



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Usability Evaluation of Spoken Humanoid Embodied Conversational Agents in Mobile Serious Games

Danai Korre



Doctor of Philosophy

The University of Edinburgh

2019

Declaration of originality

2019

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

Danai Korre



17/10/2019

Abstract

The use of embodied conversational agents (ECAs) and spoken dialogue systems in serious games offers theoretical advantages such as a more natural interaction with an agent displaying characteristics like personality, engagement, enjoyment, trust and emotions. Despite these theoretical advantages, according to recent studies, the interaction with spoken dialogue systems, either in the form of an embodied agent or not, is still inferior compared to other approaches that allow a direct manipulation of the system. However, the way users interact with mobile devices is rapidly changing, since the latest generation of mobile devices include voice driven virtual assistants (Apple Siri, Google Now, Samsung S Voice, Amazon Alexa). Previous research has focused on the design aspects of ECAs but there are limited empirical evaluations regarding their effectiveness in serious games and mobile serious games. In an era where usability has become an integral part of the development process, introducing ECAs in these environments without proper evaluation can be problematic. Thus, there is a strong reason to examine if ECAs enhance usability over current interaction paradigms in serious game environments, even more so in mobile devices as there is a recent trend towards mobile serious games.

The research presented here investigates, across a series of two large scale experiments and a survey, the extent to which spoken Humanoid Embodied Conversational Agents (HECAs) can foster usability in mobile serious game (MSG) applications. The aim of the research is to assess the impact of multiple agents, serious game approaches and illusion of humanness on the quality of the interaction.

The first experiment (pilot study 1) investigates whether the portrayal of an application as a game (with game-like implicit feedback) influences the overall usability of a virtual application. The main purpose of this study is to act as a methodological sandbox to inform the methodology approaches of the main experiment. Qualitative analysis of the experiment shows that 78% of participants

prefer the game version. Also, the game version was perceived as more fun, enjoyable and stimulating.

Results of the survey (study 2) show that 83% of the participants have played games within the last 6 months and 60% of them play games on their smartphone even though they also own laptops as well as desktops and game consoles. The device of choice for everyday activities is also the smartphone. Moreover, most participants replied that they owned a smartphone with a screen size over 5". The data collected from the preliminary studies informed the hardware, methodological and design decisions of the main experiment.

The main mobile experiment investigates two styles of agent presentation, an agent of high human-likeness (HECA) and an agent of low human-likeness (text). The purpose of the experiment is to access how agents of high human-likeness can evoke the illusion of humanness and affect usability. Agents of high human-likeness were designed by following the ECA design model that is a proposed guide for ECA development. The results of the experiment show that users prefer to interact with the HECA. The difference between the two versions is statistically significant with a large effect size and many of the participants justifying their choice by saying that the human-like characteristics of the HECA made the version more appealing. This research provides key information on the potential effect of HECA on serious games, which will likely impact the design decisions regarding spoken HECA and the design of future mobile serious games.

Lay summary

Embodied conversational agents (ECAs) are virtual characters that can communicate with people through voice and/or text. Even though they have been around for a while, their complexity and flexibility to be used in different contexts allows for further investigation. One context that ECAs have not been tested much in are mobile serious games (MSGs). Mobile serious games are mobile games but with a purpose. Sometimes it is education, others training etc. Many big companies invest in serious games and there is a trend towards MSGs that makes evaluation in the field valuable.

However not all ECAs are equal. For example, there are ECAs with whom you can communicate using voice like Amazon Alexa and others that use text. Another example is the way these agents are presented. One agent might look like a realistic animated human while another one might look like an animal or a fantastic figure.

Therefore, we are studying the effect that specific types of ECAs have on usability. More specifically how the "illusion of humanness" evoked by a human-like ECA can affect the usability of the application. We will be using a serious game called "Moneyworld" which includes two ECAs with different roles and we will investigate their effects on the usability and the participants.

This research provides empirical data on the potential effect of humanoid ECAs on serious games, that can be potentially used to inform design decisions regarding spoken ECAs and the design of future MSGs.

Acknowledgements

I would like to express my very great appreciation to my supervisors, Professors Austin Tate and Judy Robertson for their useful feedback and constructive recommendations as well as guidance and valuable support during my Ph.D. Your help and understanding during some difficult times are very much appreciated.

I would like to offer my special thanks to Professor Mervyn Jack for giving me the opportunity to do this research and all the CCIR staff for their help during my time there.

I am particularly grateful for the assistance given by Dr. Simon Doolin for his advice and support and Adam Clayden for the technical aid he provided.

I wish to acknowledge the help provided by Dr. Hazel Morton and Dr. Nancie Gunson during the first pilot study, your input was greatly appreciated.

I would like to thank Lloyds TSB for funding this research.

Advice given by my officemates has been a great help during analysis.

I would also like to thank my parents for their support and encouragement throughout my study. Also, I would like to thank my friends for cheering me on along the way.

Last but not least, I would like to thank my partner John for all his love and support. I couldn't have done this without you.

List of tables

Table 1 Definitions of SGs.....	70
Table 2 Within subject design (repeated measures) based on a 2x2 factorial design.....	114
Table 3 Summary of Numerical Values Assigned to each of the 7-Point Likert Scale Categories.....	118
Table 4-Usability attributes.....	119
Table 5 Summary of Numerical Values Assigned to each of the 5-Point Likert Scale Categories.....	120
Table 6 The API (Agent Persona Instrument).....	122
Table 7 Data categories.....	125
Table 8-Technical information of experimental setup.....	139
Table 9-Experiment versions.....	141
Table 10 Summary table of usability evaluation: Implicit – Explicit feedback.....	142
Table 11-Descriptive statistics.....	146
Table 12-Significant differences of overall means.....	148
Table 13 One-way ANOVA.....	149
Table 14-Descriptive Statistics-Game version by order of experience.....	150
Table 15-Significant differences in individual attributes.....	150
Table 16-Significant differences of overall means based on first experiences.....	153
Table 17 - Game genre preference by gender.....	182
Table 18 - 2x2 factorial design table for the main experiment.....	190
Table 19-Input parameters for power analysis.....	192
Table 20-Output parameters for power analysis.....	192
Table 21 Within subject design (repeated measures).....	193
Table 22 Summary Table of Usability Evaluation: Presence of Humanoid Animated Agents in Mobile Serious Game.....	194
Table 23 Usability attributes.....	198
Table 24 The API (Agent Persona Instrument) attributes (Baylor and Ryu, 2003).....	198
Table 25-Descriptive statistics.....	204
Table 26-Paired samples test	206
Table 27-Cohen's d and omega-squared rules of thumb and reported effect sizes for this experiment.....	207
Table 28 Sample t-test summary after Bonferroni correction.....	210

Table 29-Descriptive statistics for the collaborator agent.	215
Table 30-Descriptive statistics for the instructor agent.....	216
Table 31-Paired samples t-test for collaborator agent version means.	217
Table 32-Paired samples t-test for instructor agent version means.	218
Table 33-Cohen's d and omega-squared rules of thumb and reported effect sizes for the collaborator agent persona.	219
Table 34-Cohen's d and omega-squared rules of thumb and reported effect sizes for the instructor agent persona.	219
Table 35-Mean scores and results of paired t- tests on Individual Agent Persona Instrument for version - Collaborator agent.	221
Table 36-Mean scores and results of paired t-tests on Individual Agent Persona Instrument for version – Instructor agent.	226
Table 37-A priori sample size calculation for regression analysis.	232
Table 38-Descriptive statistics.	234
Table 39-ANOVA for shopkeeper- interaction partner agent.	236
Table 40-Hierarchical Multiple Regression Analyses for the Shopkeeper Agent of the Embodied Conversational Agent Version.	238
Table 41-Descriptive statistics	240
Table 42-ANOVA table for Alex-instructor agent.	242
Table 43-Hierarchical Multiple Regression Analyses for Alex Agent of the Embodied Conversational Agent Version.	244
Table 44-Summary of findings	266
Table 45-One-Sample Kolmogorov-Smirnov Test for instructor agent version means.	348

List of figures

Figure 1-Map of Evaluations	7
Figure 2-Contributing disciplines to present research.....	10
Figure 3 - Evolution of interactional systems. Adapted from Nishida, (2014).	16
Figure 4 Forecast for virtual digital assistants.	19
Figure 5 - Contributing disciplines to the field of ECAs.	21

Figure 6 - An agent in its environment. The agent takes sensory input from the environment and produces as output actions that affect it. The interaction is usually an ongoing, non-terminating one (Wooldridge, 1999).....	23
Figure 7 - Humanoid Embodied Conversational Agents.....	25
Figure 8 - Categories of ECA Design Model (ECADM).....	29
Figure 9 - The Knowledge Navigator was the first appearance of an anthropomorphic embodied conversational agent.	30
Figure 10- Jennifer James auto-sales person.....	32
Figure 11- Mori's axis of uncanny valley.....	51
Figure 12 Spectrum of application interface design in relation to human likeness.	59
Figure 13 - Disciplines contributing to serious games (Adapted from Dörner et al., 2016).	66
Figure 14- Differences between Serious Games and Gamification (Marczewski, 2013).	73
Figure 15 Feedback loop.....	79
Figure 16 Introduction to Moneyworld, Time machine chamber.	105
Figure 17 Corner store layout with shopkeeper ECA.....	106
Figure 18 Alex shown in the virtual portal.	107
Figure 19 Coin Tray.....	108
Figure 20 Coin submission.	109
Figure 21 Usability versus accuracy of speech recognition.....	116
Figure 22 Sample Attitude Statement and 7-point Likert Scale.	117
Figure 23-Screen shot of learn mode feedback.....	138
Figure 24-Screen shot of gaming mode feedback.	139
Figure 25- Estimated marginal means Version by Order.	152
Figure 26- Activities on mobile devices daily.....	183
Figure 27-Spectrum of application interface design in relation to human likeness.....	186
Figure 28-ECA design decisions that result in high human-likeness	187
Figure 29 - Neutral text instructor.....	188
Figure 30 - ECA instructor.	189
Figure 31 ECA collaborator.	189
Figure 32-Error plot.....	211
Figure 33-API Profile – API items for the collaborator agent and the difference between designs.	222
Figure 34-API Profile – API items for instructor agent.	227

Figure 35-Factors and latent factors as presented by the author of the API questionnaire. (Baylor and Ryu, 2003)	231
Figure 36-ECA design decisions that result in high human-likeness and in turn illusion of humanness	289
Figure 37-Scatterplot of the dependent variable and the regression standardised predicted value from the full model (9 independent variables).	360
Figure 38-Scatterplot showing that homoscedasticity has been met.	362
Figure 39-Histogram and normal P-P plots showing the normal distribution of the residuals.	363
Figure 40-Scatterplot of the dependent variable and the regression standardized predicted value from the full model fort Alex- instructor agent (9 independent variables).	365
Figure 41-Histogram and normal P-P plots showing the normal distribution of the residuals.	366
Figure 42-Scatterplot showing that homoscedasticity has been met.	367

List of equations

Equation 1-Cohen's formula for calculating effect size in multiple regression (Selya, et al., 2012).	238
---	-----

Abbreviations

SG: serious games	TTS: text-to-speech
MMI: multimodal mobile interfaces	MSG: mobile SG
UI: user interface	VR: virtual reality
MSG: mobile serious games	AR/MR: augmented reality/mixed reality
GUI: graphical user interface	GBL/DGBL: game-based learning/digital game-based learning
ECA: embodied conversational agent	SGI: serious game initiative
HECA: humanoid embodied conversational agents	MMORPG: massively multiplayer online role-playing games
HCI: human computer interaction	ANOVA: analysis of variance
NLI: natural language interaction	IV/DV: independent variable/depended variable
AI: artificial intelligence	OLS: ordinary least squares
VDA: virtual digital assistants	API: agent persona instrument
ECADM: ECA design model	CASA: computers as social actors
2D/3D: two dimensional/three dimensional	
PA: personal assistant	

Table of Contents

Declaration of originality.....	i
Abstract	iii
Lay summary.....	v
Acknowledgements.....	vii
List of tables.....	ix
List of figures	x
List of equations	xii
Abbreviations	xiii
Chapter 1 Introduction	1
1.1 Introduction	1
1.2 Motivation.....	2
1.3 Objectives	5
1.4 Research questions	6
1.5 Thesis Outline.....	6
1.6 Thesis Contribution.....	10
Chapter 2 Research Background.....	13
2.1 Contemporary conversational systems.....	14
2.1.1 Conversational systems and voice enabled technologies	14
Embodied conversational agents (ECAs) and relevant literature	20
2.2 Embodied conversational agents (ECAs) and relevant literature	20
2.2.1 Embodied conversational agents (ECAs)	20
2.2.2 Definition of ECAs	21
2.2.3 Humanoid embodied conversational agents.....	24
2.2.4 Requirements and characteristics of ECAs	25
2.2.5 Brief history of ECAs	30

2.2.6	Context of use and roles of ECAs	35
2.2.7	Multiple agents	37
2.2.8	Presumed benefits and challenges of ECAs.....	41
2.2.9	Theories on embodied conversational agents related to this research.....	45
	Serious games (SGs) and mobile serious games (MSGs)	63
2.3	Serious games (SGs) and mobile serious games (MSGs)	63
2.3.1	Serious games.....	63
2.3.2	Mobile Serious Games.....	89
	Usability engineering.....	92
2.4	Introduction to Usability engineering	92
2.4.1	Usability and mobile devices.....	93
2.4.2	Usability and serious games.....	96
2.4.3	Usability and ECAs.....	97
2.5	Summary	99
Chapter 3	Methodology.....	103
	Introduction	103
3.1	System description and technology used.....	104
3.1.1	System description.....	104
3.1.2	Technology used.....	110
3.2	Experimental design and experimental procedure.....	111
3.2.1	Experimental design	111
3.2.2	Data collection method.....	112
3.3	Evaluation Metrics	114
3.3.1	Quantitative data collection	114

3.3.2	Qualitative data collection	122
3.4	Statistical Analysis of Experiment Data.....	123
3.4.1	Hypothesis testing	123
3.4.2	Statistical analysis for multiple linear regression	128
3.5	Sample Size Justification.....	129
3.5.1	Sample size for t-test.....	129
3.5.2	Sample size in regression	132
3.5.3	Sample size for technographic survey (Study 2)	132
3.6	Ethical Procedure	132
3.7	Summary.....	133
Chapter 4	Preliminary work	135
4.1	Pilot study 1	135
4.1.1	Introduction.....	135
4.1.2	Experiment Interface Design	137
4.1.3	Experiment Design.....	140
4.1.4	Experiment Procedure	142
4.1.5	Results	145
4.1.6	Discussion and conclusions	161
4.2	Study 2: Survey on the use of Mobile Devices and game playing....	166
4.2.1	Introduction.....	166
4.2.2	Purpose of the research	166
4.2.3	Questionnaire Design	167
4.2.4	Survey Methodology.....	168
4.2.5	Results	169

4.2.6	Discussion.....	181
4.2.7	Summary.....	183
Chapter 5	Main Experiment and Evaluation.....	185
5.1	Introduction	185
5.2	Experimental Interface Design	186
5.2.1	Materials	188
5.3	Experimental Design.....	190
5.3.1	Hypothesis testing	191
5.3.2	Sample size	191
5.3.3	Participants	193
5.3.4	Materials	195
5.4	Experimental Procedure	195
5.4.1	Questionnaires.....	196
5.5	Results.....	201
5.5.1	Quantitative analysis	202
5.5.2	Qualitative analysis	245
5.6	Summary	260
Chapter 6	Discussion and Conclusions - Research Contributions and Design Implications with Respect to Embodied Conversational Agents in Mobile Serious Games	263
6.1	Introduction	263
6.2	Key findings	264
6.3	Preliminary work.....	267
6.4	Main experiment.....	268
6.5	Limitations	282
6.6	Future work and suggestions	285

6.7	Implications for developers.....	286
6.8	Conclusions.....	290
References.....		293
Appendices		329
Appendix A		331
Appendix B		337
Appendix C		341
Main experiment: Usability questionnaire	341	
Assumption testing	341	
Appendix D		345
Main experiment: API questionnaire.....	345	
Descriptive statistics and assumption testing	345	
Appendix E.....		359
Main experiment: Regression analysis	359	
Regression model assessment for collaborator and instructor agents.....	359	
Shopkeeper-collaborator agent.....	359	
Alex- instructor agent.....	364	
Appendix F.....		369
Preliminary work: Pilot study 1.....	369	
Examination of individual attributes	369	
Appendix G		373
Exit questionnaire sample	373	
Appendix H		379
Participation Acceptance Form.....	379	

Chapter 1 Introduction

1.1 Introduction

The aim of this research is to examine how spoken humanoid embodied conversational agents (HECAs) can foster usability in a mobile serious game (MSG) and contribute empirically to the area of conversational agents. By employing usability engineering methods, it is attempted to tackle issues that surround the use of ECAs on an MSG application. In more detail, this multidisciplinary examination illustrates the effect that humanising ECAs has on usability (taking into consideration the roles of the agents) while using speech recognition in order to interact with the ECAs in MSGs. To conduct this research, users' perception of the ECAs are examined through usability and agent persona evaluations to determine whether ECAs are beneficial in this context of use. This chapter introduces the motivation for this research, the thesis overview, the objectives and the significance of this investigation.

Usability is one of the many layers that influence the overall user experience (UX). Usability is concerned with the "effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments" (ISO 9241-11) (ISO, 1998), while user experience is concerned with "all aspects of the user's experience when interacting with the product, service, environment or facility" (ISO 9241-210) (ISO, 1998).

1.2 Motivation

Technological breakthroughs in personal mobile devices have transformed them into computers of exceptional power. The portability and wireless access to the internet makes mobile devices a tool of great potential for formal and informal development. However, there is a lack of studies regarding the use and the effectiveness of mobile devices for this purpose (Sung et al., 2016).

The latest generation of mobile devices has the capabilities of supporting more complex applications in terms of technical and interactive features. However, these applications share the same user interface (UI) principles as the graphical user interface (GUI) first introduced in desktop environments (Doumanis et al., 2015). The multi-touch nature of the mobile interaction along with the smaller screen size and the human fingertip call for a more compact information architecture (IA) with cleaner user interfaces and a smaller number of steps (Doumanis et al., 2015).

The way users interact with mobile devices is changing again since the latest generation of mobile devices includes voice-driven virtual assistants (Siri, Google now, S voice) (Santos-Perez et al., 2013). Even though there are plentiful mobile devices with touch keyboards, text entry is still slow and error prone (Thomas et al., 2015), while hands-free voice control and voice control in general is a reality for the latest generation of smartphones (Motorola, "Hey Siri") and home virtual assistants (Amazon Echo, Google Home). Hands-free interfaces promise greater convenience to the user; a recent study shows that speech recognition is three times faster than texting (Ruan et al., 2016). Voice-controlled intelligent personal assistants such as Amazon Alexa and

Google Home are now gaining momentum, making research in the area even more significant and contemporary. One driver behind the use of speech as an interaction mode is that the user is offered a different channel of communication with a complex system. Spoken dialogue systems (SDS) have other benefits such as hands/eyes free interaction, intuitive interaction and ease of use; therefore, the usability of those systems becomes a significant issue for the success of the interaction (Feng, 2006).

Embodied conversational agents (ECAs) are virtual characters with the ability to converse with a human through verbal (speech) and/or non-verbal communication (text and/or gestures) (Cassell et al., 2006). There are many theoretical advantages in favour of ECAs and spoken dialogue systems (systems that use speech as input) and it is assumed they provide a more "natural interaction" (Weiss et al., 2015, Takeuchi and Naito 1995). A lot of work has been done regarding the interaction between ECAs and users but not so much is dedicated on the usability. Part of the ECA research focuses on users' perception of the ECAs and is regarded to be a very important aspect of the interaction. Increased believability and perceived trustworthiness are a major goal in ECA research. To achieve that, human-like virtual agents are often developed; this human-like aspect makes ECAs subject to social conventions (Gris-Sepulveda, 2015).

