparative research on augmented reality in education," *Computers & Education*, vol. 195, p. 104711, 4 2023.

[127] W. van der Aalst, "Process Mining," in *Process Mining: Data Science in Action*, p. 33, Berlin, Heidelberg: Springer Berlin Heidelberg, 2016.

[128] P. Mukala, J. Buijs, M. Leemans, and W. van der Aalst, "Exploring Students' Learning Behaviour in MOOCs using Process Mining Techniques," *Computing Conference*, pp. 1–12, 2015.

[129] G. Sedrakyan, J. De Weerdt, and M. Snoeck, "Process-mining enabled feedback: "tell me what i did wrong" vs. "tell me how to do it right"," *Computers in Human Behavior*, vol. 57, pp. 352–376, 2016.

[130] R. Cerezo, A. Bogarín, M. Esteban, and C. Romero, "Process mining for self-regulated learning assessment in e-learning," *Journal of Computing in Higher Education*, vol. 32, no. 1, pp. 74–88, 2020.

[131] S. Heirweg, M. De Smul, E. Merchie, G. Devos, and H. Van Keer, *Mine the process: investigating the cyclical nature of upper primary school students' self-regulated learning*, vol. 48. Springer Netherlands, 2020.

[132] R. K. Hambleton and R. W. Jones, "Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development," *Educational Measurement: Issues and Practice*, vol. 12, pp. 38–47, 10 1993.

[133] T. L. Kelley, "The selection of upper and lower groups for the validation of test items.," *Journal of Educational Psychology*, vol. 30, pp. 17–24, 1 1939.

[134] K. S. Taber, "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education," *Research in Science Education*, vol. 48, no. 6, pp. 1273–1296, 2018.

[135] G. Guest, K. MacQueen, and E. Namey, *Applied Thematic Analysis*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., 2012.

[136] R. Vallat, "Pingouin: statistics in Python," *Journal of Open Source Software*, vol. 3, no. 31, p. 1026, 2018.

[137] JASP Team, "JASP (Version 0.17)[Computer software]," 2023.

[138] J. R. Lewis and J. Sauro, "Item Benchmarks for the System Usability Scale," *Journal of Usability Studies*, vol. 13, no. January, pp. 158–167, 2018.

[139] J. L. Domínguez Alfaro, S. Gantois, J. Blattgerste, R. De Croon, K. Verbert, T. Pfeiffer, and P. Van Puyvelde, "Mobile Augmented Reality Laboratory for Learning Acid-Base Titration," *Journal of Chemical Education*, vol. 99, pp. 531–537, 2 2022.

[140] I. Radu, B. Macintyre, and S. Lourenco, "Comparing children's crosshair and finger interactions in handheld augmented reality: Relationships between usability and child development," *Proceedings of IDC 2016* -

The 15th International Conference on Interaction Design and Children, pp. 288–298, 2016.

[141] T. S. Tullis and J. N. Stetson, "A Comparison of Questionnaires for Assessing Website Usability," in *Usability Professional Association Conference*, no. June, pp. 1–12, 2004.

[142] W. De Schepper, "Modeling the learning process of students in an AR application using process mining techniques (Master Thesis)," 2022.

[143] J. Buchner, K. Buntins, and M. Kerres, "The impact of augmented reality on cognitive load and performance: A systematic review," *Journal of Computer Assisted Learning*, vol. 38, no. 1, pp. 285–303, 2022.

[144] E. E. Goff, K. L. Mulvey, M. J. Irvin, and A. Hartstone-Rose, "Applications of Augmented Reality in Informal Science Learning Sites: a Review," *Journal of Science Education and Technology*, vol. 27, no. 5, pp. 433–447, 2018.

[145] P. Sommerauer and O. Müller, "Augmented reality in informal learning environments: A field experiment in a mathematics exhibition," *Computers and Education*, vol. 79, no. 2014, pp. 59–68, 2014.

[146] A. Tang, C. Owen, F. Biocca, and W. Mou, "Comparative effectiveness of augmented reality in object assembly," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, vol. 9, (New York, NY, USA), pp. 73–80, ACM, 4 2003.

[147] M. Dunleavy, C. Dede, and R. Mitchell, "Affordances and limitations of immersive participatory augmented reality simulations for teaching and learning," *Journal of Science Education and Technology*, vol. 18, no. 1, pp. 7–22, 2009.

[148] M. B. Ibáñez, A. Di-Serio, D. Villarán-Molina, and C. Delgado-Kloos, "Support for Augmented Reality Simulation Systems: The Effects of Scaffolding on Learning Outcomes and Behavior Patterns," *IEEE Transactions on Learning Technologies*, vol. 9, no. 1, pp. 46–56, 2016.

[149] H.-Y. Chang, T. Binali, J.-C. Liang, G.-L. Chiou, K.-H. Cheng, S. W.-Y. Lee, and C.-C. Tsai, "Ten years of augmented reality in education: A meta-analysis of (quasi-) experimental studies to investigate the impact," *Computers & Education*, vol. 191, p. 104641, 12 2022.

[150] D. Debeer, S. Vanbecelaere, W. Van Den Noortgate, B. Reynvoet, and F. Depaepe, "The effect of adaptivity in digital learning technologies. Modelling learning efficiency using data from an educational game," *British Journal of Educational Technology*, vol. 52, no. 5, pp. 1881–1897, 2021.

[151] V. J. Shute, M. Ventura, and F. Ke, "The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills," *Computers & Education*, vol. 80, pp. 58–67, 1 2015.

[152] J. Wang, A. Stebbins, and R. E. Ferdig, "Examining the effects of students' self-efficacy and prior knowledge on learning and visual behavior in a physics game," *Computers & Education*, vol. 178, p. 104405, 3 2022.

[153] K. E. Godwin, H. Seltman, M. Almeda, M. Davis Skerbetz, S. Kai, R. S. Baker, and A. V. Fisher, "The elusive relationship between time on-task and learning: not simply an issue of measurement," *Educational Psychology*, vol. 41, no. 4, pp. 502–519, 2021.

[154] R. B. Brünken, J. L. Plass, and D. L. Leutner, "Direct Measurement of Cognitive Load in Multimedia Leaming," *Educational Psychologist: A Special Issue of educational Psychologist: Volume 38*, vol. 38, no. May 2013, pp. 53–61, 2016.

[155] J. Y. Lee, J. Donkers, H. Jarodzka, G. Sellenraad, and J. J. van Merriënboer, "Different effects of pausing on cognitive load in a medical simulation game," *Computers in Human Behavior*, vol. 110, p. 106385, 9 2020.

[156] N. Sevcenko, M. Ninaus, F. Wortha, K. Moeller, and P. Gerjets, "Measuring Cognitive Load Using In-Game Metrics of a Serious Simulation Game," *Frontiers in Psychology*, vol. 12, 3 2021.

[157] G. Makransky and G. B. Petersen, "The Cognitive Affective Model of Immersive Learning (CAMIL): a Theoretical Research-Based Model of Learning in Immersive Virtual Reality," *Educational Psychology Review*, vol. 33, no. 3, pp. 937–958, 2021.

[158] Ø. Anmarkrud, A. Andresen, and I. Bråten, "Cognitive Load and Working Memory in Multimedia Learning: Conceptual and Measurement Issues," *Educational Psychologist*, vol. 54, pp. 61–83, 4 2019.

[159] J. Buchner, K. Buntins, and M. Kerres, "A systematic map of research characteristics in studies on augmented reality and cognitive load," *Computers and Education Open*, vol. 2, no. April, p. 100036, 2021.

[160] E. N. Wiebe, E. Roberts, and T. S. Behrend, "An examination of two mental workload measurement approaches to understanding multimedia learning," *Computers in Human Behavior*, vol. 26, no. 3, pp. 474–481, 2010.