Embodied conversational agents are considered due to the linguistic, extra-linguistic and non-verbal information they convey, anthropomorphic entities. The anthropomorphisation of interfaces evokes an illusion of humanness from the user's behalf that can affect the interaction and subsequently the usability.

The social behaviour that virtual agents exhibit along with their presence in the virtual environment can play a motivational role as their expressiveness makes them more engaging to the user (Lester et al., 1997). Previous research has shown that in applications with pedagogical purposes these ECAs can increase the learning effectiveness (Lester et al., 1997) while virtual humans enhance the presentation of information and can significantly provoke learner motivation and retention (Moreno-Ger et al., 2012).

According to recent studies, the interaction with spoken dialogue systems, either in the form of an embodied agent or not, is still inferior compared to other approaches that allow a direct manipulation of the system to which the user responds instinctively, despite the theoretical advances of ECAs and dialogue systems (Weiss et al., 2015).

Although there is a growing pool of empirical data relevant to the effects of ECAs, there is still lack of empirical evaluations suggesting that ECAs are more usable on mobile devices. As the empirical evidence supporting the use of ECAs in mobile devices is limited, especially in the context of SGs, additional research is needed to establish their impact to those applications (Doumanis et al., 2015). Even though most of the literature has focused on the design and implementation of ECAs, there is still lack of empirical evaluation of their effectiveness (Guo et al., 2014).

Given the lack of evidence on the potential effect of ECAs on SGs, there is a major risk related to the introduction of ECAs in SG mobile applications. In fact, the ECA might not enhance the application or might not be appropriate in this context of use. For example, an ECA in a car navigation system might have a negative effect on the effectiveness of the system, while the use of an ECA in a tutoring system could have a positive one. By introducing an ECA

without taking into consideration the context of use and the purpose of the system could lead to a poor performance by the user as the ECA might act as a distraction rather than a helpful element and the interaction may be frustrating for the user (Doumanis et al., 2015). Therefore, whether usability and quality are to be enhanced by using an ECA in a multimodal human-machine interface must be decided for each application anew (Weiss et al., 2015).

Although, there are still a lot of questions surrounding the effectiveness of using ECAs in user interfaces and additional research is needed to evaluate the impact of the combination of games and ECAs, the existing findings reveal their strong potential in provoking enhanced player learning in serious applications (Doumanis et al., 2015).

This is also a strong reason to examine if and how ECAs enhance usability over current interaction paradigms in SG environments, even more so in mobile devices as there is a recent trend towards MSGs (Gamelearn, 2015; Adkins S., 2015).

1.3 Objectives

The objectives of this thesis are summarised below:

- Examine the impact on usability of a humanoid ECA (HECA) to a mobile serious game (MSG).
- Examine the extent to which the presence of a humanoid ECA (HECA) affects the quality of the interaction for the given domain and task.

- Identify which attributes of the humanoid ECAs (HECA) contribute to the overall usability, and in what way.
- Examine the effect that ECAs with different roles have on users' perception of usability.
- Explain the results obtained in terms of relevant theories, particularly the "illusion of humanness".

1.4 Research questions

The research questions which this study attempts to answer are:

R1: To what extent do HECA affect the usability of a mobile serious game (MSG)?

R2: To what extent do users perceive a difference in agent persona between HECA and neutral text presentation as measured by the agent persona instrument (API)?

R3: Which factors relating to the HECA's persona attributes account for variability in usability, and to what extent?

1.5 Thesis Outline

The rest of this thesis is organised as follows. The second chapter includes the literature review on ECAs, speech recognition systems, SGs, usability and the technology used in this research. The chapter offers a broad review of literature on spoken dialogue systems and mobile speech recognition software, a historical analysis and an overview of research into the use of

ECAs in user interfaces, SGs and mobile devices, as well as research into SGs for mobile devices. Previous research on user evaluations regarding MSGs, mobile ECAs and mobile speech recognition systems is also addressed. The chapter introduces a research background to facilitate an understanding of how human-to-human interaction behaviours migrate to human-to-agent interaction in a MSG. From this literature review, the need for empirical evaluations of humanoid ECAs for MSGs is confirmed.

The thesis describes two distinctive empirical evaluations that progressively assess the presentation and usability of ECAs (Figure 1). Chapter 3 introduces the principles of usability engineering and experimental methods used to conduct the empirical evaluations. This chapter provides a detailed discussion on usability evaluation methods, experimental design, regression analysis and hypothesis testing, followed by the evaluation metrics used throughout this research, test procedures, ethical issues in experiments involving humans and the data retrieval and statistical analysis methods used.

Series of evaluations

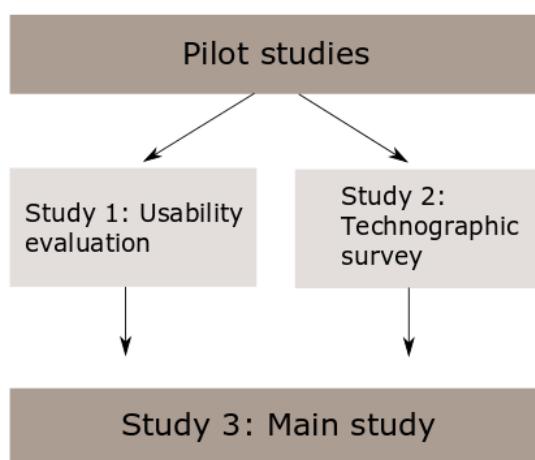


Figure 1-Map of Evaluations

The following two chapters describe the two empirical evaluations along with a large-scale pre-test technographic¹ survey. The aim of the first two evaluations are to test the effectiveness of the application and assess the user acceptability of the gamification element and agent representation in a SG. The final experiment then concentrates on the main aim of this thesis, which is to contribute to the research field of ECAs, and to provide insight on how ECAs can be integrated in an effective and usable way with a focus on MSGs involving speech recognition technology.

Chapter 4 presents the rationale, aims and findings of the pilot experiment (Study 1) and the technographic survey (Study 2). Study one is developed as a means of testing the Unity-based application along with the Pocket Sphinx speech recognition engine. Also, this evaluation acts as a methodological sand box which helps determine the methodology approach adopted for the main experiment while establishing that a SG is a suitable environment for the main experiment along the way. The experiment is designed as a simple double cell within subjects' design. With this study, it is assessed how a SG affects usability. Two versions of an application are compared. One is a SG providing implicit feedback and the other is not gamified, providing explicit feedback. While the results of the comparison do not provide a statistically significant difference among the versions, in the user stated preferences there is significant preference towards the gamified version.

Chapter 4 also reports a study (Study 2) that looks at the demographics, motivations and derived experiences of users through online survey data that are compiled from users over a one-year period. This is an important step as Michael and Chen (2006) highlight from a design and development

¹ Technographic data show the hardware and software technologies used by a population.

perspective that the SG market, unlike the entertainment industry, has outdated and less optimal hardware. Also, this market includes gamers of all levels, from experts to first-time players and the games must therefore be even more accessible (Susi, 2007). Identifying which are the devices that people use the most along with their technology and game-playing habits can help testing the software in a relevant to today's user platform.

This study collects data on the use of mobile devices and game playing. This chapter includes the questionnaire design, the quantitative results and discussion of the results.

In Chapter 5 the rationale, aims and findings of the main experiment are given. The experimental design is altered from the evaluation of the previous two studies in order to provide more robust data. Two versions of an MSG are compared (neutral text conversational agent vs HECA). Two agent questionnaires are analysed; one for each agent in the interaction (instructor agent, collaborator agent). Also, qualitative analysis regarding the two agents and the use of speech recognition software is reported. The main experiment examines whether the illusion of humanness influences the overall usability of an MSG application. Another aspect that is examined through this experiment is the effect that agents with distinct roles have on usability and identify which aspects of these agents contribute more to the overall usability. The results of the experiment show that users do exhibit a statistically significant preference for the HECA version where the effect size according to Cohen's thresholds is large thus indicating a real-life difference.

In Chapter 6, a summary of the main findings and the conclusions is provided based on the empirical evidence presented in the thesis. The interface design

implications are also presented in this chapter. Finally, the chapter offers suggestions of future work that arise from the research presented.

1.6 Thesis Contribution

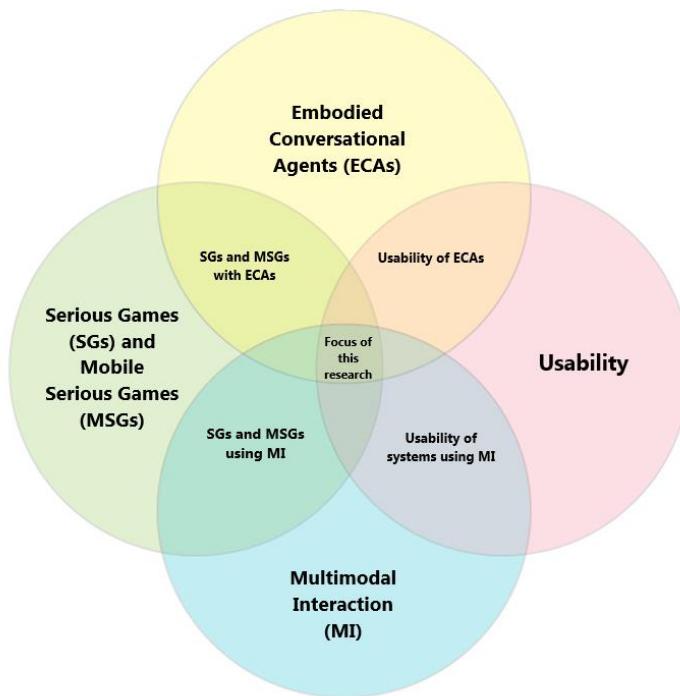


Figure 2-Contributing disciplines to present research

This research makes three contributions to knowledge in the area of ECAs and SGs/MSGs.

The first contribution is a set of design guidelines called the “Embodied conversational agent design model” (ECADM) that can be used to support the design and development of ECAs.

The second contribution is the introduction of the term "illusion of humanness" as the users' involuntary mental response that the interface possesses human attributes and/or cognitive functions. The illusion of humanness has been in the focus of the main experiment where it is confirmed that users respond to the interface socially and involuntary attribute cognitive functions due to the high human-likeness of the interface which in turn affects usability in a positive way.

The third contribution is adding to the empirical evaluations on the effectiveness of ECAs in SGs and MSGs that are limited in the literature. A mixed-method approach of both quantitative and qualitative methods is adopted to investigate the overall effectiveness and usability and provide insight into adults' perceptions and attitudes of using the system. The results of the evaluations can be used to inform the design decisions and the development of effective and efficient SGs, and MSGs grounded in interdisciplinary literature on ECAs, SGs, multimodal interaction and usability as seen in Figure 2.

Chapter 2 Research Background

A review of a wide range of literature with respect to the main research topics is addressed in Chapter 2. The chapter offers a broad review of literature on conversational interfaces, spoken dialogue systems, multimodal interaction, a historical analysis and an overview of research into the use of embodied conversational agents (ECAs) in user interfaces and serious games. It also covers research into serious games for mobile devices and a review of commercial applications using speech recognition and embodied or disembodied conversational agents. Previous research on user evaluations regarding mobile serious games, mobile ECAs and mobile speech recognition systems is also addressed. The chapter introduces background research to assist with the understanding of how human-to-human interaction behaviours migrate to human-to-agent interaction in a mobile serious game. From this literature review, the need for empirical evaluations of humanoid ECAs (HECAs) for mobile serious games is confirmed.

2.1 Contemporary conversational systems

2.1.1 Conversational systems and voice enabled technologies

2.1.1.1 Introduction to conversational systems

There are two information processors in human computer interaction (HCI) – a computer and a human – that try to communicate with each other through a restricted interface. Studying the design of the interface is important to overcome its limitations (Perez-Martin and Pascual-Nieto, 2011).

Although command-line interfaces can be sufficient for an expert user, the rest of the users have in the past been limited to the use of graphical user interface elements. Any option, even a valid one for the application, which is not available in the menu is ignored by the users. Natural Language Interaction (NLI) could be a solution that can improve the communication between a computer and a human (Perez-Martin and Pascual-Nieto, 2011).

People can communicate with the computer by using NLI with their own language which may also be a more natural mode to them (Flanagan, 1995; Perez-Martin and Pascual-Nieto, 2011). Dialog and conversation are embedded in the human psyche which makes conversational systems quite appealing.

Conversational computers were a dream of futurists from the beginning of the computing era. A testament of this is the Turing Test of computational intelligence that imagined a computer that could converse in fluent English and is indistinguishable from a human (Cohen and Oviatt, 1995).

First, a definition of dialog systems is in order. Dialog systems or conversational agents or conversational systems are according to Gulz et al., (2011) "Computer systems that interact with a user using spoken or written language, and possibly other modalities (or even a combination of them), in a connected dialogue consisting of several turns".

According to Jurafsky and Martin, (2017), those systems generally fall into two classes: task oriented dialog agents and chatbots.

The task-oriented dialog agents are designed for a specific task and are set up for short conversation (for a single to perhaps a few dozen interactions) in order to get information from the user to complete the task. They are based on a domain ontology in which the ontology defines one or multiple frames with each frame being a collection of slots and defines the values that each slot can take. This frame-based architecture was introduced for the first time in the GUS system for travel planning (Bobrow et al., 1977) (as cited in Jurafsky and Martin, 2017) and is the base for most modern digital assistants. The ECAs of this research and the digital assistants both on mobile devices (Siri, S voice, Google now etc.) as well as in smart speakers (Amazon Alexa, Google Home etc.) belong in this category. The commercial digital assistants can control home appliances, note appointments on a calendar, give directions and information, send texts and make calls. Conversational agents have an important role in the human-robot interaction and companies deploy them on their websites for customer service (Jurafsky and Martin, 2017).

The second class of dialogue system is chatbots. Chatbots are systems that can handle more extensive conversations that aim to mimic the unstructured nature of conversations in human-to-human interaction. These systems can

be designed for entertainment (Microsoft's 'XioIce') or more practical purposes. In fact, the very first chatbot ELIZA (Weizanbaum, 1966) was purposed for psychological counseling. There are two classes for chatbot architectures, the rule-based systems and the corpus-based systems (Jurafsky & Martin, 2017).

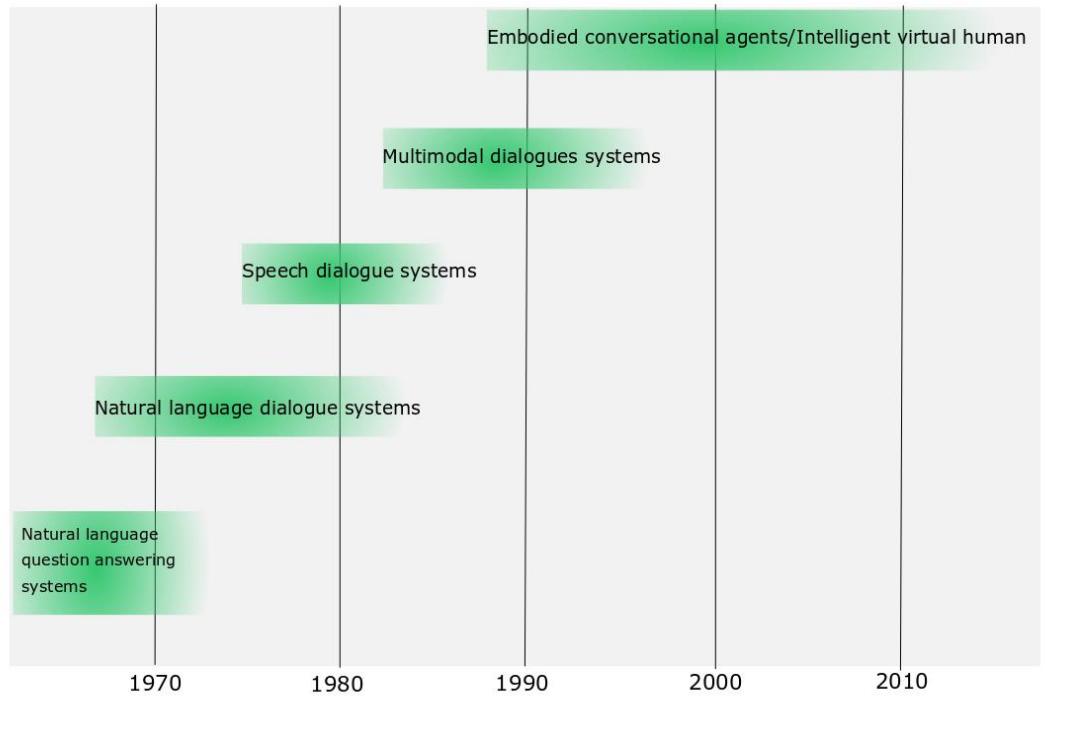


Figure 3 - Evolution of interactional systems. Adapted from Nishida, (2014).

As seen in Figure 3, the early natural language dialogue systems were text based question answering machines such as Baseball (Green et al., 1961), the aforementioned ELIZA (Weizanbaum, 1966), LUNAR (Woods, 1973) and SHRDLU (Winograd, 1972) (Nishida et al., 2014). These systems transformed the user input into database queries which were then used to answer the question. Even though those systems could handle only simple sentences, they were very impressive at that time. In the pursuit of better HCI, artificial

intelligence (AI) researchers extended them as interactional systems and in 1980s speech recognition systems such as HEARSAY-II (Erman et al., 1980) were developed. Also in the 1980s, speech recognition systems extended to multimodal interfaces an example of which is Put-That-There (Bolt, 1980). A concept video with the title "The Knowledge Navigator" was released by Apple Inc. in 1987. The video showed how an artificial intelligence system could help people via an embodied conversational agent. That inspired researchers to build agents that bore similar characteristics to the agent in the video such as anthropomorphism and verbal-non verbal interaction which marked the beginning of the field of ECAs and intelligent virtual agents (Cassell et al., 2000; Nishida et al., 2014). More on ECAs can be found in the section Embodied Conversational Agents of the Background chapter.

2.1.1.2 Contemporary conversational systems and voice enabled technologies

Using speech as an interaction modality with machines has been proposed long ago. The reason is that speech is used by humans as the main way of communication (Weiss et al., 2015); an inherent people's ability to listen and speak. Thus, this modality emulates human-to-human interaction (Yankelovich et al., 2007).

Previous work has shown that the interaction with spoken dialogue systems, either in the form of an embodied agent or not, is still inferior to other approaches that allow a direct manipulation, despite the theoretical advances of ECAs and dialogue systems (Weiss et al., 2015). The reasons for the reluctance of using ECAs in multimodal HCI are multi-fold. On the human side of interaction, when communicating audio-visually via speech people convey extra-linguistic information instinctively. In HCI this information that

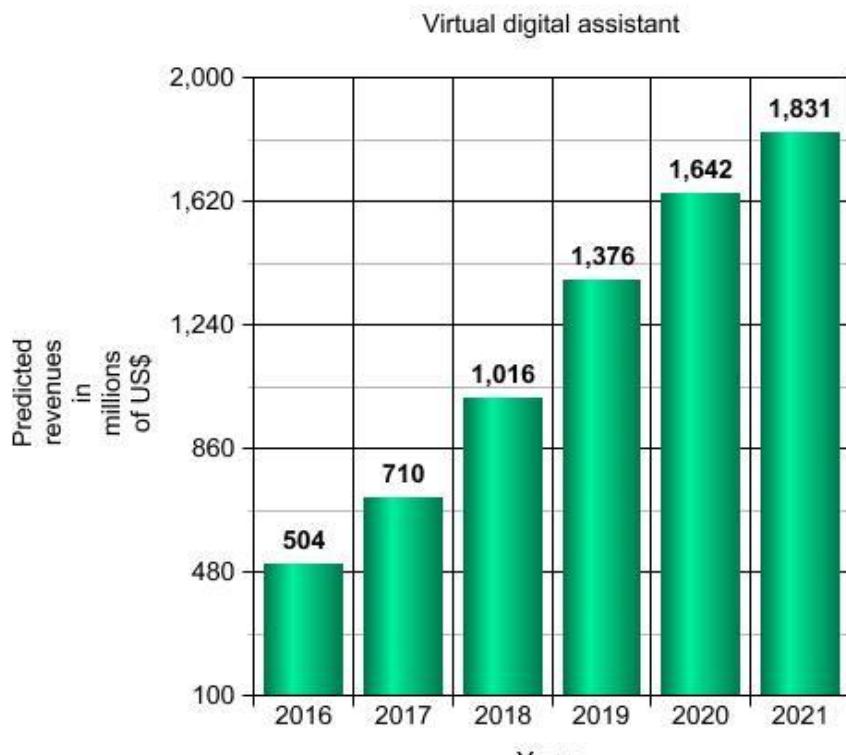
works complimentary to speech is not reliably communicated since machines are usually not able to interpret and extract this type of information. On the machine side of interaction, current technology is still somewhat limited on generating extra-linguistic information the way humans do. Furthermore, such attempts are not always interpreted correctly by users (Weiss et al., 2015).