[161] S. G. Hart, "NASA-task load index (NASA-TLX); 20 years later," *Proceedings of the Human Factors and Ergonomics Society*, pp. 904–908, 2006.

[162] E. Lampen, J. Teuber, F. Gaisbauer, T. Bär, T. Pfeiffer, and S. Wachsmuth, "Combining simulation and augmented reality methods for enhanced worker assistance in manual assembly," *Procedia CIRP*, vol. 81, pp. 588–659, 2019.

[163] J. Alves, B. Marques, M. Oliveira, T. Araújo, P. Dias, and B. S. Santos, "Comparing Spatial and Mobile Augmented Reality for Guiding Assembling Procedures with Task Validation," *19th IEEE International Conference on Autonomous Robot Systems and Competitions, ICARSC 2019*, pp. 0–5, 2019.

[164] T. de Jong, "Cognitive load theory, educational research, and instructional design: Some food for thought," *Instructional Science*, vol. 38, no. 2, pp. 105–134, 2010.

[165] P. Gerjets, K. Scheiter, and R. Catrambone, "Designing instructional examples to reduce intrinsic cognitive load: Molar versus modular presentation of solution procedures," *Instructional Science*, vol. 32, no. 1-2, pp. 33–58, 2004.

[166] P. Gerjets, K. Scheiter, and R. Catrambone, "Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations?," *Learning and Instruction*, vol. 16, no. 2 SPEC. ISS., pp. 104–121, 2006.

[167] A. Field, *Discovering Statistics Using IBM SPSS Statistics*. Sage Publications, 2013.

[168] I. S. Mackenzie, *Human-Computer Interaction: An Empirical Research Perspective*. Elsevier, 2013.

[169] W. J. Conover, *Practical Nonparametric Statistics*. John Wiley, 1999.

[170] H. Hoijtink, J. Mulder, C. van Lissa, and X. Gu, "A Tutorial on Testing Hypotheses Using the Bayes Factor," *Psychological Methods*, 2019.

[171] H. Hoijtink, *Informative Hypotheses*. New York: Chapman and Hall/CRC, 10 2011.

[172] C. J. Van Lissa, X. Gu, J. Mulder, Y. Rosseel, C. Van Zundert, and H. Hoijtink, "Teacher's Corner: Evaluating Informative Hypotheses Using the Bayes Factor in Structural Equation Models," *Structural Equation Modeling*, vol. 28, no. 2, pp. 292–301, 2021.

[173] X. Gu, H. Hoijtink, J. Mulder, and C. van Lissa, "bain: Bayes factors for informative hypotheses," 2021.

[174] Y. Rosseel, "lavaan: An R Package for Structural Equation Modeling," *Journal of Statistical Software*, vol. 48, no. 2, pp. 1–36, 2012.

[175] M. C. Johnson-Glenberg, D. A. Birchfield, L. Tolentino, and T. Koziupa, "Collaborative embodied learning in mixed reality motion-capture environments: Two science studies," *Journal of Educational Psychology*, vol. 106, no. 1, pp. 86–104, 2014.

[176] M. C. Johnson-Glenberg and C. Megowan-Romanowicz, "Embodied science and mixed reality: How gesture and motion capture affect physics education," *Cognitive Research: Principles and Implications*, vol. 2, no. 1, 2017.

[177] M. Z. Iqbal, E. Mangina, and A. G. Campbell, "Current Challenges and Future Research Directions in Augmented Reality for Education," *Multimodal Technologies and Interaction*, vol. 6, p. 75, 9 2022.

[178] G. Makransky, P. Wismer, and R. E. Mayer, "A gender matching effect in learning with pedagogical agents in an immersive virtual reality science simulation," *Journal of Computer Assisted Learning*, vol. 35, no. 3, pp. 349–358, 2019.

[179] D. N. Eh Phon, M. B. Ali, and N. D. A. Halim, "Collaborative augmented reality in education: A review," *Proceedings - 2014 International Conference on Teaching and Learning in Computing and Engineering, LATICE 2014*, pp. 78–83, 2014.