However, voice enabled technologies are gaining momentum in the digital ecosystem. Recent technological progress has shown that speech recognition is working better than before due to deep learning and big data that train the deep neural networks (Landay, 2016). Advances in the contributing fields of automatic speech recognition and synthesis, artificial intelligence, computational linguistics and machine learning improved significantly the capabilities of conversational systems. These improvements resulted in a booming market of voice enabled systems such as Google Home, Amazon Echo and Apple Home Pod (Hamilton, 2017).

Voice enabled systems are ideal for eyes-free and hands-free interaction such as driving or cooking.

According to Tractica², use of virtual digital assistants (VDAs) (mobile and stationary), that in their majority use speech recognition, is going to increase exponentially in the next years. The forecast shows that the unique active consumer VDA users will grow from 390 million in 2015 to 1.8 billion worldwide by the end of 2021 and the predicted revenues from over half a million in 2016 to 1.8 billion by 2021 as shown in Figure 4.

² <https://www.tractica.com/newsroom/press-releases/the-virtual-digital-assistant-market-will-reach-15-8-billion-worldwide-by-2021/>



Source: Tractica

Figure 4 Forecast for virtual digital assistants.³

2.1.1.3 Voice assistants on mobile devices

Even though today's mobile devices have evolved tremendously compared to their early ancestors, they still have some challenges to overcome. The human fingertip along with the limited screen size, makes the interaction more challenging. This limitation has been addressed by several mobile systems that provide aggregated information, presented through a single medium and requires minimum user intervention. Consequently, the way users interact with mobile devices is called to change since the latest generation of

³ Source: Tractica

mobile devices includes voice driven virtual assistants (Siri, Google Now, S voice) (Doumanis and Smith, 2015).

A study by Stanford University has shown than when comparing speech and keyboard text entry for short messages in two languages (English and Mandarin Chinese) on touchscreen phones, speech recognition had an input rate of 2.93 times faster (153 vs. 52 WPM) for English and 2.87 times faster (123 vs. 43 WPM) for Mandarin Chinese than the keyboard (Ruan et al., 2017).

Voice assistants are common in premium tier smartphones (over US\$300) with 97% of them sold worldwide in 2017 having one out of the box. Voice assistants will soon be integrated to lower priced smartphones (over US\$100) with an estimate that 80% of all smartphones will have a voice assistant integrated natively in 2020 (Hyers and Mawston, 2017).

This marks a new era for human-smartphone interaction with voice input becoming a common mode.

Embodied conversational agents (ECAs) and relevant literature

2.2 Embodied conversational agents (ECAs) and relevant literature

2.2.1 Embodied conversational agents (ECAs)

As discussed in 2.1, researchers have been dreaming of a computer with conversational capabilities for decades. Many researchers believe that the

mainstream of conversational system development is a path toward ECAs or intelligent virtual humans (Nishida et al., 2014). Embodied conversational agents in their present form are the result of many contributing disciplines. Those disciplines differ for each ECA depending on its capabilities and modes of interaction. For example, an ECA that uses speech input needs speech recognition and speech-to-text technology, while ECAs with text input do not. In general, ECAs are by their nature multidisciplinary as shown in Figure 5.

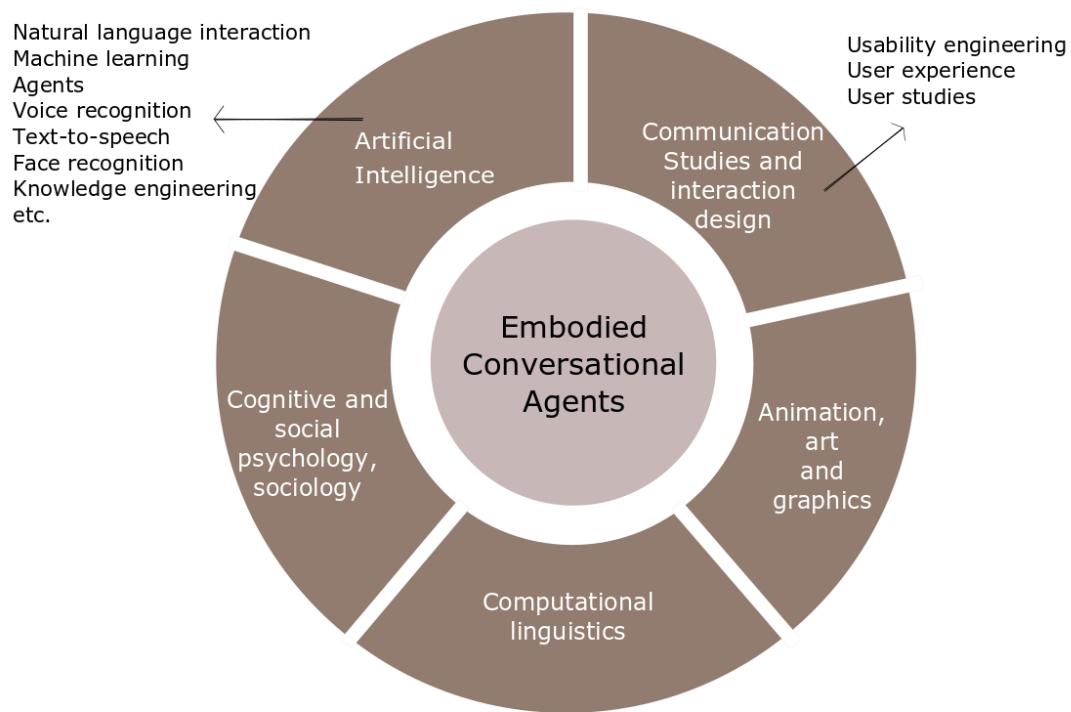


Figure 5 - Contributing disciplines to the field of ECAs.

2.2.2 Definition of ECAs

The field of embodied conversational agents is multidisciplinary and constitutes a subcategory of conversational agents.

The term "Embodied Conversational Agent" was coined by Justine Cassell in 2000 and is defined as follows: "computer interfaces that can hold up their end of the conversation, interfaces that realise conversational behaviours as a function of the demands of dialogue and as a function of emotion, personality, and social conversation" (Cassell et al., 2000).

Also, according to Cassell, embodied conversational agents (ECAs) are virtual characters with the ability to converse with a human through verbal (speech) and/or non-verbal communication (text and/or gestures). The main difference of ECAs compared to other artificial intelligence (AI) entities is the human abilities they share. Examples of such abilities are the recognition and response to verbal and non-verbal input; the generation of verbal and non-verbal output; the relation to conversational functions and provision of signals that indicate the conversation state; and the contribution of new propositions to the discourse (Cassell et al., 2001).

Embodied conversational agents are comprised of three elements, the embodiment, the conversation and the agent. Breaking down the term and analysing the components will provide a better understanding of what ECAs are or can be.

Embodiment: The term embodiment with a broad meaning is used to describe all the low-level aspects that contribute to the physical⁴ appearance of the agent. Those aspects include the head, the design of the agent, the rendering of the agent, the animation (hand gestures and facial expressions) and the quality of the corresponding motions (gesture and lip synching) (Ruttkay et al., 2004).

⁴ According to the Cambridge dictionary physical means *relating to the body*, which in this case refers to the virtual body.

Conversation: According to the Cambridge Dictionary⁵, conversation is described as “a talk between two or more people in which thoughts, feelings and ideas are expressed, questions are asked and answered, or news and information is exchanged”. In the case of ECAs, the conversational aspect is fulfilled by the communication between the user and the ECA using verbal and/or non-verbal modalities.

Agent: Wooldridge (1999) defines agents as: “a computer system that is situated in an environment and is capable of autonomous action in this environment in order to meet its design objectives” (see Figure 6).

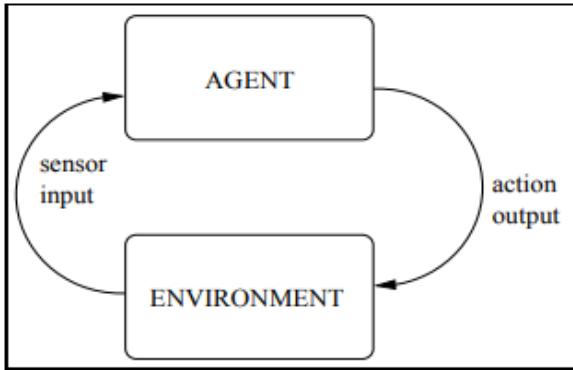


Figure 6 - An agent in its environment. The agent takes sensory input from the environment and produces as output actions that affect it. The interaction is usually an ongoing, non-terminating one (Wooldridge, 1999).

The term “agent” itself may have a dual meaning that can be conflated (Erickson, 1997);

- 1) the agent-metaphor, which refers to the presentation of the character on-screen and

⁵ As found in Cambridge Dictionary:
<https://dictionary.cambridge.org/dictionary/english/conversation>

2) AI aspects in software that are not always visible on screen, e.g. intelligence, additivity, responsiveness.

2.2.3 Humanoid embodied conversational agents

Interface agents such as ECAs are agents that have some form of a graphical/visual representation on the interface and are capable of autonomous actions without explicit directions from the user (Doumanis and Smith, 2015). In their broader meaning, that of humanlike conversational AI, ECAs have 161 synonyms.⁶ The terms mostly used interchangeably with ECAs are: virtual character, intelligent agent or social agent (Veletsianos and Miller, 2008).

Another term related to ECAs is that of virtual humans. Virtual humans are the result of the emergence of different fields around computer science such as artificial intelligence, computer animation, computer graphics, human-computer interaction and cognitive science (Kasap and Magnenat-Thalmann, 2008). These characters can play the role of the guide, the trainer, the teammate, the rival or a source of motion in virtual space (Brogan et al., 1998). However, virtual humans along with their complexity, can vary diametrically as each of them has a specific role and purpose depending on the goal of the application. The main difference between ECAs and virtual humans is that virtual humans always have the appearance of a human and they do not necessarily possess any intelligence or communication skills. An example is the non-interactive characters in games that are used to populate a scene. When virtual humans are combined with ECAs, the result is a HECA (Figure 7). In this context, the word *humanoid* is used with the definition

⁶ As found in: <https://www.chatbots.org/synonyms/>*

given by the Cambridge Dictionary⁷ as "a machine or creature with the appearance and qualities of a human".

Virtual Human + Embodied Conversational Agent (ECA) = Humanoid Embodied Conversational Agent (HECA)

Figure 7 - Humanoid Embodied Conversational Agents

2.2.4 Requirements and characteristics of ECAs

Embodied conversational agents need to possess the following abilities which complies to the modelling of regular autonomous agents. First, they should perceive verbal and/or nonverbal input from the user and the user's environment. Second, they should translate the inputs' meaning and respond appropriately through verbal and/or nonverbal actions. Last, those actions should be performed by an animated computer character in a virtual environment (Huang, 2018).

According to De Vos, (2002), ECAs share the following five features:

1. Anthropomorphic appearance

Anthropomorphism is defined by the Oxford Dictionary⁸ as "*The attribution of human characteristics or behaviour to a god, animal or object.*".

⁷ As shown in Cambridge Dictionary:
<https://dictionary.cambridge.org/dictionary/english/humanoid>

⁸ As found at Oxford Dictionary:
<https://en.oxforddictionaries.com/definition/anthropomorphism>

In the case of ECAs, the agent is visually represented by some form of anthropomorphic embodiment which can be either an animal, a human or a fantasy figure. The rendering of this visual manifestation of the ECA can either be animated or static, 2D or 3D, photorealistic or stylised or any other form that can convey conversational functions. More on anthropomorphism as a concept in section 2.2.9.

2. Virtual body that is used for communication purposes

Embodied conversational agents should be able to use their embodiment to either communicate messages or enhance the communication through other modes of interaction. This is called non-verbal communication and can be achieved using body posture, body movements, facial expressions, gestures etc.

3. Natural communication protocols

Usually ECAs use different communication protocols than those of classic HCI which rely more to menus and buttons. These protocols are based on human-to-human interaction, and ECAs use NLI as a natural mode of interaction to better emulate human-to-human interaction.

4. Multimodality

In its basic form, multimodality is an interdisciplinary approach based on social semiotics and communication that does not rely merely on the language.

Embodied conversational agents should be able to communicate through various channels that are typically used in face-to-face interaction such as gestures, speech and other modes of interaction.

According to Wik (2011), "Task-based, interactive exercises and the use of sound, pictures, agents and games, will not only enrich learning by making it a more worthwhile experience to learn. By presenting content to be learned in a rich multimodal environment, a more robust memory trace is also created and thus the retention will be increased. Motivational and cognitive factors may hence fuse during learning activities and influence the outcome of the skill building."

Also, interaction with multimodal interfaces has been encouraged by the fact that those systems present increased human-likeness which has been shown to support cognitive functions such as learning and information comprehensiveness (Dehn and Van Mulken, 2000).

5. Social role

Embodied conversational agents are different from other computer systems in the sense that they try to emulate human-to-human interaction in a believable manner and, therefore, have a social standing. The concept of believability is described by Bates, (1994) as "one that provides the illusion of life, and thus permits the audience 's suspension of disbelief". In ECA research, the concept of believability is approached in two ways. One way is that higher believability can result by implementing more NL functions (Cassell and Stone, 1999). The other way is that believability is more a matter of personality and emotions supported by the significant roles that portayal of emotions plays in creating "believable" characters by Disney (Bates, 1994). The work presented in this thesis uses ECAs that express personality and emotions as part of the " illusion of humanness " of the system (more in section 2.2.9).

A model that organises ECAs' characteristics into three categories, is proposed in this thesis called the ECA Design Model (ECADM) and is given in Figure 8. Firstly, on the presentation level, ECAs can be depicted as either human or non-human characters, animated or static, photorealistic or more stylised, 2D or 3D, they can have a full body, only a head, a bust or a torso and finally their physical properties can vary (hair colour, clothes, body type, accessories, age etc.) (Haake and Gulz, 2009; Gulz and Haake, 2006; Veletsianos and Miller, 2008; Clarebout and Heidig (née Domagk), 2012). Secondly on the interaction level, decisions on the input and output modalities of the ECA must be taken. Multimodality is a basic feature of ECAs; this means that ECAs can employ one or more of the inputs and output modalities such as voice and text.

Finally, the persona level of the ECA is constituted by features related to the perceived by the user character of the ECA. Just like in real life as well as with virtual assistants, voice plays a major role in forming opinions about someone's personality. The agent's voice along with their role in the application and the personality they adopt form a cluster of personality pointers. These personality pointers are also informed by non-verbal and extra-linguistic information.

Those categories are general and can be broken down to specifics, for example under the Interaction level one may add the number of agents within the application.

The model serves a dual function: 1) inform design decisions for designers and 2) act as a guide to categorise ECA research which will allow for better comparisons and analyses.

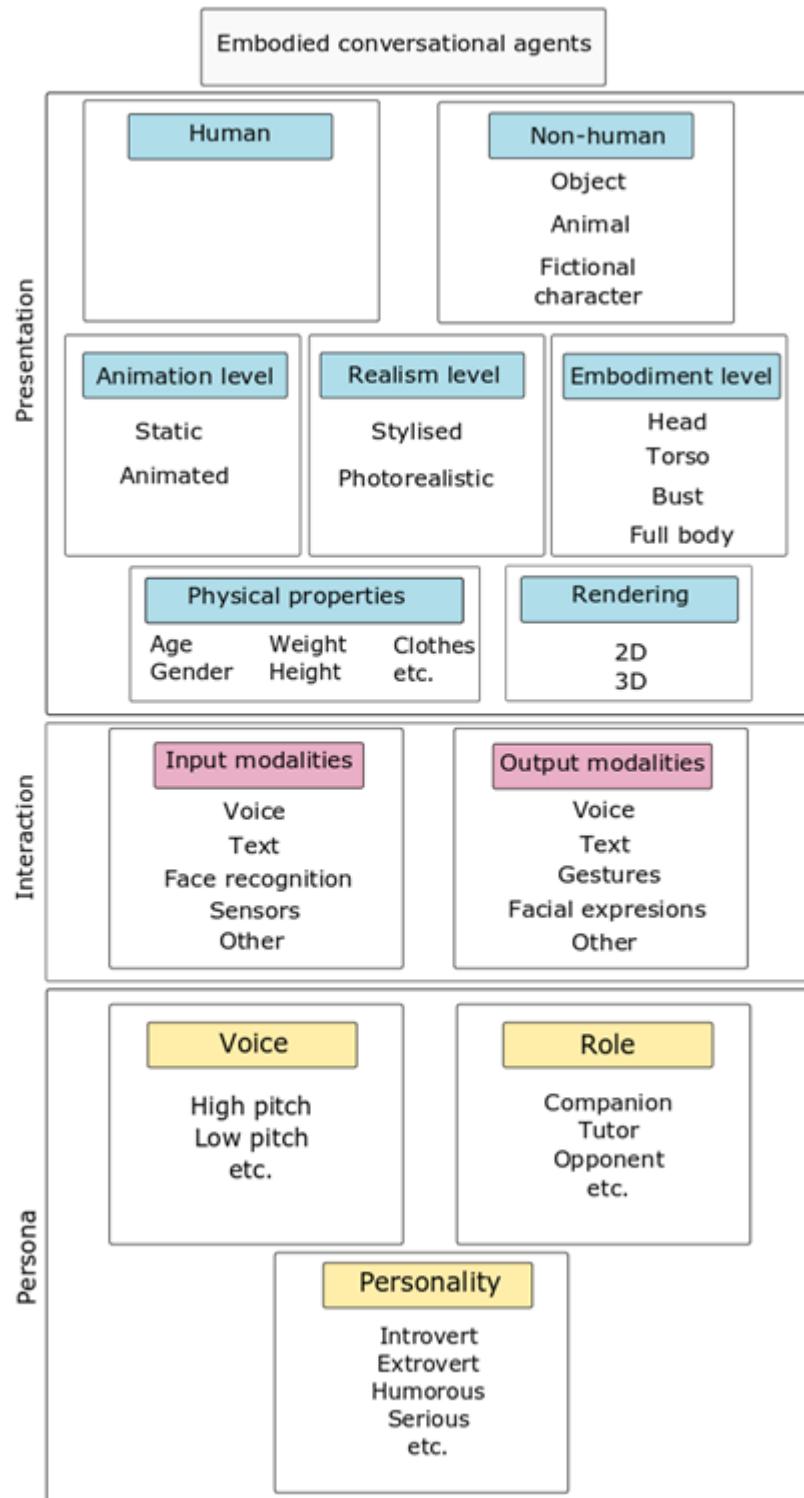


Figure 8 - Categories of ECA Design Model (ECADM).

2.2.5 Brief history of ECAs

One might argue that embodied interfaces have existed as a general concept for hundreds of years. Examples include non-electronic moving machines such as automata and the first visual personification of a talking machine which is a static face on the side of Euphonia, the 1830's Faber's Talking Machine (McBreen, 2002). Those systems were embodied but not interactive.

The evolution of ECAs follows that of conversational systems discussed in 2.1., but a milestone to what is known today as ECAs was a concept video by Apple Inc. in 1987 titled "The Knowledge Navigator" as shown in Figure 9. The Knowledge Navigator demonstrated how the future technology could look, but at that time computers did not have the capabilities to support such interaction.

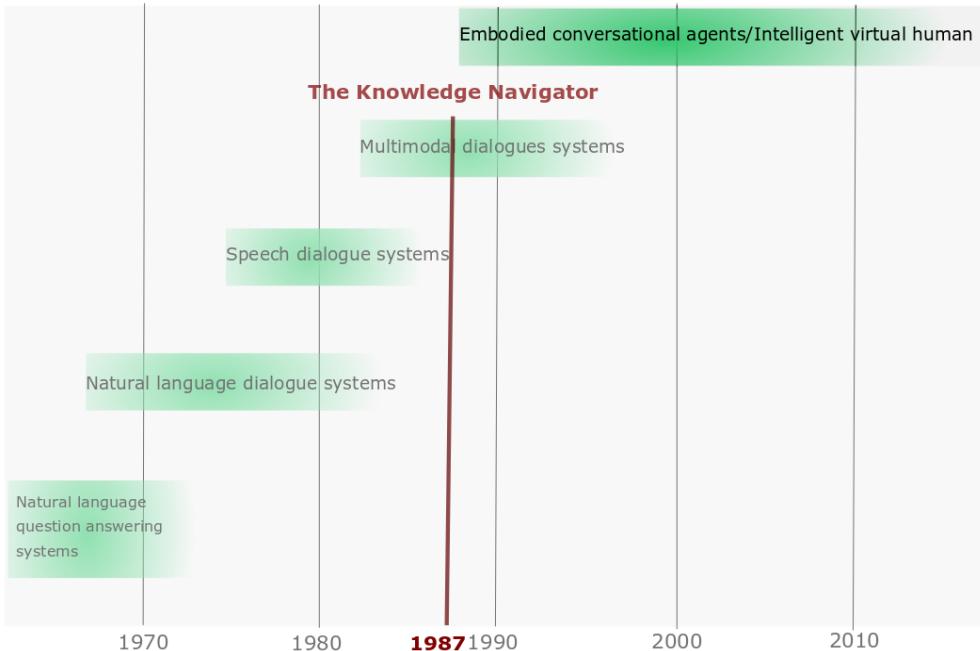


Figure 9 - The Knowledge Navigator was the first appearance of an anthropomorphic embodied conversational agent.

In this video, an anthropomorphic intelligent agent with human-like characteristics under the name Phil is shown. Phil was presented as an animated talking bust of a male thus allowing the user to identify the agent as a character. Phil could also identify the user, allowing "him" to provide a customised experience. This agent symbolically embodied an intelligent agent, capable of social interactions that can use multimodal communication. This concept video was a milestone on how computers can be used as social agents (Nishida et al., 2014).

One of the first embodied multimodal interfaces was Peedy. Peedy was the result of The Persona project at Microsoft Research that started in 1992. The Persona project aimed at the development of a lifelike computer assistant with spoken language input and expressive visual presence. The user could interact verbally with Peedy (a 3D parrot), who acted as a music assistant. Peedy could listen to requests in English and respond verbally or with gestures (Ball et al., 1997).

Another example of early ECA applications is Olga. Olga was a 3D HECA that provided users with consumer information about microwave ovens. The novelty of this system compared to similar ones was that it combined verbal input, 3D animated gestures and facial expressions, lip-synchronisation during speech synthesis and direct manipulation of the system through a graphical interface (Beskow and McGlashan, 1997).

Thorisson (1996) researched multimodal interaction with a 2D animated conversational agent by the name Gandalf. Gandalf was a pedagogical conversational agent who was discussing the solar system. Gandalf could process input from many sources, such as speech and gaze, to model mainly

the social aspects of multimodal dialogue interaction (Cassell, et al., 1999).

The novelty of this research was that it demonstrated that the ECA could engage in conversation effectively by displaying realistic verbal and non-verbal behaviour. Also, Thorisson successfully demonstrated the importance of discourse structure when Gandalf's facial expressions were used appropriately for the content of the conversation.

Hayes-Roth, (1998) also developed an ECA named Jennifer James, as shown in Figure 10 Jennifer James was a 3D ECA presented as an ex-NASCAR driver who acted as a salesperson in a virtual auto show and was engaging in five to ten-minute conversations with the user. Jennifer could engage in free dialogue with the user. During the interaction, she used multimodal communication such as dialogue, facial expressions and animated gestures.



Figure 10- Jennifer James auto-sales person.

One of the most well-known ECAs is Rea (Cassell, et al., 1999). Rea is a HECA and her name stands for Real Estate Agent. Rea could interact with the user

multimodally; she had an articulated 3D body, she was able to sense the user through verbal input and cameras and communicate through intonated verbal output, gestures and face expressions. The novelty of this system was the architecture used that allowed Rea to conduct mixed initiative conversation; hence, she could be interrupted and take turns with the user, as could happen in a human-to-human interaction, allowing for a more natural interaction. Also, the significance of Rea lies in the fact that the system implemented conversational interactions as opposed to using conversation as an interface metaphor. (Nishida et al., 2014).

One agent that was designed to explore task-oriented collaboration in virtual worlds was Steve. Steve was also one of the first conversational pedagogical agents. This agent was able to teach students how to operate naval ship specific machinery but was not domain limited. Steve also used verbal and non-verbal modes of interaction to collaborate with the user and the user could communicate with him through speech recognition (Rickel and Johnson, 2000).

MACK was another project by Cassell et al., (2002). MACK stands for Media lab Autonomous Conversational Kiosk and it was an embodied conversational kiosk that built upon the research on ECAs and information displays in mixed reality with the purpose of showcasing spatial intelligence. MACK was also capable of multimodal input and output (speech, gestures, intonation and gaze) with a visible torso able to use arms and hand gestures to indicate spatiality (Cassell et al., 2002).

GRETA was also a HECA capable of multimodal interaction. It was developed by Poggi et al. (2005) in the context of the EU project MagiCster, and they tackle the issue of believability of ECAs' communicative behaviours. According

to Poggi et al., the multimodal signals that are displayed by the ECA are determined by a series of aspects such as the culture, style, content of discussion, emotions, user sensitivity, context and personality. A dynamic combination of the above was used to determine how and what the ECA would say. GRETA was a 3D head and one of the first affective ECAs with her own personality and social role capable of expressing emotions that are consistent with the context of the conversation.

Max (Multimodal Assembly eXpert) was an affective HECA developed by Bielefeld University (Kopp et al., 2003). Max was able to demonstrate assembly procedures to the user in a CAVE-like virtual-reality environment. Max was fully embodied, has almost a human size and was capable of multimodal interaction (synthetic speech, gestures, facial expressions and gaze). Max has been adopted in various other roles such as an assistant (Kopp et al., 2003) and science museum guide (Kopp et al., 2005).

Earlier research on ECAs focused more on their fundamental functions such as face animation and dialogue processing. In more recent years, due to the advancements on computer graphics and mobile devices with incorporated sensors, the focus shifted to human-agent interaction in a deeper level of communication and affective agents (Huang, 2018). Some of those projects are MARC, a multimodal affective and reactive ECA (Courgeon, 2008) and TARDIS (Training young Adult's Regulation of emotions and Development of social Interaction Skills), an EU project that employs an ECA to help users train for job interviews (Ilagan, 2014).

2.2.6 Context of use and roles of ECAs

Embodied conversational agents have been used in a wide variety of roles from tutoring to sales. A look at the past use of ECAs shows the capabilities of such systems (Cassell et al., 2002).

Previous research has identified the possibilities of using ECAs for e-commerce and finance (Collin et al., 2004; Matthews et al., 2008; Foo et al., 2008), as sales agents (Andre et al., 2000; Cassell et al., 1999; Hayes-Roth, 1998), as TV style presenters (Noma et al., 2002), as medical advisers (Pelachaud et al., 2002; Poggi et al., 2005), as companions (Cavazza et al., 2010), as museum guides and in installations (Kopp et al., 2008; Kopp, et al., 2005; Bickmore, et al., 2013), for mission rehearsal training (Hill et al., 2003), for military leadership and cultural training (McCollum et al., 2004; Raybourn et al., 2005), for psychological support (Hayes-Roth et al., 2004) and in various other roles.

When used in an educational context, ECAs have served as animated pedagogical agents (more in section 2.2.6.1) in a variety of roles (Andre et al., 2000; Lester et al., 1997; Moreno, et al., 2001; Moundridou and Virvou, 2002). According to Kim and Baylor, (2008), pedagogical agents have been identified in four major roles: 1) an expert who provides information, 2) a mentor who advises, 3) a motivator who encourages, and 4) a companion who collaborates.

Embodied conversational agents have become a key element within the thriving sector of the video game industry as well, as they enhance storytelling and create more immersive experiences for the player. Thus, this market focuses on the improvement of interactive non-player characters (Gris

Sepulveda, 2015). It has been suggested that ECAs are well suited within entertainment applications as they can be natural interaction partners and can act in a socially appropriate way (Rhem and Wissner, 2005).

2.2.6.1 **Pedagogical agents**

A rather interesting subcategory of ECAs is pedagogical agents (PAs). Even though pedagogical agents are not necessarily animated and can also be presented as static images or videos of human tutors (Clarebout and Heidig (née Domagk), 2012), this review focuses on the PAs that possess similar characteristics with ECAs. In this light, the definition of Veletsianos and Miller (2008) is adopted where a PA is defined as “a conversational virtual character employed in electronic learning environments to serve various instructional goals (for related definitions see Adcock and Van Eck, 2005; Baylor, 2002; Gulz, 2004).” The popularity of PAs started in the early 1990s when their effectiveness and educational perspective was examined in the first studies (Clarebout et al. 2002). What distinguishes ECAs from pedagogical agents is that the latter focus on supporting learning and instruction (Veletsianos and Miller, 2008). They are referred to as “learning partners” or “virtual tutors” and are used to facilitate learner motivation and learning outcomes (Clarebout and Heidig (née Domagk), 2012).

According to Veletsianos and Miller, (2008) most research on pedagogical agents has been experimental and quasi-experimental. These studies tend to explore cause and effect relationships, mostly evaluating the impact of the agents features such as gender, instructional role etc. on quantitative variables such as performance and engagement. Even though such research is methodologically important, the lack of consistency on research designs

makes comparisons among experiments and drawing conclusions difficult (Clark and Choi, 2005). Also, a review on pedagogical agents (Heidig and Clarebout, 2011) reveals that even though the number of empirical studies on the effectiveness of PAs is significant, there are still many open questions. This review emphasises the fact that not much is known about the effective design of PAs despite the large number of studies. There are mainly two reasons for this: 1) some aspects have only been evaluated once and 2) the complexity of PAs, as many variables come into play during their implementation (Clarebout and Heidig (née Domagk), 2012).

Moreno, (2005) does attempt to propose principles for the design of PAs, but even though all these principles are based on empirical studies, most of them are based on single studies only (Clarebout and Heidig (née Domagk), 2012).

Clarebout and Heidig (2012), also point out that there is a methodological issue regarding the state of the art on PAs and it is that most studies do not have a control group. Studies with more than one agent group are needed in order to answer how a PA should be designed in order to facilitate motivation and learning. The comparison to a control group is rather important as PAs are rather time and resource consuming and whether to be included into a multimedia learning environment needs to be decided based on empirical studies.

2.2.7 Multiple agents

Research involving ECAs is not new, research on the use of multiple ECAs though is still limited and quite fragmented. Even though there are multiple applications with more than one agent, the effect of their number and roles on usability has not been explored much. Only one study has been identified

to be specifically focused on the usability of multiple ECAs. Tracking the literature on multiple ECAs was challenging as the literature was fragmented and the referencing patterns inconsistent. Even from studies included in this review, most did not focus on the fact that they used more than one ECAs, as the focus was not on the number of ECAs but the effect of the application on the users. There were though, some notable exceptions that will be discussed below.

There are several applications that include more than one agent. Some examples of serious games specifically that have a single user interacting with multiple conversational agents are Tactical Language and Culture Training (Johnson and Valente, 2008), Crystal Island (Rowe, et al., 2010), Operation ARIES (Helpern, et al., 2012), Coach Mike (Lane, et al., 2011) and StoryStation (Robertson and Wiemer-Hastings, 2002) among others.

There are multiple reasons why developers and researchers might be interested in having multiple agents in an application. Empirical studies have shown that the delivery of information in the form of a dialogue instead of a monologue can be more effective for persuasion (Suzuki and Yamada, 2004) and education (Craig et al., 2000).

One of the first groups to use multiple agents was the ThinkerTools research group. A long-term goal of the group has been teaching general inquiry skills to middle school science students (Shimoda et al., 1999; White, 1993; White and Frederiksen, 1998). The group found that by using a model called the Inquiry Cycle, the student's inquiry skills improved. The Inquiry Cycle is a general model of how one does inquiry, starting with Question, then moving to Hypothesize, Investigate, Analyse, Model and Evaluate. In the applications developed by the group, the parts of this cycle are represented by an agent.

The research argues that metacognitive processes are understood more easily in a multiagent system even though there are no empirical evaluations to support it.

Later, people from the same group developed the Inquiry Island (White et al., 2002) where they argue that in order to develop expertise for lifelong learning for the students, they need to reify, reflect on, and improve their cognitive, social, and metacognitive processes. In that light, Inquiry Island, houses a community of software advisors, including an Investigator, Collaborator, Reflector, and Reviser. In this application the agents were represented by an inanimate humanoid cartoon face. The advisors are given a background story, speak as if self-aware and talk about having personal goals. Again, apart from a trial no conclusive empirical data exist to support the claims.

Amy Baylor and her team are one of the few groups who looked at the effects of the number of ECAs (PAs in their case). Baylor and Chang (2002) found that having two agents was more preferable to one when the system provided non-adaptive and just-in-time (compared to summative), feedback. Also, further research indicated that having two agents with distinct roles, one with expertise (Expert) and one with motivational support (Motivator), had a significantly more positive impact on the perceived value of the agents and learning. The study built upon previous research and the assumption that characteristics such as interaction, control and choice can be afforded by the presence of multiple agents. The study was conducted to measure learning effectiveness and motivation with 48 participants. The two-agent condition was Expert and Motivator and the one-agent condition was Mentor (which was designed as an aggregated version of Expert and Motivator). The results from this study support the notion that having two agents has a positive

effect on facilitating learning. Also, separating the agents' roles in a two-agent condition, reduced the learner's cognitive load. The effect of greater learning due to dividing the agents persona by functionality into two agents is known as the split persona effect (Baylor and Ebbers, 2003).

An extension of the previous study looked at the effectiveness of PAs roles for promoting learning and motivational outcomes. In this research 73 participants, in a between-subject design, were called upon to interact with one of three agents (Mentor, Expert, Motivator) while learning about instructional planning. The three contrast comparisons were: comparing the agents; value with and without motivation, with and without expertise and overall value of Mentor (a combination of Expert and Motivator). Results indicated that the motivational agents (Motivator & Mentor) were significantly more engaging, human-like and facilitative of learning compared to the Expert agent, but also less credible. Significantly more credible and better on the transfer measure were the agents with expertise (Expert and Mentor) compared to the Motivator agent, but they were also less human-like and supportive. Last, the Mentor was perceived as significantly more engaging and facilitative of learning and significantly better transfer performance compared to the other two agents (Baylor, 2003).

Even though the results from the aforementioned studies cannot be overlooked, the magnitude of these effects is not known as no effect sizes were reported.

Collin et al., (2004) presents an empirical study on the customer attitude to the usability of two interfaces that employ embodied conversational agents (ECAs), one version includes a single ECA, while the other version has multiple (three) ECAs in a banking scenario. For this study, he used a repeated

measures within-subjects design, with balanced exposure and 32 participants along with exit questionnaires to collect qualitative data. The user could interact with the ECAs using speech and the ECA responded verbally. The results from both quantitative and qualitative data indicated statistically significant support for the usability of the single agent version when compared to the multiple agent version. Even though the findings supported the use of a single ECA, the author highlighted that there are interesting areas for further research into multiple agent scenarios. At this point two important clarifications ought to be made. First, the study does not mention the effect sizes and 32 participants is a relatively low number of participants for the results to be conclusive. Second, the effects of ECAs on usability are application domain specific and cannot be assumed as transferable to other domains. Thus, the usability of ECAs must be examined for each application anew as there are multiple contributing factors that can affect it (Weiss, et al., 2015).

After reviewing research on multiple ECAs it was found that none of the previous studies assessed the usability of each ECA within the application but rather the usability of multiple agents collectively. This is one of the issues tackled in the research presented in this thesis.

2.2.8 Presumed benefits and challenges of ECAs

Albeit the technological advances in computing (in terms of software and hardware) that allowed the animated agents to become even more visually appealing for the user during real-time problem-solving advice, the effects of ECAs are debatable (Beun et al., 2003).

There is an assumption from those in favour of ECAs, that when included in the interaction they will increase the efficiency of human-computer interaction due to their anthropomorphism (Beun et al., 2003). Research indicates that such animated agents provide key benefits that enhance learning environments (Cassell et al., 2002). Also, ECA advocates claim that peoples' cognitive resources can be spent on the primary task because they will not have to learn a new way of communicating with the system, instead they will communicate with it as they would with any other person (Van Mulken et al., 1998). Another characteristic that ECAs possess that is not available with traditional interfaces is that users can communicate in parallel with the system by conveying non-verbal cues along with verbal instruction (Sepulveda, 2015). Also, the user might perform additional nominal tasks due to the face-to-face interaction between the ECA and the user (Kipp et al., 2006). Using ECAs in an interface can affect the way users realise the believability of a system. According to Dehn and Van Mulken, (2000) if a system is perceived as both intelligent and competent, then it is more likely for the user to attribute more believability to it (Doumanis, 2013).

According to Doumanis (2013), the arguments in favour of ECAs are that they can improve certain cognitive functions through enhanced motivation; positively affect learnability; positively affect the believability of a system and enhance trust-building with a user.

On the other hand, opponents to ECAs argue that the interaction will be hindered by the presence of a humanoid agent, since cognitive resources will be consumed in processing the visual information and speech (Walker et al., 1994).

Shneiderman is one of the biggest critics of ECAs. He argues that humanising the system may induce false mental models (Shneiderman and Maes, 1997). An example is that anthropomorphic agents may lead the user to believe that the system is also human-like in terms of cognitive aspects. That can make the user have expectations from the systems that it does not possess and may result in a negative experience (Doumanis, 2013).

At this point it must be noted that Shneiderman and Plaisant, (2004) are inconsistent on the definition of anthropomorphism in their arguments. At times they separate the anthropomorphism of the computer (i.e. making the user believe that the computer is an anthropomorphic entity) with the anthropomorphism of the software that gives feedback. They also suggest to the designers in their guidelines to prefer using 'appropriate humans for audio or video introductions or guides' (Shneiderman and Plaisant, 2004). However, Maes implies that the computer is a machine and the feedback given to a user (in that case of an interface agent) can be anthropomorphised (Maes, 1997; Murano, 2006).

According to Doumanis (2013), the main arguments against ECAs are that ECAs can induce false mental models of a system; reduce the sense of user control; might lead to cognitive overload and distract the user from the task. According to the media equation theory users respond socially to computers with minimal social cues. Thus, ECAs are redundant and tricking users into simulated relationships with ECAs is unethical (Doumanis, 2013).

The benefits of ECAs are also debatable relative to cognitive load theory (Sweller et al., 1998).