[180] O. A. Meyer, M. K. Omdahl, and G. Makransky, "Investigating the effect of pre-training when learning through immersive virtual reality and video: A media and methods experiment," *Computers and Education*, vol. 140, no. December 2018, p. 103603, 2019.

[181] M. Perifanou, A. A. Economides, and S. A. Nikou, "TeachersâĂŹ Views on Integrating Augmented Reality in Education: Needs, Opportunities, Challenges and Recommendations," *Future Internet*, vol. 15, p. 20, 12 2022.

[182] M. Heintz, E. L.-C. Law, and P. Andrade, "Augmented Reality as Educational Tool: Perceptions, Challenges, and Requirements from Teachers," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12884 LNCS, pp. 315–319, Springer International Publishing, 2021.

[183] N. Alalwana, L. Cheng, H. Al-Samarraie, R. Yousef, A. Ibrahim Alzahrani, and S. M. Sarsam, "Challenges and Prospects of Virtual Reality and Augmented Reality Utilization among Primary School Teachers: A Developing Country Perspective," *Studies in Educational Evaluation*, vol. 66, p. 100876, 9 2020.

[184] A. Dengel, M. Z. Iqbal, S. Grafe, and E. Mangina, "A Review on Augmented Reality Authoring Toolkits for Education," *Frontiers in Virtual Reality*, vol. 3, pp. 1–15, 4 2022.

[185] R. De Levie, "Explicit expressions of the general form of the titration curve in terms of concentration: Writing a single closed-form expression for the titration curve for a variety of titrations without using approximations or segmentation," *Journal of Chemical Education*, vol. 70, no. 3, pp. 209–217, 1993.

[186] M. Gao, P. Kortum, and F. L. Oswald, "Multi-Language Toolkit for the System Usability Scale," *International Journal of Human-Computer Interaction*, vol. 36, no. 20, pp. 1883–1901, 2020.

[187] D. Koning, *Ontwerp van een online zelf-assessment voor het meten van de fysieke activiteit bij ouderen tussen de 55 en 75 jaar.* PhD thesis, 2016.

[188] A. Bemer, *Slimmer door de Smartphone?: Onderzoek naar Informatieverwerking met Gebruik van Verschillende Devices.* PhD thesis, Open Universiteit, 2019.

Curriculum vitae

Jessica Lizeth Domínguez Alfaro was born in Mexico City and completed her bachelor's degree in Chemistry at Universidad Nacional Autónoma de México (UNAM). She then pursued a master's degree in Chemistry with a specialisation in Science-Based Business at Leiden University, where she was supervised by Prof. dr. M. van der Stelt. In 2019, Jessica began her PhD candidacy at KU Leuven, working in the Softmatter, Rheology and Technology (SMaRT) group.

Her research project was part of the Chemical Engineering Immersive Learning (CHARMING) European Training Network, where she focused on exploring how mobile augmented reality can be used to learn chemical practical skills. The scientific contributions are presented in this thesis, which was funded by the European Union's EU Framework Programme for Research and Innovation Horizon 2020 under Grant Agreement No 812716. Moreover, Jessica has also made the application developed for this dissertation available to the public, and it can be found at https://alfajess92.github.io/marlab/.

In addition to research, Jessica has also worked in the private sector. She held several positions, including Graduate Trainee at Croda International, Business Intelligence Intern at DSM Sinochem Pharmaceuticals, and Marketing Specialist at International Flavours and Fragrances (IFF).

Following the completion of her PhD, Jessica will take on the role of Junior Manufacturing Project Manager at Signify, where she will continue to leverage new digital tools for dealing with new challenges in Industry 4.0.