On one hand, according to cognitive load theory (Kalyuga et al., 1999), even the presence of an animated PA can add an extraneous cognitive load as it

will divide a user's attention to multiple information sources within the learning application (split attention effect). The modes of interaction employed by the agent provide very similar information (voice, facial expressions, text), which may result in a redundancy effect (Clark and Choi, 2005). Also, according to Valetsianos et al. (2008) contextually irrelevant pedagogical agents may increase extraneous cognitive load as contextual irrelevance can hinder learning as learners will have to attend to more than one schema.

On the other hand, an ECA that is well designed can reduce cognitive load as it can help focus the users on what they need to pay attention to. Also, the multiple modalities that the ECA provides may result in a modality effect⁹ (Louwerse et al., 2008).

Embodied conversational agents though cannot be labelled as good or bad as there is a plethora of contributing factors that can affect the interaction with an ECA. All the categories from the ECA design model (ECADL) (model described in section 2.2.4) can be altered and affect the interaction. A more advanced ECA (for example ECAs using multimodal input such as face recognition and natural language) can be more believable than their simplistic counterparts; the complexity of those agents though, comes with challenges as these systems are prone to mistakes (e.g. misinterpreting semantics of natural language) and demand more development time and expertise. One way to tackle these problems is to use more simplified approaches (e.g. decision tree mechanisms or simplistic graphics) but they also make for a less realistic experience. Trade-offs, such as the ones

⁹ "Improved learning that occurs when separate sources of non-redundant information are presented in alternate, auditory, or visual forms. The effect is explained by increased working memory capacity when using more than one modality." (Pagani, 2009)

mentioned above, makes finding the optimal approach in a specific setting a nontrivial task (Provoost et al., 2017).

2.2.9 Theories on embodied conversational agents related to this research

This section will discuss relevant theoretical work that can be linked to the body of research presented in this thesis. More specifically theories and possible effects related to the social role of ECAs, theories related to how humans process information and the illusion of humanness that is related to certain human-like characteristics possessed by an agent as experienced by the user. The following part of this review moves on to describe in greater detail the effects and theories related to the way users perceive ECAs as social beings.

2.2.9.1 Embodied conversational agents and social responses: CASA, media equation and ethopoeia

The main rationale for the use of ECAs in mobile serious games and in general is to provide an interaction metaphor that mimics human-to-human interaction. As mentioned previously, one of the advantages of ECAs is that the user can communicate with the system in a natural and intuitive way (Cassell and Stone, 1999). Thus, the added visual element to the interaction allows for a more natural gameplay (Doumanis and Smith, 2015). However, to

what extent the metaphor of human-to-human interaction can migrate to serious game HCI is still an open question (Doumanis, 2013).

Rationale for ECAs in general can be found in social psychology. One of the major theoretical foundations of virtual character and ECA research is the media equation. Nass, et al. (1994) proposed the "Computers as Social Actors (CASA)" approach that is now known as the media equation theory. It implies that people tend to interact with computers and media in an inherently social way. Even though the users know that the computer is a medium rather than a human being, they treat it in a social way as they would in human-human interaction (Nishida et al., 2014).

Experimental demonstration of this effect was carried out by Reeves and Nass (1996) showing that humans treated computers and media in an inherently social way although not consciously. The users rated seemingly "polite" computers as more favourable even though computers are not capable of expressing politeness. As a result, human-like interfaces such as virtual agents, pedagogical agents and ECAs would also be in principle subjected to social rules (Veletsianos, 2010). According to Kramer et al., (2015) the effects of ECAs can be described as "social" if they can evoke to the participant similar emotional, cognitive or behavioural reactions to the ones evoked by other humans.

Further research by Nass and colleagues (Nass et al., 1997; Nass and Moon, 2000) used the term "Ethopoeia" to describe the phenomenon that occurs during the interaction between a human and a virtual agent. The "Ethopoeia" explanation suggests that people unconsciously apply social rules when interacting with a virtual agent in a similar way they would with other humans. Additionally, they reject the hypothesis that people consciously

anthropomorphised computers thus they replied consciously as participants, but when asked denied doing so. The explanation according to Nass and colleagues can be found in the way the human brain has evolved to automatically recognise emotive reactions from humans (Kramer et al., 2015).

Studies supporting this notion have provided evidence that users/participants ethnically identify with virtual agents, respond politely and apply gender stereotypes to them (Scott et al., 2015).

Based on the media equation and the ethopoeia explanation, ECAs can be programmed to appear polite, extrovert, humorous or affective. Inherently people are subjected to pre-conceptions, stereotypes, first impressions and expectations; previous experimental work has shown that some of these biases migrate to human-ECA interaction. An example is that virtual characters may be stereotyped based on their appearance which can be used as an indication of their intelligence, competence and aptitude (Norman, 1997; Veletsianos, 2006; Veletsianos, 2010). First impressions may also influence the user's perception of the agent. Previous research conducted by Veletsianos (2010) and Cafaro et al., (2006) show the importance that first impressions play during the interaction both in terms of verbal and non-verbal behaviour. People process non-verbal cues by an ECA as information on which they make assumptions or draw conclusions. These assumptions consequently may influence the user's perceptions and behaviour. Another example can be found in the work of (Moreno and Flowerday, 2006) who explored the similarity-attraction hypothesis as applied to virtual characters. Their conclusions were that users preferred characters that were similar to them.

It should be mentioned that not all studies evaluating people's reactions towards virtual agents found social effects in the way that is demonstrated by CASA related studies (Kramer et al., 2015).

A full discussion of the biases that are introduced by the addition of an ECA in an interaction scenario lies beyond the scope of this study. Although they are mentioned as they are variables that affect the user's perception of the agent.

As discussed, previous research on whether the media equation theory can be applied to ECAs has provided evidence supporting this notion (Doumanis, 2013). Although, the evidence for the adoption of ECAs is encouraging, it is not a testament of whether their use can improve HCI or not (compared to text and menu GUI). Also, most experimental work related to the media equation and ECAs has been conducted on desktop computers. Mobile users may have a different reaction towards ECAs as mobile devices are most commonly used in places with ambient noise and crowds. Another issue is that research on mobile ECAs dates to early 2000 while the mobile and computer graphics technology has seen tremendous changes in recent years and most users are more technology literate. Based on these observations, further research on mobile ECAs is necessary.

2.2.9.2 **Persona effect**

Embodied conversational agents (ECAs) have the benefit over other interaction models in HCI that they provide an "intuitive" interaction which is embedded to human-to-human communication (Weiss et al., 2015). Takeuchi

and Naito (1995) put it simply as less effort being put by the user to learn new technical details of new interfaces, services and products because of the ECA.

According to Weiss (2015), the benefit of such a “natural” interaction can be empirically observed as the addition of anthropomorphic human-like interfaces (even as an additional feature of traditional interfaces such as web sites) can result in higher performance (measuring efficiency and effectiveness through time and scores) and higher quality (subjective evaluation). This effect is described as the persona effect.

“Persona effect” is a term coined in 1997 by Lester et al. The persona effect is described as the affective impact of animated pedagogical agents on students’ learning experience. More specifically, the persona effect has been revealed through the observation that “the presence of a lifelike character in an interactive learning environment -- even one that is not expressive — can have a strong positive effect on students’ perception of their learning experience”. Even though empirically studied, it is worth mentioning that this initial study did not include a control group with no agent.

The persona effect has been further empirically studied by Mulken et al., 1998 where the subjective measures results support the persona effect (the presence of an agent had a positive effect on the participants’ perception of the presentation). The presentation was perceived as less difficult and more entertaining even though the presence of an agent had no effect in comprehension. It must be noticed that the number of participants was 30. Findings as such highlight the strong effects of emotional communication shown by virtual personifications (Scott et al., 2014).

However, the persona effect is debated as it appears to be highly dependable on the conditions such as the task and system (Dehn and van Mulken 2000; Foster 2007; Yee et al. 2007).

Compared to spoken dialogue systems, ECAs can facilitate interaction (Dohen 2009) as the multimodal interaction can benefit human processing of information in higher neural activity and decreased load (Stein et al. 2009). This translates to ECAs being possibly less demanding to interact with, if the load of tasks is not large and the non-verbal signals are communicated properly (Weiss et al., 2015).

Although natural interaction has become a target, it should be taken into consideration that some users might not like to interact with a human-like agent but rather directly manipulate the interface.

Koda and Maes (1996) supported the notion that the presence of an ECA in a game application may result in increased entertainment. Also, non-verbal interaction that comes along with ECAs (such as eye contact) may increase attention (Takeuchi and Naito, 1995). Moreover, social effects such as social facilitation¹⁰ and politeness have been observed indicating that ECAs can indeed create social situations where phenomena of social psychology appear involuntarily (Sproull et al. 1996; Weiss et al., 2015).

¹⁰ “The tendency for people who are being watched or observed to perform better than they would alone on simple tasks” according to the encyclopaedia of PhycCentral:
<https://psychcentral.com/encyclopedia/social-facilitation/>

2.2.9.3 Uncanny valley

A risk when developing human-like interfaces is the uncanny valley effect. The theory behind the effect was developed by the Japanese robot designer Masahiro Mori. He claims that adding human characteristics in a robot (also applicable to virtual humans and in this research HECAs) made it quite charming to people till the point that these humanoid robots appeared to be quite close to human realism. Mori noticed that observer's reaction to these humanoid robots was "that when a person looks at this character there will be an instinctive feeling of uneasiness" (Mori, 1970).

In order to summarise his observations on how the characters' degree of realism can affect observer's impression of the character, Mori introduced a graph between the degrees of realism and how pleasant it is for the human observer (Figure 11).

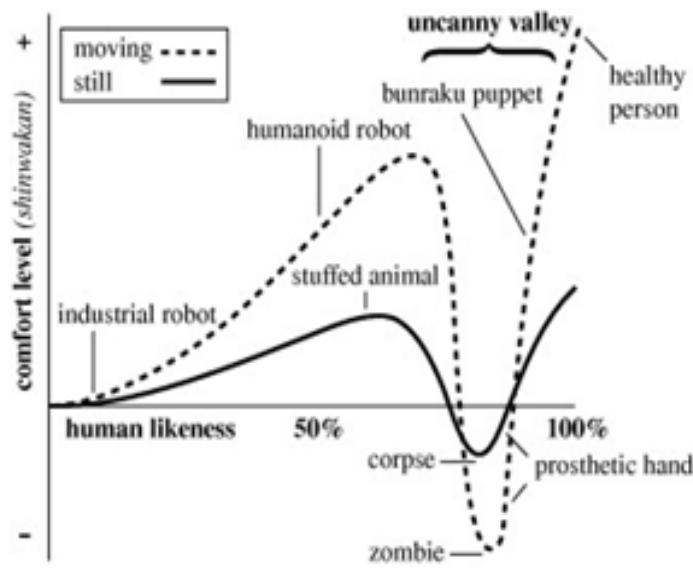


Figure 11- Mori's axis of uncanny valley.

Mori's axis offers a way to conceptualise human likeness in terms of a character's form (e.g. texture, shape), interactivity (e.g. timing) and dynamics (e.g. motions, facial expressions, speech) (MacDorman and Ishiguro, 2006).

Human observer's reaction to synthetic humans is believed to have an evolutionary origin. Steckenfinger and Ghazanfar (2009) enhanced this hypothesis with their paper "Monkey visual behaviour falls into the uncanny valley" where non-human primates tend to prefer looking at real photographs than computer-generated images of monkeys.

The uncanny valley effect can even be observed in neural activity (Saygin et al. 2012). The conclusion that the authors derived is that the effect can be based on perceptual mismatch where ultra-realistic human-like robots are expected to behave in an equally realistic human-like way. Therefore, the strength of the effect relies on how high the expectation is (Weiss et al., 2015).

This theory is believed to be true not only in robotics but in any type of artificial human-like objects (dolls, avatars, computer generated characters etc.); it is therefore considered during the creation of virtual humans and more relevant to this research, HECAs. Nevertheless, it must be pointed out that Mori amended his theory in order to acknowledge that game characters are less likely to evoke this feeling of uneasiness in contradiction with the virtual actors who lack interactivity (Aldred, 2011).

The technological advances in the area of computer graphics and gaming allowed designers and developers to approximate realism that can potentially fall into the uncanny valley. As people are very familiar with other human beings it is very easy for them to observe any irregularities and, therefore, the result may plunge into the uncanny valley. This uncanny feeling though can be provoked for multiple reasons. "If you are interacting with an android and

the timing of its speech and gestures is off, this will be uncanny for a different reason than if its eyes are too far apart" according to Dr. Karl MacDorman from the School of Informatics at Indiana University. Overcoming the uncanny valley has become, especially for video game designers, a pious hope.

According to Gratch, gesturing without any facial expression can look peculiar and vice versa. The same is true for the movement of hands without any involvement of the torso. Moreover, the facial expressions should accomplish any attempts of emotions because, otherwise, the lack of facial involvement could detract from the expected result even if the character's speech and gestures are synched (Gratch et al., 2013).

Uncanny valley though has not always been deemed as a negative. Robot designer David Hanson notes that realistic representations of humans have been the artistic subject of various artists from ancient Greece to contemporary art. Artifacts as such have been considered masterpieces instead of evoking uncanny feelings. Moreover, he indicates that realistic representations of humans can be used as tools, so a better understanding in human cognition and perception can be achieved (Hanson et al., 2005). By extending his observations from robotics to virtual humans, they can apply in every form of human representation (Korre, 2012).

2.2.9.4 **The illusion of humanness effect**

2.2.9.4.1 *Definition of illusion of humanness*

For the purposes of this research the illusion of humanness is defined as the user's notion that the system possesses human attributes and/or cognitive

functions. The illusion of humanness is not to be confused with anthropomorphism which is more related with the attribution of human properties to non-human entities or humanoid which almost always refers to having the appearance of a human as it was defined earlier in the chapter. When it comes to anthropomorphism, "attribution" is a key term as it implies that giving human characteristics to non-human agents is a conscious action from humans' side while "the illusion of humanness" is an involuntary reaction to a humanoid and anthropomorphic interface. The illusion of humanness is an extension of the "ethopoeia" explanation and persona effect but not limited to the unconscious application of social rules or an affective impact on learning but rather a determining factor on users' performance and perceived usability. It refers more specifically to systems that present information by utilising one or more human-like attributes (ex. voice, gaze, gestures, body) thus giving an illusion of "humanness" to the user. These attributes can be presented in textual, auditory and/or visual form. These attributes can be in the form of:

- gesturing
- facial expression
- eye gaze
- human-like movement
- voice
- embodiment
- behaviour (ex. using pronouns, personality, politeness, humour)

Isbister and Doyle, (2002) claim that in order to make a powerful visceral reaction to the agent – evoke the "illusion of life" – the character should have an appearance along with sound and movement. Studies in this area are not

limited to the “realism” rather amplifying the user’s reaction to the agent. A few examples would be enhanced realism of the agent’s movement; creating the right visual style for specific applications; create natural sounding speech for the character. Bates (1994) writes, “To our knowledge, whether an agent’s behaviour produces a successful suspension of disbelief can be determined only empirically.”

2.2.9.4.2 Anthropomorphism

The illusion of humanness is related to anthropomorphism but is not synonymous. Anthropomorphism is the attribution of human characteristics to non-human entities and is a combination of the Greek words for human and form/appearance (ἀνθρώπος + μορφή). The word “anthropomorphism” etymologically is more relevant to the appearance, but it has also been used in the past to describe human-like behaviour in the field of HCI or even an umbrella term for human-like interfaces therefore it will be briefly explored as the factor the evokes the “illusion of humanness” effect.

In psychology the term “anthropomorphism” has been used rather loosely to describe a range of different things from deductions about non-human agents to almost any type of dispositional assumptions about non-human agents. The loose use of the term does not fit with the dictionary definition of the word which is “attributing human characteristics or behaviour to a god, animal, or object” (Soanes and Stevenson, 2005). Thus, anthropomorphism goes beyond dispositional assumptions about non-human agents and requires attributing human-like appearance or mind to an agent. Hence, anthropomorphism is attributing characteristics that are considered uniquely

human to non-human agents such as mental characteristics (emotion, cognition etc.) (Waytz, et al., 2014). The presence of a mental state establishes enough condition for humanness as the presence of a humanlike face or body implies humanlike cognitive state as well (Johnson, Slaughter, and Carey, 1998; Morewedge, Preston, and Wegner, 2007).

When it comes to computer UI, anthropomorphism involves an entity which is usually part of the UI, that exhibits some human characteristic (De Angeli, Johnson and Coventry, 2001). Not all human characteristics need to be present for the entity to be considered anthropomorphic but can have one or two, for example a stick figure with eyes and mouth.

The psychology of anthropomorphism was examined by Adam Waytz (Harvard University) and Nicholas Epley (University of Chicago). This neuroscience research revealed that when people think of humans and non-human entities, the same brain areas are activated. This result is an indication that anthropomorphism utilises the same processes as the ones used when thinking of other people. Thus, anthropomorphism can evoke a certain mental response (illusion of humanness) where people think of non-human entities as human consequently render them worthy of consideration or moral care (Waytz et al., 2014).

Although there is a tendency from humans to anthropomorphise, they do not attribute human characteristics to every object they come across. This selectivity is partly due to the factor of similarity. If an entity possesses many human-like traits, such as human-like facial features and movements, then it is more likely to be anthropomorphised (Nauert, 2010).

Since 2000, there have been a range of different terms defined and affiliated with anthropomorphic interfaces. The initial introduction came with the CASA

model that was discussed earlier, by Nass et al. (1994). CASA associated social cues with HCI. The most commonly format of anthropomorphic interfaces researched are ECAs along with human likeness interfaces, avatars and human-like computer interfaces (Tuah, 2018).

Beun (2003), supported that ECAs have the capacity to improve multimodal HCI due to their humanlike communicative behaviour and appearance. He adds that since human to human interaction comes naturally to people then anthropomorphising the agents would improve the process of communication which implies that the agent would have social effect to the user. Although his opinion is not always shared by other researchers as discussed earlier.

Sepulveda (2015), even though recognising that rapport is not exhibited only by humans but also by some animals, based his research on the premise that rapport is an exclusively human state of interaction. Thus, he supports that anthropomorphic agents are preferred to establish, evaluate and analyse rapport. Consequently, ECAs create a stronger bond with the user as they exhibit human traits that are more easily understood.

Rapport is also connected with believability and trust as ECAs can affect the way in which users perceive the believability of the system. Dehn and van Mulken (2000) found that when a system is perceived as competent and intelligent then it might be perceived as more believable by the user. Cassel and her associates also believed that the more natural the conversation with the ECA the higher the believability (Cassel and Stone 1999). Trust is also connected with believability; especially when information is provided, trust becomes a key factor (Doumanis, 2013).

Cassel's approach in building trust in ECAs is by establishing and maintaining social relationships with ECAs. The justification is that interaction paradigms in human to human interaction such as small talk and greetings along with the embodiment of the agent and speech can users to think that the system is more knowledgeable and reliable and thus could be trusted more (Cassel and Bickmore 2000).

To avoid any confusion the research presented in this thesis has been concerned with anthropomorphic agents within an application and not anthropomorphising the computer itself.

2.2.9.4.3 Levels of anthropomorphism and human likeness

Up until the time this thesis was written no levels of anthropomorphism were found in the literature of ECAs and HCI. Thus, three levels of anthropomorphism are proposed.

Anthropomorphism can take on various forms at the user interface. The simplest form is textual, another form concern using auditory cues while visual cues of multiple manifestations can be used and typically would involve using text and/or voice audio (Murano, 2006). Figure 12 shows the spectrum of human likeness in application interface design.