List of publications

Publications in peer-reviewed academic journals

- S. Solmaz, **J. L. Domínguez Alfaro**, P. Santos, P. Van Puyvelde, and T. Van Gerven, "A practical development of engineering simulation-assisted educational AR environments," Education for Chemical Engineers, vol. 35, pp. 81-93, 4 2021.
- **J. L. Domínguez Alfaro**, S. Gantois, J. Blattgerste, R. De Croon, K. Verbert, T. Pfeiffer, and P. Van Puyvelde, "Mobile Augmented Reality Laboratory for Learning Acid-Base Titration," Journal of Chemical Education, vol. 99, pp. 531-537, 2 2022.
- **J. L. Domínguez Alfaro**, M. Arztmann, J. Jeuring and P. Van Puyvelde, "AR for Science Education: Students' Behaviour Patterns and the Relationship between Cognitive Load, Knowledge Acquisition and Performance," British Journal of Educational Technology, *Submitted*.
- D. Cermak-Sassenrath, S. Solmaz, **J. L. Domínguez Alfaro**, and C. Udeozor, "Bunno's Fabulous Soap-Making Challenge-Educational Concept, Game Design and Initial Evaluation," Chemical Engineering & Technology, pp. 1-15, 2 2023.

International conference proceedings peer-reviewed – full papers

- **J. L. Domínguez Alfaro** and P. Van Puyvelde, "Mobile Augmented Reality Apps in Education: Exploring the User Experience Through Large-Scale Public Reviews," in Lecture Notes in Computer Science, vol. 12980, L. T. De Paolis, P. Arpaia, and P. Bourdot, Eds. Cham: Springer International Publishing, 2021, pp. 428-450.

- **J. L. Domínguez Alfaro** and P. Van Puyvelde, "Analysis of the Learning Experience in a Chemical AR Application Using Process Mining," in Lecture Notes in Networks and Systems, vol. 633, M. E. Auer, W. Pachatz, and T. Rüütmann, Eds. Cham: Springer International Publishing, 2023, pp. 770-777.
- M. Arztmann, **J. L. Domínguez Alfaro**, J. Blattgerste, J. Jeuring, and P. Van Puyvelde, "Marie's ChemLab: a Mobile Augmented Reality Game to Teach Basic Chemistry to Children," in 16th European Conference on Games Based Learning, (Lisbon), 2022.
- **J. L. Domínguez Alfaro**, C. Udeozor, S. Solmaz, and D. Cermak-Sassenrath, "Designing a Mobile Game for Introducing Learners to a Soap Making process," in 16th European Conference on Games Based Learning, (Lisbon), pp. 1-7, 2022.
- C. Udeozor, **J. L. Domínguez Alfaro**, J. Glassey, "Assessment framework for immersive learning: application and evaluation," in 9th International Conference of the Immersive Learning Research Network, CA, 2023, *short paper accepted.*

International conference proceedings- abstracts/presentations

- **Domínguez Alfaro, J.L.**, Van Puyvelde, P. (2022). Mobile Augmented Reality: a new way to train in the chemical lab! Presented at the Biennial Conference on Chemical Education (BCCE), West Lafayette, Indiana, 31 Jul 2022-04 Aug 2022. (Oral Presentation)
- **Domínguez Alfaro, J.L.**, Van Puyvelde, P. (2022). Analysis of the Learning Experience in a Chemical AR Application using Process Mining. Presented at the International Conference on Interactive Collaborative Learning, Vienna, Austria, 27 Sep 2022-30 Sep 2022. (Oral Presentation)
- **Domínguez Alfaro, J.L.**, Solmaz, S., Cermak-Sassenrath, D., Arztmann, M., Udeozor, C. (2021). An Experimental Casual Game to Educate Players in Soap Production. (Abstract No. 31624). Presented at the 13th European Congress of Chemical Engineering and 6th European Congress of Applied Biotechnology, Online, 20 Sep 2021-23 Sep 2021. (Oral Presentation)



FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF CHEMICAL ENGINEERING
SOFT MATTER RHEOLOGY AND TECHNOLOGY
Celestijnenlaan 200J box 2424
B-3001 Leuven
jessicalizeth.dominguezalfaro@kuleuven.be
http://www.cit.kuleuven.be/smart