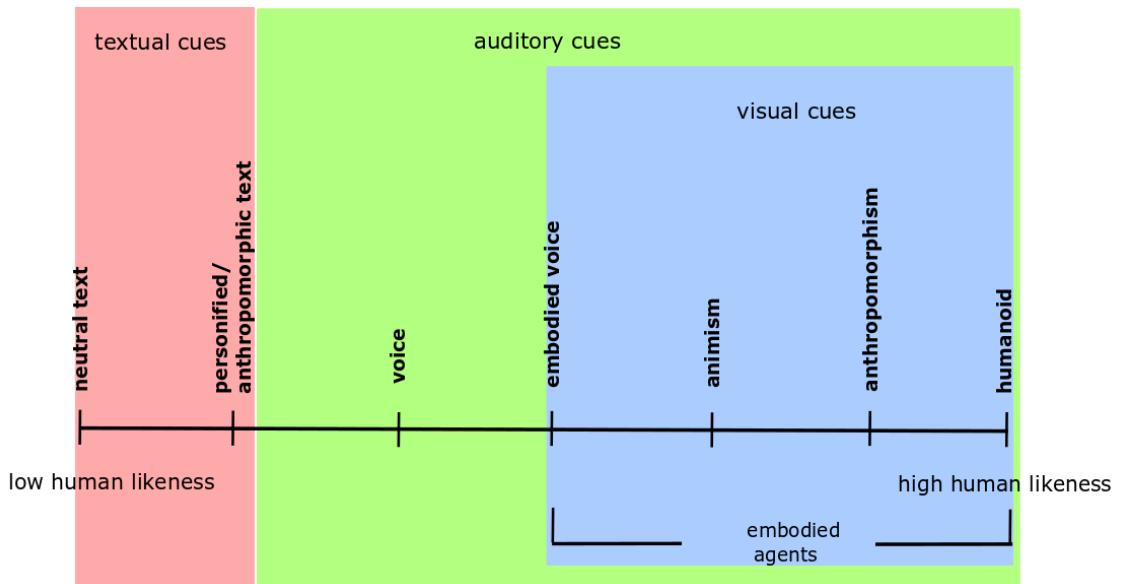


Figure 12 Spectrum of application interface design in relation to human likeness.

Textual cues

The anthropomorphised aspect of this type of feedback is the way text is written on the screen, i.e. using pronouns such as "I". Some chatbots are also an example of displaying textual anthropomorphism or personification. One of the most well-known and one of the earliest chatterbots is ELIZA that was developed by Weizenbaum (1976). Eliza is a mock Rogerian psychotherapist that uses human-like communication paradigms such as trying to engage to conversation and using pronouns such as "I" when referring to itself. Other most recent examples of textual cues are different purpose chatbots such as the system ALICE was developed by Dr. Richard Wallace way back in the early

Internet in 1995, Endurance: A Companion for Dementia Patients¹¹, Casper: Helping Insomniacs Get Through the Night¹², MedWhat: Making Medical Diagnoses Faster¹³ and many more.

Machine learning advancements in recent years have seen chatbots becoming more popular as they are more responsive, clever and helpful. What all the above systems have in common is that they mimic human to human interaction (refer to themselves with pronouns, having names, engage with the person they converse, sometimes use humour etc.).

Auditory cues

Auditory cues or “voice” are usually expressed in the form of Text-to-speech (TTS) technology or dynamically loaded voice clips of humans. The system may also use pronouns such as “I” to refer to itself. An example of a system using auditory cue is the virtual assistants that have recently became rather popular. Home virtual assistants such as Amazon Alexa and Google home but also mobile virtual assistants such as Siri, S voice and ok Google are a few examples of virtual assistants that speech recognition and voice output of a TTS form. Some of these systems have names associated with them such as Alexa and Siri which give the illusion of an identity and further anthropomorphises the system.

The mere existence of voice expresses anthropomorphism. Even for the systems with no allocated names, the mere fact that they have a human-like voice gives an illusion of persona or identity due to extra-linguistic data provided through the voice beyond the context of the message such as

¹¹ <http://endurancerobots.com/azbnmaterial/a-robot-companion-for-senior-people-and-patients-with-alzheimer-s-disease/>

¹² <http://insomnobot3000.com/>

¹³ <https://medwhat.com/>

intonation, gender etc. When humans hear a voice, they can in most cases understand emotion based on the tone that is being used. Speech can reveal cues on the speaker's personality, beliefs (for example hesitation), cognitive process, social membership etc. (Zara, 2007). Thus, how something is said is also of importance.

When human voice along with NLI is used by the agent then the users might forget the limited capabilities of the system and expect human-like ones.

When developing usable spoken multimodal systems, the appropriateness of speech interaction must be decided for each application anew based on the purpose and environment of the application (Dybkjær et al. 2004).

Visual cues

Non-verbal communication and extra-linguistic information are also of importance and can be anthropomorphic. Developing ECAs that mimic humanlike non-verbal behaviours reinforces the understanding that the inclusion of non-verbal behaviour enhances the human-agent interaction. Images that are characterised as anthropomorphic can range from simple stick drawings to hyper realistic 3D characters (Murano, 2005). This includes video clips of humans (Bengtsson, 1999). Non-verbal behaviour includes but is not limited to lip-synching that is accurate with ECA speech output, gesturing, facial animations such as eyebrow raising and change of eye gaze. Face animations (rising of eyebrow, smiling etc.) have been used successfully to communicate emotion and signal speech input from the user (Doolin, 2014). Appearance influences people's cognitive assessments (Nass, 2000).

According to Zara et al. (2007) these anthropomorphic characteristics involve the same set of modalities as the expressions of emotion:

- According to Knutson (1996), the face reveals personality while Baron-Cohen et al. (1996) support that it reveals mental state as intention or beliefs. Facial animations (rising of eyebrows, smiling etc.), have been successfully used in ECA development to portray emotion and acknowledge speech input from application users. Also, experiments contacted by Dryer (1999), showed that characters designed with round shapes, big faces and happy expressions were perceived by the participants as extroverted and agreeable while characters designed with big bodies, bold colours and erect postures were perceived as extroverted and disagreeable. According to Gultz and Haake, (2006) here is almost no research that involves systematic studies of different facial looks.
- According to Baron-Cohen et al. (1996) eyes reflect cognitive activity and provide context of the nature of interpersonal relationship (Hall et al, 2005). In interactions involving more than one user at a time, eye gaze has been used to signal who should speak (Bohus and Horvitz, 2010).
- Argyle (1980), claims that gestures are physical representation of beliefs, intention and so on. Through a series of experiments Foster (2007) found that when speech is combined with appropriate hand gestures, the usability of human-ECA interaction is significantly enhanced.

Serious games (SGs) and mobile serious games (MSGs)

2.3 Serious games (SGs) and mobile serious games (MSGs)

2.3.1 Serious games

Computer games are undisputedly popular in modern society. Statistics show that the games industry is the fastest growing entertainment industry with 2.2 billion people playing games around the world. The global games market in 2017 was expected to increase by 10.7% compared to 2016, with a value of \$116 billion and a projected compound annual growth rate (CAGR) of 8.2% by 2020. Also, mobile gaming is expected to represent more than 50% of the total games market in 2020.¹⁴ In 2014, a market research report showed that the average gamer is 31 years old with 59% of US Americans playing video games and 51% of U.S. households own a dedicated device for playing games. Also, for the same year the report shows that consumers spent \$21.53 billion in 2013 on the game industry.¹⁵ Forward three years, in 2017 the data show that the average gamer is 35 years old and 67% of U.S. households own a dedicated device for playing games. Additionally, the money spent on the game industry rose from \$21.53 billion in 2013 to \$30.4 billion in 2016.¹⁶ The UK market is also rising with estimates suggesting that it will worth £5.2

¹⁴ Reported by Newzoo: <https://newzoo.com/insights/articles/the-global-games-market-will-reach-108-9-billion-in-2017-with-mobile-taking-42/>

¹⁵ Reported by ESA (Entertainment Software Association) ESSENTIAL FACTS About the computer and video game industry 2014: http://www.theesa.com/wp-content/uploads/2014/10/ESA_EF_2014.pdf

¹⁶ Reported by ESA (Entertainment Software Association) ESSENTIAL FACTS About the computer and video game industry 2017: http://www.theesa.com/wp-content/uploads/2017/09/EF2017_Design_FinalDigital.pdf

billion by 2021, becoming the largest market in Europe.¹⁷ Also, video games, virtual reality (VR) and e-sports will continue to be among the top preferences for consumers with a remarkable estimated growth for the next 5 years.¹⁸ The amount invested in educational technology companies between 1997 and 2017 was \$37.8 billion in total. More than half of this investment (62%) took place in the last three years alone (2015 - 2017)¹⁹. Regarding the game-based learning market specifically, in 2016 the worldwide revenues reached \$2.6 billion. The global five-year CAGR is 22.4% with a forecast showing that the revenues will rise to \$7.3 billion by 2021.²⁰

This fast growth is attributed to the popularity of games especially among younger people making them a great medium to obtain information and knowledge (Lenhart et al., 2008; Seng and Yatim, 2014). The areas of computer graphics, video games and interactive visual simulation were drastically affected by the technology advancements and especially the affordable prices of high-performance graphics hardware (Encarnacao, 2009).

The combination of information and curricular material with games and later computer games has been long proposed. Computer and video games were originally designed for entertainment, but were repurposed for training, promotion and education due to the growing general familiarity with games

¹⁷ Reported by:

<https://ukie.org.uk/sites/default/files/UK%20Games%20Industry%20Fact%20Sheet%20February%202018.pdf>

¹⁸ Reported by PWC: <https://www.pwc.co.uk/industries/entertainment-media/insights/entertainment-media-outlook.html>

¹⁹ Reported by: Metaari (http://users.neo.registeredsite.com/9/8/1/17460189/assets/Metaari_s-Analysis-of-the-2017-Global-Learning-Technology-Investment-Pat27238.pdf)

²⁰ Reported by: Ambient Insight: <http://seriousplayconf.com/downloads/the-2016-2021-global-game-based-learning-market/>

and gaming techniques (Kapp, 2007). All the advancements resulted in the development of serious games (SGs).

While there is rapidly increasing interest in SG, the empirical data informing best practices for their design remain relatively limited (Dörner et al., 2016). Further research is required to determine the best way to develop SGs that fulfil their potential in different learning contexts. Serious games are used for purposes such as ergonomics analysis, training, simulation and learning. Dörner et al also argue that apart from the technological advances, other aspects contribute to making games attractive for purposes different than entertainment. According to Freeman (2003), some of those aspects are advanced methodologies such as emotionally involving players with the game. According to John and Srivastava (1999), again apart from the technological advancements, researchers have identified that the aesthetic presentation and narrative are also important factors for game enjoyment (Dörner et al., 2016).

With the increasing role of SGs in the corporate pipeline, the importance of usability analysis of these systems increases accordingly. In this new concept, it is essential that not only the system but also the user is taken into consideration. The human factor in SGs is quite important, as it is a medium of interaction, information and training (Korre, 2012).

Serious games are the result of many contributing disciplines as shown in Figure 13.

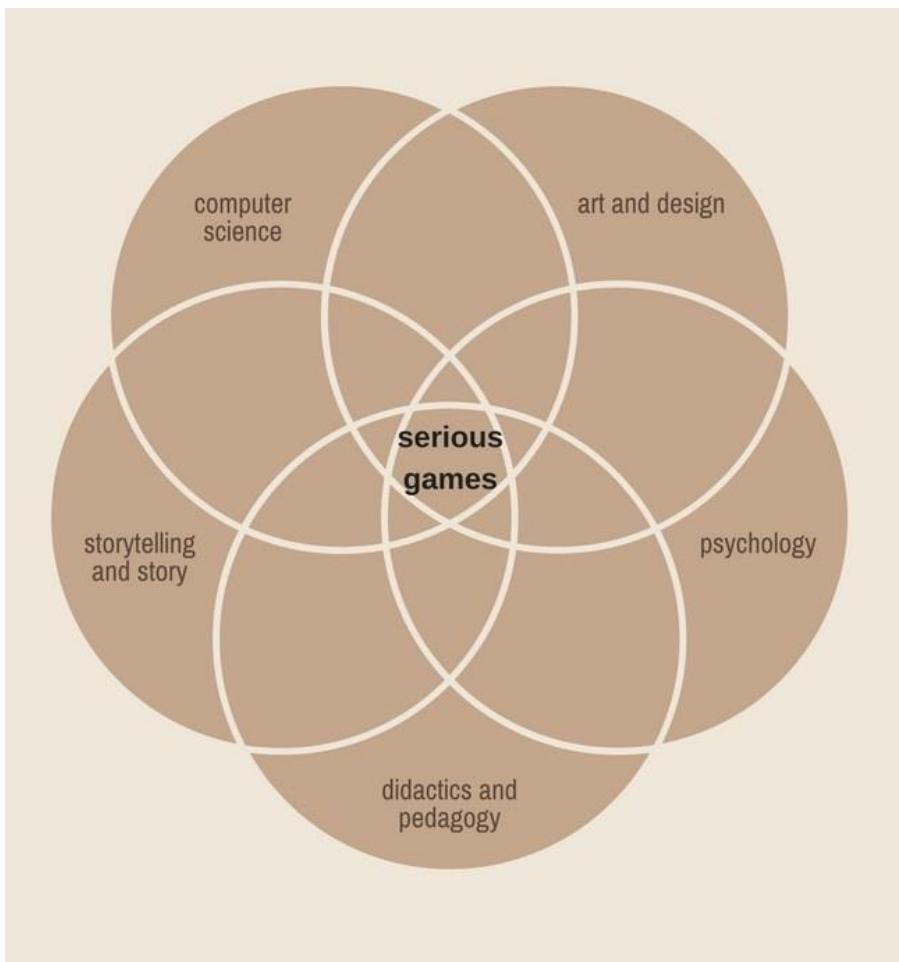


Figure 13 - Disciplines contributing to serious games (Adapted from Dörner et al., 2016).

2.3.1.1 Definition of Serious Games

A search in the literature reveals that definitions of SGs have been disputed and there are many definitions of the term. IGS Global give the astonishing number of 42 different definitions for SGs.²¹ What most agree on though, is

²¹ IGS Global definitions for Serious Games: <https://www.igi-global.com/dictionary/serious-games/26549>

the principle that SGs are digital games that are developed not just for entertainment reasons (Susi et al., 2007; Michael and Chen, 2005).

To define SGs, we first need to define games. Many scholars worked on the classification of computer games and their potential (Garris et al., 2002; Malone, 1981; Prensky, 2003) but according to Botturi and Loh, (2008), defining games is also a very disputed topic. An exhaustive description of every definition for games is beyond the scope of this review; it is necessary though to adopt a definition for the process of defining SGs. For this research the author adopts Wouters's approach. A game has to be interactive (Prensky, 2001; Vogel et al., 2006); to have a clear goal often set by a challenge (Malone, 1981); to be based on a set of agreed rules and constraints (Garris et al., 2002) and to provide feedback so the players can monitor their progress (Prensky, 2003). Wouters also argues that a competitive element is part of games but is not a necessity for SGs. Wouters's approach was adopted as the most suitable definition for what a game is. This definition of games is used as the basis on which the definition of SGs adopted in the present research is based.

From the various definitions of SGs, the most popular are presented chronologically in Table 1.

Author	Definition
(Abt, 1970)	<i>"Games may be played seriously or casually. We are concerned with serious games in the sense that these games have an explicit and carefully thought-out educational purpose and are not</i>

	<i>intended to be played primarily for amusement. This does not mean that serious games are not, or should not be, entertaining."</i>
(Sawyer, 2003;2007)	<i>"Serious games are those games produced by the video game industry that have a substantial connection to the acquisition of knowledge."</i>
(Zyda, 2005)	Zyda (2005) in his attempt to define serious games, also suggested a classification of games and video games. According to Zyda (2005), a game is " <i>a physical or mental contest, played according to specific rules, with the goal of amusing or rewarding the participant.</i> " A Video Game is " <i>a mental contest, played with a computer according to certain rules for amusement, recreation, or winning a stake</i> " and last a Serious Game is " <i>a mental contest, played with a computer in accordance with specific rules that uses entertainment to further government or corporate training, education, health, public policy, and</i>

	<i>strategic communication objectives."</i> Zyda's definition is an expansion of Sawyer's definition of the term.
(Michael and Chen, 2006)	<i>"Games that do not have entertainment, enjoyment or fun as their primary purpose"</i>
(Bergeron, 2006)	<i>"a serious game is an interactive computer application, with or without a significant hardware component, that: has a challenging goal, is fun to play and/or engaging, incorporates some concept of scoring, and imparts to the user a skill, knowledge, or attitude that can be applied in the real world."</i>
(De Freitas, 2006)	<i>"The serious games movement is a trend towards designing and analysing the use of games (and simulations) for supporting formal educational and training objectives and outcomes. The movement aims to meet the significant challenge of bringing together games designers and educationalists to ensure</i>

	<i>fun and motivation as well as demonstrating educational value."</i>
(Dörner et al., 2016)	<i>"A serious game is a digital game created with the intention to entertain and to achieve at least one additional goal (e.g., learning or health). These additional goals are named characterizing goals ."</i>

Table 1 Definitions of SGs.

Building upon the definition of games given by Wouters, the author adopts a combination of the definitions given by Michael and Chen, Bergeron, De Freitas and Zyda as the definition of SGs used in this research. According to these definitions, SGs are games, therefore interactive, with a clear goal, based on a set of rules and provide feedback (Wouters, 2013); whose primary purpose is not entertainment or enjoyment (Michael and Chen, 2005); yet they are fun to play and/or engaging, have a scoring system (feedback) and teach a skill, knowledge or attitude to the user that can be then used in the real world (goal) (Bergeron, 2006b); and "a mental contest, played with a computer (interactive) in accordance with specific rules (rules) that uses entertainment to further government or corporate training, education, health, public policy, and strategic communication objectives" (Zyda, 2005). De Freitas definition refers to SGs as a movement and not as a game style or type which underlines that SGs can be designed as any type of game (De Freitas, 2006). These definitions work cumulatively and characterise SGs as entertaining game applications, not bound to specific game genres or

styles, with rules and scoring systems that aim to communicate a transferable skill. Bergeron and Zyda are the only ones referring to a scoring system or reward as an important element of a SG and game respectively while Michael's and Chen's general definition is widely accepted and succeeds to differentiate SGs from entertainment games. The definition given by Dörner et al. (2016) is also very interesting as it mentions the characterising goals of SGs that will be explored later. He also suggests that the purpose of a SG in some cases can be defined by the player as well as the developer. For example, the game Doom can be used by the player for training motor skills and thus becoming a SG even though it was not developed for this purpose. It is also important to note that SGs are not a game genre but more like an approach (Dörner et al., 2016b).

The characterising goals of SGs according to Dörner et al., "can be matched to *competence domains*, e.g., cognition and perception, emotion and volition, sensory-motor control, personal characteristics, social attitudes, and media use" (Dörner et al., 2016).

2.3.1.2 **Serious games and similar concepts**

Serious games and game-based learning (GBL) are used interchangeably in the literature. SGs have been developed for a broader spectrum of purposes that is not limited to education or learning (Sawyer and Smith, 2008) and GBL is actually a subcategory of SGs (Hainey et al., 2011). Even though GBL is more specific to learning, as a sub-category of SGs the author includes the most relevant research on GBL in this review.

Serious games can be used to change attitudes and behaviours (Bogost, 2007) and for skill acquisition and training (Boyle et al., 2011) or intentionally

for learning (Boyle, 2014). In a systematic literature review by Connolly et al. (2012) the learning and behavioural outcomes of SGs are classified as follows: affective and motivational outcomes, behaviour change, knowledge acquisition/content understanding, motor skills, perceptual & cognitive skills, physiological outcomes, social/soft skill outcomes.

Even though the term "serious games", just like the term "game based learning", can technically be used for non-digital games, after the work of Sawyer and Rejeski (2002) and the Serious Games Initiative (SGI) -- founded by the Woodrow Wilson Centre in Washington DC (Susi et al., 2007), the term is used to describe digital SGs games (Wilkinson, 2016). For GBL the distinction is clearer as for the digital manifestation of those games, the term "digital game-based learning" (DGBL) is used.

Serious games are often developed for learning, but they can be used for other purposes such as the acquisition of skills. As mentioned previously, SGs can have other characterising goals and can be divided into categories according to those goals e.g. exergames, advergames etc. (Dörner et al., 2016b).

As a concept, SGs are an umbrella term for simulation games, (digital) GBL, mobile-based learning etc. The basic distinction of SGs from edutainment²² and simulations is that for an application to be characterised as SG it must be a game with at least some of the characteristics that games possess.

Sawyer (2007) claims that "too often SGs are defined only as those which the definer does!" (Sawyer, 2007), while Smith and Sawyer (2008) argue that "most labels define a specific output ignoring the larger possibility space for

²² Education through entertainment

SGs. This implies the possibility space for SGs only equals that specific label" (Smith and Sawyer, 2008).

As stated in the white paper "Why serious games work" (PIXEL learning, 2011), SGs, immersive learning simulations or game-based learning have the same meaning which is the use of computer game techniques integrated into traditional learning methods but are not one and the same.

Edutainment is another term associated with SGs, Michael and Chen though claim that SGs "are more than just 'edutainment'" (Michael and Chen, 2006). Edutainment and SGs overlap when edutainment is delivered in the form of a digital game in which case it becomes an SG.

The difference between SGs, gamification, games and gameful design is also illustrated in Figure 14.

	Game Thinking	Game Elements	Game Play	Just for Fun
Gameful Design				
Gamification				
Serious Game / Simulation				
Game				

Figure 14- Differences between Serious Games and Gamification (Marczewski, 2013).

2.3.1.3 History of serious games

Even though SGs appear to be a relatively recent phenomenon, it is not a new concept. Plato, for example, explored the role of play and identified its effects on the development of children into adults (D'Angour, 2013). Also, he regarded philosophy to be a joyful game but serious nonetheless (Ardley, 1967).

The term "SGs" was criticised for its literal meaning; it is an oxymoron because games are supposed to be inherently fun and not serious according to Newman (as cited in Ritterfeld et al., 2009). Despite this paradox many academics and professionals think that SGs can be both fun and educational, purposeful, impactful, meaningful and engaging (Ritterfeld et al., 2009).

It can be claimed that what we know today as SGs is a modern manifestation of eons of practices and theories; SGs exist in a non-digital form for centuries such as chess where a militaristic metaphor was applied to a board game (Wilkinson, 2016).

The term "SGs" was coined by Clark C Abt in 1970 (Djaouti et al., 2011; Susi et al., 2007) referring to applications of game theory in areas such as economics, management, training and education (Abt, 1970).

Although there are examples of SGs in a non-digital format such as "The New Alexandria Simulation: A Serious Game of State and Local Politics" and Abt's work is mostly about analogue simulation games, the current use of the term refers to digital games/applications (Wilkinson, 2016). Analogue SGs are beyond the scope of this research therefore, they will not be mentioned

extensively in this review. Further information on analogue SGs can be found in the work of Djaouti et al. (2011) and Wilkinson (2016).

Computer based simulations have been used by the US military since 1948 with Air Defence Simulation, but it was the rise of arcade games and game consoles towards the end of the 20th century that boosted the SGs development. This is attributed to the popularity of commercial games.

In the 1980s, existing games were repurposed for advertising (Pepsi invaders) (Spence, 1988) and new ones developed by the military along with Atari (The Bradley Trainer) (Wilkinson, 2016). Also, games were used in healthcare for rehabilitation (Griffiths, 1997; 2003) and psychotherapy (Gardner, 1991; Spence, 1988).

With its contemporary use though, the field of SGs was "resurrected" by Sawyer and Rejeski (2002) with their paper "Serious Games: Improving Public Policy Through Game Based Learning and Simulation" who associated SGs with video games, the game America's Army by the US Army and the Serious Games Initiative (SGI), founded by the Woodrow Wilson Centre in Washington DC, all in 2002 (Susi et al., 2007). The year 2002 is also considered to be the starting point of the current wave of SGs (Djaouti et al., 2011). Djaouti found that 65.8% of SGs before 2002 were educational, 10.7% were advertising and 8.1% of them were orientated to ecology. After 2002, the landscape of SGs changed with advertising taking the lead (30.6%) closely followed by education (25.7%) and healthcare (8.2%) (Djaouti et al., 2011).

The field of SGs emerged from the tremendous technical, cultural and business growth of the game industry in the last decades (Ritterfeld et al., 2009). It is a relatively new concept, which allows the use of digital games in several applications for educational, learning and informative purposes (Susi

et al., 2007; Korre, 2012). Serious games have gained momentum in the past decades both in academic research (Ritterfeld et al., 2009) and the entertainment industry (Alvarez and Michaud, 2008; Susi et al., 2007).

We are going to use the term SGs with its contemporary use, hereafter, to describe only digital SGs. A comprehensive historic overview of SGs can be found in the work of Djaouti et al. (2011) and Wilkinson (2016).

2.3.1.4 **Context of use**

In SGs, the user can be either represented by an avatar and interact with the environment or be an observer who experiences a certain scenario (Magnenat-Thalmann and Kasap, 2009).

Serious games can take advantage of the latest games technologies to create virtual spaces for interactive experiences. These games can exist in various forms such as: web-based applications, mobile applications, more sophisticated stand-alone computer games (Cassell et al., 2001), virtual reality (VR) and augmented or mixed reality (AR/MR).

It is argued that through SGs the users can improve their perception, attention and memory as they remain strongly engaged (Tramonti et al., 2014). As a result, SGs are useful tools for promoting the cultural heritage (e.g. 'Olympic Pottery Puzzle' (Gaitatzes et al., 2004)); and marketing promotion activities (Milka Biscuit Saga²³), corporate training, educational activities and social campaigns.

²³ <http://serious.gameclassification.com/FR/games/44282-Milka-Biscuit-Saga/index.html>

A recent systematic literature review on SGs evaluation (Calderon and Ruiz, 2015) categorised the application areas for SGs as health and wellness, culture, professional learning and training, social), support (SGs developed to support and help people in life's decisions) and education. Health and wellness include games for improving the quality of life and create awareness. Culture refers to SGs used for cultural training. Learning and training includes SGs used by companies to train and teach their staff. Serious games used for social skills training are in the social category; the Money world application which was used in this work is an example of a social serious game. Support refers to SGs developed to support and help people in their lives' decisions. Most of the SGs were classified as educational (53.52%) followed by health and wellness (20.2%), professional learning and training (18.18%), culture (5.5%), social (4.4%) and support (1.1%).

Another categorisation was made to identify the types of SGs that have been assessed throughout the years. The majority of SGs were computer based (58%) followed by web based (10%), videogames (9%), virtual world (8%), mobile (6%), board games (5%), massively multi-player online role-playing games (MMORPG) (2%) and LEGO-based (2%). Although these categories do not appear as though they would be mutually exclusive, they provide an insight on the media used to deliver SGs. Additional information is given on the quality characteristics with most primary studies assessing the learning outcomes followed by usability and user experience. It is also revealed that 55% of the studies has a sample size of 1 to 40, 22% of 41 to 80, 8% of 81 to 120 and 15% of more than 120 participants. The number of participants is important to assess the effect size of the observed phenomenon. This paper concludes that more randomised control trials are needed in the field of SGs

to provide more rigorous evidence of their effectiveness (Calderon and Ruiz, 2015).

2.3.1.5 Feedback and serious games

A growing body of research reveals that games (video and computer based) have a strong instructional value which is enhanced by the fact that they are cost effective, safe and meaningful to today's users (Susi et al., 2007). Even though SGs sound like a good idea in paper, if the execution is poor, the effects on the user might also be poor. Kiili (2006) has suggested that many SGs for educational purposes lack the distinctive interactivity that games possess and simply resemble digital exercise books which in turn can lead to reduced motivation.

One way to tackle this issue is by providing feedback. A common concept in game design regarding motivation is feedback loops. Depending on the context of the game, feedback loops can be positive or negative. Either way, their main components are the same (see Figure 15):

- User performs an action
- Feedback is given
- User experience is modified
- Repeat

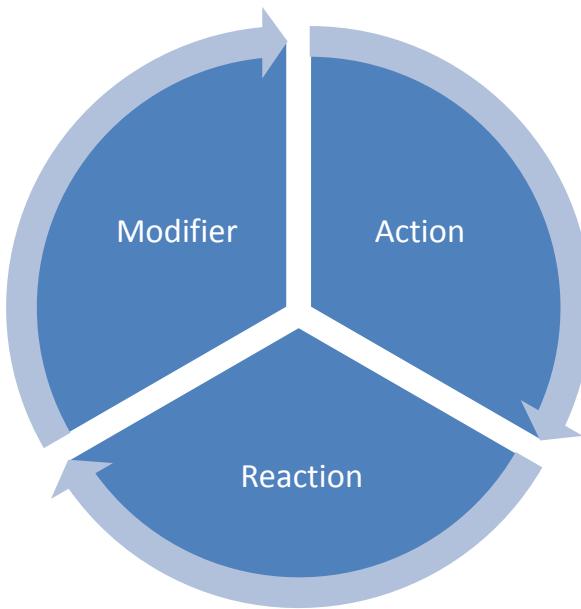


Figure 15 Feedback loop.

A positive feedback loop will amplify motivation while a negative loop will reduce it (Marczewski A., 2013).

A game mechanism that is frequently used for feedback is rewards. Badges, stars and points are a few of the rewards given to players. While intrinsic²⁴ motivation has been considered by learning designers as more valuable, research supports that extrinsic motivation reward mechanisms found in games helped where the learning content is not perceived as valuable and/or interesting by the players (Lepper, 1988). Research also indicates that rewards subject to performance (extrinsic motivation) in games can result in intrinsic motivation and positively change attitudes to certain tasks (Harackiewicz and Manderlink, 1984, Eisenberger et al., 1999).

²⁴ "In Self-Determination Theory (SDT) (Deci & Ryan, 1985) we distinguish between different types of motivation based on the different reasons or goals that give rise to an action. The most basic distinction is between intrinsic motivation, which refers to doing something because it is inherently interesting or enjoyable, and extrinsic motivation, which refers to doing something because it leads to a separable outcome." (Ryan & Deci, 2000)

It has been indicated that reward mechanisms provide a sense of fun to the user/player by promoting intrinsically rewarding experiences that can be equally or even more important than extrinsic rewards.

Also, it is important when clarifying a short-term goal that the feedback is quick, something that reward systems can accomplish (Wang and Sun, 2011).

Reward system heuristics can be applied to real-life settings such as workplaces. Reeves and Read (2009) have provided evidence on how digital-game mechanisms and elements can be used to improve employee performance and satisfaction showing how transferable those tactics can be in real life (Wang and Sun, 2011).

Nonetheless, rewards can reduce intrinsic motivation when they are used incorrectly (Deci et al., 1999). When it comes to SGs, rewards should be well managed and relevant to the context of the game as frequent use of rewards out of context can result in losing their motivational value (Donovan, 2012).

2.3.1.6 Review of evidence of benefits of serious games

A particularly controversial subject for SGs and similar concepts such as DGBL, GBL, e-learning and edutainment is their claimed positive effects.

A growing body of research reveals that games (video and computer based) can have an educational value and assisted the learning process effectively (Chiang et al., 2011; Yang et al., 2010), while another strong claim is that video games are good for learning (Gee and Morgridge, 2005; Gee, 2003; Shaffer et al., 2005).

Thirty-six learning principles that make games good for learning are identified by Gee (2003). Five of those principles are distinguished by Hainey and Connolly (2011): interaction, production, risk taking, customisation and identity.

Malone (1980) emphasises how important intrinsic motivation is in learning improvement. Shabanah (2014) defends that computer games are also based on intrinsic motivation since players are willing to spend significant amount of time learning the rules of the game merely for the benefit of playing. Thus, the act of playing computer games is not motivated extrinsically but rather intrinsically since the players engage for entertainment reasons and not for rewards (Shabanah, 2014). Dempsey et al. (1993) have also stated that "games result in significantly higher levels of motivation, reduce training time and may improve retention of what is learned" (Dempsey et al., 1993).

Educationalists (e.g., Betz, 1996; Gee, 2003; Gredler, 1996; Kafai, 1994; Malone, 1981; Prensky, 2001; Rieber, 1996; Squire, 2003) have long proposed and investigated the benefits of the application of computer games to learning.

Many of the features that motivate the use of games in an educative context can be found in Shabanah's work. Shabanah's research is based on the work of Garvey (1990) who defines games as a "pleasurable, spontaneous, and voluntary" activity and Gee (2003) who uses the term "the cycle of expertise" to describe the enjoyable process of gaining knowledge by playing.

Shabanah also claims that computer games simplify evaluation based on the work of Thiagarajan (1978) (as cited in Shabanah, 2014) who states that games can be used for evaluation of knowledge as in this context it is "obviously superior to any paper-and-pencil test and is easier to administer."

Also, based on later work from Thiagarajan (2003), Shabanah claims that computer games add an emotional element to the interaction as events that evoke strong emotive responses such as games, simulations and role play are beneficial to the long-lasting learning (Thiagarajan, 2003), (Shabanah, 2014)²⁵.

According to Magnenat-Thalmann, games became a great medium not only for entertainment but also to obtain and share information and knowledge (Magnenat-Thalmann and Kasap, 2009). The growth of the games industry could be attributed to their popularity among people, most of which have access to at least one platform due to the technological advances and affordable devices they can use to play (e.g. smartphones) (Korre, 2012).

Serious games could appeal to the present day's digital native generation and be a solution to bridge the gap between their ever-online lifestyles and the somewhat static educational environments (Arnab et al., 2012). According to Gee (2007), the characteristics that games possess such as interactivity, customization, strong identities, well-ordered problems, pleasant frustration, built around the cycle of expertise, "fair" and "deep" suggest that video games can be excellent learning platforms (Gee, 2007; Blanchard et al., 2012). The same notion is shared by Ulicsak who claims that digital games can be used as a teaching tool because they are immersive, interactive and engaging (Ulicsak, 2012).

Abt identifies that one of the key considerations for SGs adoption is not just their effectiveness, but their cost effectiveness (Wilkinson, 2016). This notion is also shared by other researchers who support that SGs are cost effective, safe and meaningful to today's learners and produce deep long-term learning (Corti and Gillespie, 2015; Squire and Jenkins, 2003; Susi et al., 2007).

²⁵ Some of the work cited by Shabanah could not be accessed by the author.

Jennett also claims that the increased interest in SGs is attributed to their potential of providing an engaging context for learners (Jennett et al. 2008). It is also argued that through SGs the users can improve their perception, attention and memory as they remain strongly engaged (Tramonti, et al., 2014). Another advantage of SGs is the magic circle; the magic circle is a voluntary activity that occurs in a safe place (Linser et al., 2008).

In addition to obvious advantages, like allowing players to experience situations that are otherwise inaccessible or even impossible in real life due to safety, cost, time, etc. (Corti, 2006; Squire and Jenkins, 2003), it is argued that SGs have other advantages as well such as the development of different skills. Thus, the purpose of the SG should be taken into consideration when designing such a system as not all games are ideal for all purposes (van Eck, 2006; Susi, 2007). Computer games in general are often accompanied with negative characteristics such as addiction. However, there are studies showing that for the same curriculum, the students could acquire the knowledge and skills much more efficiently when taught using a computer game versus a conventional teaching method (Seng and Yatim, 2014). It should be noted though, that more research needs to be done in order to determine if and to what extent the curriculum affects these results.

A series of meta-analyses shows that SGs, computer games for teaching or training and GBL – the last two are subcategories of SGs -- are broadly more effective than traditional methods, although the quality of studies is variable which in turn affects the reliability of the results.

Wolfe's meta-analysis (1997), in an examination using seven studies on computer games for teaching strategic management, found that the results were in favour of the computer game based method showing improved

learning outcomes and knowledge gains compared to traditional learning methods (Wolfe, 1997).

Hays, (2005) conducted a meta-analysis of 48 empirical studies, 26 review articles and 31 theoretical studies focusing on the instructional effectiveness of games. The objectives of his research were to review the empirical research on the effectiveness of games and provide conclusions and suggestions for their use. The conclusions from his research were: the empirical research on the instructional effectiveness of games is fragmented with methodological flaws and ill-defined terms; just because one type of SGs work for a specific purpose and a specific audience does not mean that all games will work for all purposes and all audiences; debriefing and feedback can make instructional games more effective; there is no evidence that games are the ideal instructional medium for all purposes; and the players/learners can focus better on the instructional information when they are provided with information on how to use the game (e.g. tutorial).

Vogel's meta-analysis (2006) of thirty two studies showed "significantly higher cognitive gains and better attitudes toward learning" for those who used games or simulations compared to those who used more conventional teaching methods (Vogel et al., 2006).

Sitzmann's meta-analysis (2011) showed that the interactive cognitive complexity theory suggests that computer-based simulation games were more effective than conventional instructional methods because they manage to engage the learners with cognitive and affective processes. The examination of the instructional effectiveness of these games showed that the procedural knowledge was 11% higher, the post training self-efficacy 20% higher and the procedural knowledge 14% higher than the control group,

where trainees were taught using traditional methods. A very important observation made by Sitzmann was that there is strong evidence of publication bias²⁶ in games research (Sitzmann, 2011).

Connolly's et al. (2012) systematic literature review examined the potential positive effects of games and SGs on users of 14 years old and over with respect to learning, engagement and skill enhancement. The search identified 129 empirical evaluations. The findings of the review showed that playing video games is linked to a series of cognitive, behavioural, perceptual, motivational and affective impacts and outcomes with the most frequent being knowledge acquisition and affective and motivational outcomes (Connolly et al., 2012).

Pieter Wouters's et al. (2013) meta-analysis focused on the cognitive and motivational effects of SGs. Consistent with their hypotheses that SGs affect the motivation and cognitive processes, they found them to be more effective in terms of learning ($d = 0.29$, p less than 0.01) and retention ($d = 0.36$, p less than 0.01) than conventional instruction methods. The case was not the same for motivation though ($d = 0.26$, p greater than 0.05). It is worth mentioning at this point that the reported effect sizes²⁷ (d) are considered according to Cohen as "small" meaning that the relative size of the effect is rather small. Additional analyses revealed SGs users who had many training sessions, worked in groups and had additional instruction methods learned more compared to people taught with traditional methods (Wouters et al., 2013).

²⁶ "Publication bias is often referred to as the "file drawer problem" and occurs when the probability that a study is published is dependent on the magnitude, direction, or significance of a study's results (Begg, 1994)." (Sitzmann, 2011)

²⁷ Effect size is a standardised, scale free measure of the relative size of the effect of an intervention which quantifies and emphasises the size of the difference between two groups (Coe, 2002). More on effect size in Methodology chapter.

Clark's et al. (2015) systematic review and meta-analysis focused on comparisons of game versus non-game conditions and augmented versus standard games. The results from these comparisons revealed that digital games enhanced student learning significantly compared with non-game conditions. Also, additional analysis indicated that the effects varied across different game mechanics, visual and narrative characteristics which highlights the importance of design beyond the medium (Clark et al., 2016).

A systematic literature review by Boyle et al. (2016) focused on 143 papers with the most frequently occurring outcome for games being knowledge acquisition. The importance of a systemic programme for empirical work was also highlighted on the examination of which game features are most effective in promoting engagement and supporting learning; this is a focus of the research presented in this thesis.

While the above meta-analyses are quantitative, qualitative meta-analyses were also conducted by Ke (2009) and Vlachopoulos and Makri (2017). Ke's analysis included 89 gaming studies with one of the most important findings being that computer games can be used to develop higher order thinking skills. She also found that motivation and attitude were improved by GBL across different cohorts and domains (Ke, 2008). Vlachopoulos and Makri's (2017) literature review focused on the effect of simulations and games on achieving specific learning goals. The results from this review show that games and simulation have a positive impact on learning and that the learning outcomes from the integration of games into the learning process are cognitive, affective and behavioural.

This does not support that video games are the solution all by themselves. It depends on how they are used and what sorts of wider learning systems they are part of (Gee and Morgridge, 2005).

According to Hays (2005) there was a lack of evidence supporting an across the board use of games for instruction although based on recent reviews this might have changed. The findings from his work show that because GBL worked effectively for a specific domain or under a specific context, that does not mean that it will be effective under a different domain or context. That leads to exploring some of the drawbacks of SGs and similar concepts.

Although most focus on the positive effects of SGs, Wouters mentions that there is a school of thought that games with narrative put a cognitive load on the player/learner thus distracting them from the focus which is the learning content (Wouters et al., 2013). It has also been suggested that games with educational context strongly resemble digital exercise books, without utilising the characteristics of computer games (Kiili, 2005). That being said, Shaffer et al. (2004) highlight that many educational computer games are developed without the support of an underlying body of research, while Virvou et al. (2005) call attention to the fact that "the marriage of education and game-like entertainment has produced some not-very-educational games and some not very entertaining learning activities" (Shaffer et al., 2005; Virvou et al., 2005). Also, Sanford et al. (2015) found in their study that when designing SGs for children and younger population, designers should consider the experiences, expectations and perceptions of gamers so the games can be more effective. (Sanford et al., 2015). This is a particularly important observation acting as one of the starting points for the research presented in this thesis even though the focus is on adults. How the information is presented and what type of interaction is preferred is of high importance as

different types of games can have different effects (Gee and Morgridge, 2005).

Another justification for the limited integration of SGs in the learning process is the lack of tools for tracking and assessing the player/learner (Elborji and Khaldi, 2014).

According to Hainey et al. (2011), most of the disadvantages associated with GBL, focus on the lack of empirical evidence supporting GBL, destructive behaviour and attitudes (e.g. aggression, gender bias, immersion effect causing the player to alienate (Rosas et al., 2003) and logistics, cost disagreements and misconceptions around games (e.g. coverage, teachers resistance to new technology, software-hardware compatibility, curriculum inflexibility, limited budget, lack of supporting material etc. (Baek, 2008; Rosas et al., 2003) (as cited in Hainey et al., 2011).

Of course, since SGs are a subcategory of computer games, thus negative aspects associated with video games migrate to SGs. According to Susi et al. (2007), games may have a negative impact on the player. Those impacts may result in health issues such as headaches and repetitive strain injuries among others, psycho-social issues such as depression, social isolation, increased gambling and substitute for social relationships, and the effects of violent computer games such as aggressive behaviour and negative personality development. Connolly and Stansfield (2007) though, highlight that there is no general agreement on the long term effects of violence on game players (Connolly, Stansfield and Hainey, 2008). Griffiths (2002) point out that the negative effects that are usually correlated with games involve excessive users of computer games. It is worth mentioning that most of the studies on the negative effects of games focus on adolescents and not adult players.

2.3.2 Mobile Serious Games

The latest generation of mobile devices has the technical and interactive characteristics to support more complex applications. The rapid technological advancements in the field of mobile communications and devices have enabled more sophisticated functions than mere texting or calling. The early mobile devices did not have many capabilities and desktop/laptop computers were by far more capable of delivering information. Today, mobile devices have almost the same technological capabilities with personal computers (PC), making them a notable new medium for research.

Sánchez and Olivares (2011) claim that the benefits of using mobile devices for educational purposes have been pointed out by several researchers (Park, 2011; Sánchez and Olivares, 2011; Csete, 2004). The nature of these devices allows learning virtually everywhere, i.e. on the street, in the subway, on the bus etc. (Salinas and Sánchez, 2006) (as cited in Sánchez and Olivares, 2011) and, thus, creating a new era for technology-enhanced learning by allowing the learning experience to continue across environments (Chan et al., 2006).

Researchers and practitioners alike have also pointed out the advantages of the lower cost of mobile devices (Park, 2011).

Most recently, mobile designers have started integrating game mechanics and game design thinking in order to make mobile applications more playful and engaging to use. The added game play offers new opportunities for transferring skills and knowledge (Doumanis and Smith, 2015).

Mobile serious games (MSG) are a relatively new extension of SGs that run on mobile devices. This research focuses on MSGs on smartphones.

2.3.2.1 Mobile serious games and learning

While the focus of the research presented in the thesis is usability, it is worth mentioning briefly how and why mobile devices are used to foster learning because of the key role mobile learning plays in the development of SGs.

Jaldemark et al. (2017) report that due to the characteristics of mobility, connectivity and context sensitivity, mobile devices offer new opportunities for learning (Rouillard et al., 2014). The increased shift to mobile learning in the last years affected all educational levels. Mobile learning is often paired with SGs for higher education (Vlachopoulos and Makri, 2017) and with playful GBL for primary school (Hainey et al., 2016).

The distinctive technological characteristics of mobile learning deliver positive pedagogical affordances. Seven features of mobile devices that allow for their use in and out of the school context are summarised by Pea and Maldonado, (2006); these are “portability, small screen size, computing power (immediate starting-up), diverse communication networks, a broad range of applications, data synchronisation across computers, and stylus input device”. Klopfer and Squire (2008) mention that the most frequently reported characteristics of mobile learning are “portability, social interactivity, context, and individuality” with portability as the most distinctive feature that automatically sets apart mobile learning from other learning methods (Klopfer and Squire, 2008; Park, 2011).

In the work of George and Serna (2011), it is argued that mobile devices should form a set of diverse platforms by being integrated in learning systems globally. This point introduces new challenges; at a higher level, that the game content should adapt to the learner experiences and at a lower level, that the platforms should be managed dynamically (Balme et al., 2004) to increase collaborative and original aspects (Rouillard et al., 2014).

Hylén (2017) argues that there are two reasons for using mobile devices for adult learners. First is that mobile devices are more affordable than information and communication technologies (ICT) equipment available at classrooms and that they provide access to a wider range of learners. This has been the subject of various research projects such as the European project "MyMobile" that supports the idea of lifelong learning as a key concept of the European information society. Second is that mobile phones and social media encourage a learner-orientated approach (Hylén, 2017).

As Kukulska-Hulme and Pettit, (2006) note, the availability of mobile technologies increases the importance of lifelong learning and adult learners are considered to be lifelong learners (Manganello et al., 2013); thus mobile devices can be a great medium for lifelong learning (Deniozou, 2016).

Although one can argue that the lifelong learning was important before, mobile technologies can render lifelong learning easier.

Many empirical studies showed that the ownership of the mobile device involved the learners in the learning process (Park, 2011).

Even though technically personal digital assistants (PDAs) and tablets are mobile devices, this review focuses mainly on smartphones. This is due to the differences among smartphones, PDAs and tablets in terms of specifications,

i.e. size, screen size, processor, memory, resolution, connectivity etc., and the fact that smartphones are more widely used.

The technology of mobile devices evolves rapidly. There have been many studies in the past involving mobile and handheld devices, but a significant part of them is now redundant and needs updating. PwC²⁸ cites several key factors in growing the video game industry, one of which is mobile phones with relatively large screens that can download games with sophisticated graphics.

Usability engineering

2.4 Introduction to Usability engineering

With the increased complexity of new technologies and a wider part of the population being affected, usability testing becomes rather significant to HCI and UI design. Products that may be otherwise useful, risk failure if users cannot interact with and fully engage due to UI failures. (Ger et al. 2012)

Prior to evaluating the usability of either an application or a product, it is crucial to define what usability is. Even though there are numerous definitions of usability, there is one that stands out as widely accepted:

'The efficiency, effectiveness and satisfaction with which specified users can achieve specified goals in particular environments' (ISO, 1998).

²⁸ PwC stands for PricewaterhouseCoopers

Usability testing is used to measure the effectiveness and efficiency of a product that is used by specific users under a specified context of use.

Even though there are numerous definitions for usability, the common theme derived from the majority is that a product/application has multiple aspects that can affect user interaction (Gould, 1995).

Through usability engineering, it can be ensured that heterogeneous populations will be able to interact more easily with various applications, although not all applications can be measured with the same usability tools or methods (Moreno-Ger, 2012). The instruments that a usability engineer should use to evaluate a product depend greatly on its context of use as well as the users target group.

2.4.1 Usability and mobile devices

Even though laptops and tablets are technically mobile, mobile devices are most likely to be carried by people for the most part of the day. Rapid advances in software but mostly hardware allow mobile devices with rather impressive performance to mimic that of a low range or average computer. That in combination with internet connectivity evolved the use of mobile devices from calls and messages to a plethora of other uses such as navigation, emailing etc.

According to Doumanis and Smith (2015), even though mobile devices nowadays have a great competence, their UI design is still based on the graphical user interface (GUI) used in desktop computers. This is not to be confused with modes of interaction such as the touch screen. The input methods mobile phones use nowadays such as touch require a more

compressed information architecture (IA)²⁹. The human fingertip combined with the limited screen makes the human-mobile interaction more challenging.

Since mobile devices have multiple uses and the screen size is limited, the importance of having direct access to features without sacrificing usability is highlighted. Also, due to the limited screen size, traditional desktop UI is deemed unsuitable. (Findlater and McGrenere, 2008).

Text input on small screens can be awkward and slow which in turn can discourage the user using the service/application (Waycott and Kukulska-Hulme, 2003). A recent comparison by Stanford University of speech and keyboard text entry for short messages in two languages (English and Mandarin Chinese) on touchscreen phones showed that speech recognition had an input rate of 2.93 times faster (153 vs. 52 words per minute (WPM)) for English and 2.87 times faster (123 vs. 43 WPM) for Mandarin Chinese than the keyboard (Ruan et al., 2017). This came to emphasize the need for a more efficient way of interaction with speech being the closest alternative due to the advances in speech recognition technology and the rise of virtual assistants such as Amazon Alexa. Conversational interfaces such as spoken ECAs could be a viable alternative to error-prone text input. Some usability challenges of mobile applications though come with the nature of the device which is meant to be used in outdoor environments with varying light and noise levels. Simulations have been used as a result of the recognition that traditional usability laboratories and testing do not include the factors that affect mobile usability (Johnson, (1998), Graham and Carter, (1999)). This

²⁹ According to the Information Architecture Institute: “Information architecture is the practice of deciding how to arrange the parts of something to be understandable. Information architectures (IAs) are in the websites we use, the apps and software we download, the printed materials we encounter, and even the physical places we spend time in.” (IAI, n/a)

approach can be efficient for tackling the challenges related to studying mobile systems (Dahl, Alsos and Svanæs, 2009).

Zhang and Adipat (2005) and Kukulska-Hulme (2007) summarised the usability limitations of mobile devices as follows: 1) mobile device's attributes in terms of screen size, memory, performance and weight; 2) limitations in terms of software, applications and content, difficulty in adding applications and challenges in learning how to use a mobile device, lack of built-in functions; 3) connectivity, network speed, reliability; 4) issues that have to do with the environment where mobile devices are used such as using it outdoors with a variety of noise and light levels, security concerns, protectors against rain etc.; and 5) mobile device input methods are different from those for desktop computers and takes time to master. This increases the chances of an erroneous input.

Deb (2011) addressed these issues and highlighted that they must be taken into consideration when designing for mobile environments while hoping that manufacturers would address some of these issues. Indeed, within a decade the mobile industry has changed drastically. Most of the usability limitations from 2007 have been contained by technological advances. The introduction of touchscreen came with LG Prada and was popularised by Apple iPhone in 2007, the year that smartphones appeared on the market. Smartphones introduced the touchscreen, near field communication (NFC), wireless charging and later voice control, fingerprint scanning, face recognition, high-definition screen, multiple sensors such as heart rate sensors and gyroscope. Screen size, memory and performance have gradually become bigger and better compared to earlier devices. This changed the way people interacted with their mobile devices and introduced different usability challenges such as the error-prone finger typing. Also, dedicated services

such as Apple store, Google play and Android devices specifically allowed significant flexibility for the software and available applications. From 2011 onwards, fourth generation (4G) and later long-term evolution (LTE) technology and the supporting infrastructure allowed for better connectivity and network speeds. The only point that is still relevant nowadays is the use of mobile devices in public spaces especially since voice has been used as an input modality. This calls for a usability testing approach that simulates an environment where smartphones are commonly used which is why the author opted for a non-lab-based environment for conducting the main experiment.

2.4.2 Usability and serious games

According to Olsen et al. (2014) "Usability testing is an important, yet often overlooked, aspect of serious game development.". Games present some usability challenges due to their uniqueness on information presentation and the level of interactivity that they involve. Serious games aim to engage the users into activities that are not only entertaining but also purposeful. Usability testing for such media can be more challenging since more factors - such as cognitive resources- need to be taken into consideration. Adding game elements in an application does not guarantee that the desirable outcome will occur. While there is a plethora of usability testing methods and tools for productivity tools such as text editors and spreadsheets, analysing the usability of SGs presents unique challenges. Since SGs are fundamentally different from productivity tools, using the same instruments can be problematic (Moreno-Ger et al., 2012).

Usability is a non-transferable step of the game development process. It is relevant to the overall experience and can affect the players' interaction with the game. If the player cannot read the text clearly or has difficulties mastering the controls, that usability failure will distract from the game experience. Due to the introduction of SGs in many different domains and for many purposes that mainly have to do with acquiring a skill or knowledge, SGs present unique usability challenges. If the overall usability is poor, users' cognitive reserve, focus and attention may be redirected from the actual game to mastering the controls or interaction modalities (Olsen, 2014).

Mobile applications and mobile SGs specifically present unique usability challenges. Factors such as the noise levels and the light conditions can affect the way the player's ability to progress promptly. To ensure the player's comprehension and understanding, there is a need for additional communication modalities. Latest generation smartphones have plenty of sensors that can be harnessed to enable multimodal interaction with mobile applications. The use of multimodal communication in SGs can help players immerse in the scenario (Doumanis and Smith, 2015).

2.4.3 Usability and ECAs

Numerous aspects of ECAs (physical, behavioural etc.) have been evaluated empirically for several years. However, ECAs' interdisciplinary nature allows for further investigation on how they can create highly usable interfaces, as they rely heavily on technological advances such as the processing power, rendering techniques, graphic cards that are ever changing thus making previous research dated or even obsolete.

According to Weiss et al. (2015), the paradigms from ECA research till now show that there is not a universal way to evaluate ECAs. It is proposed, therefore, to design appropriate methods based on the purpose and for each application anew. The validity and reliability of the given methods are essential as they need to address the target at hand with minimum uncertainty.

Users' perspective on the task, interaction expectations and how the ECA is perceived are of major importance for comprehending the evaluation results. The social aspect that is introduced in the interaction by the ECA can increase attention as it can cause distraction from the main task thus making the evaluation for each application anew of great importance (Takeuchi and Naito 1995).

Weiss suggestions for evaluation is for at least two conditions to be compared by empirical evaluations of field tests in order to answer questions such as "Can an ECA improve the quality of the interaction for the given domain and task?"

In some cases, ECAs use spoken language to interact. Inherently they exhibit at least an anthropomorphic element which is the voice while there could also be a visual element such as a human face. Spoken dialog systems can offer an intuitive and natural way of interaction due to voice interaction. From the user's perspective, the ECA's capabilities can be enhanced by embodiment as the user could have increased expectations. An example would be believing that the ECA has social skills and cognitive function which should be mirrored in human-like communication behaviour. If such expectations are not met, user experience will be negative (Weiss et al., 2015).

Speech interaction with ECAs also needs to be evaluated for each scenario in order to produce a usable spoken multimodal system. There are parameters

such as speech recognition fidelity, application purpose and environment and channel stability that can affect the interaction (Bernsen, 1994; Weiss, 2015).

2.5 Summary

This chapter started by providing an overview of the history of interactional systems since their early days in 1960s. An introduction to conversational systems and voice enabled technologies was then provided as the predecessors of ECAs. The evolution of research on these systems revealed a trend of an ongoing attempt to make conversational systems as human-like as possible; an early example is the Turing test. Embodied conversational agents were a step to that direction.

The multidisciplinary nature of ECAs was then discussed along with the multitude of options that developers have while designing them. Building upon that, an ECA design model was introduced as an aid for developers depending on the purpose of the agent. Design decisions have the potential of affecting the interaction in a significant way. Apart from the obvious graphical representation choices, an agent can be designed to extract reactions from users in a deeper level. Embodied conversational agent designs can vary from highly anthropomorphic to very simplistic in all three levels of: persona, presentation and interaction.

A brief historical overview of ECAs was then presented aiming to introduce the field and call attention to the direction of the field towards agents with high human-likeness. This trend is due to technological advances that allow multiple input and output modes of interaction with the agent that allows for

higher information bandwidth in human-agent interaction. The social role of ECAs is then discussed as their human-like nature evoke social responses to users. Theories that are related with the social aspect of ECAs are addressed along with empirical work exploring them. The illusion of humanness effect is then discussed as a novel concept which is described as the unconscious effect humanoid ECAs have on users in relation to usability. In an attempt to establish a frame for evaluating the illusion of humanness, the levels of anthropomorphism and human likeness are then defined.

The following section introduced SGs and highlighted their possibility of being a meaningful to today's users' platform where the illusion of humanness can later be explored. Market research has illuminated a trend towards MSGs which makes research on the topic even more contemporary. Smartphone technology allowed for high quality graphics and processing power which allows the transfer of SGs from desktop to mobile.

The many theoretical advantages in favour of ECAs were explored, followed by reasoning to consider that ECA technology may also be effective in SGs.

Finally, the importance of usability is discussed along with usability challenges imposed by SGs, ECAs and mobile devices.

Although there is a growing pool of empirical data relevant to the effects of ECAs, there is still lack of empirical evaluations of the usability of ECAs on mobile devices which amplified the call for empirical research. Previous studies found that there is lack of empirical evidence on the impact of embodiment of conversational agents within mobile interfaces. Given the lack of evidence on the potential effect of ECAs on SGs, there is a major risk related to the introduction of ECAs in MSG applications. The evolution of mobile technology along with technological advances in the area of ECAs

also call for empirical research on the topic as a way of updating the literature by using contemporary technology.

The work addresses both the illusion of humanness evoked by humanoid ECAs and the effectiveness of the agents in serious games.

Chapter 3 Methodology

Introduction

The aim of this thesis is to produce empirical evidence on the effectiveness and users' perceptions of the impact of humanoid spoken embodied conversational agents on mobile serious game applications. The experiments detailed in this thesis were specifically designed and implemented to collect this empirical information.

In this thesis the usability and perceived persona of the embodied conversational agents (ECAs) was examined, the efficiency of task completion by the user was informally observed and the effectiveness of the interaction was documented and analysed with the purpose of producing a pool of empirical data as for the use of spoken humanoid ECAs (HECAs) in mobile serious games.

In total three empirical studies are presented. The pilot studies are reported in Chapter 4 and the main experiment is reported in Chapter 5.

3.1 System description and technology used

3.1.1 System description

Moneyworld

The application that is used in this thesis is called Moneyworld. Moneyworld is an application developed in 2012-2013 by the in-house team of the former Centre for Communication Interface Research (CCIR) at the University of Edinburgh where I started my PhD and was used for consistency with other findings from the lab at the time. The team who designed the application consisted of Nancie Gunson, Hazel Morton, Diarmid Marshall, Graeme Roy, Nick Anderson, Simon Doolin, Mervyn Jack and the author. Although I contributed to the development of the application, the concept and gameplay was decided by the team before I got involved as the application was not specifically designed for the evaluations reported in this thesis. The centre closed in 2014 and by using reverse engineering I adjusted the application for the remaining part of my PhD to fit the purposes of the evaluations.

The style of the application is described by the developers as a casual game. Casual games, according to Juul (2010), are easy to learn how to play, work in many different situations and fit well with many different players. In contrast with more sophisticated games, casual games require minimal training and have simple interfaces.

Moneyworld is a 3D interactive application where the user travels back in time in order to learn more about the old money system that was used in the

UK till the early 1970s. In this application, two photorealistic agents equipped with speech recognition are used. The participant partakes in a shopping experience using voice and mouse as input methods thus making the application multimodal.

First, participants are informed about the purpose of the experiment and then the introduction begins. In this introduction a female unembodied voice welcomes the user to the time machine chamber and introduces the concept of the application (Figure 16).

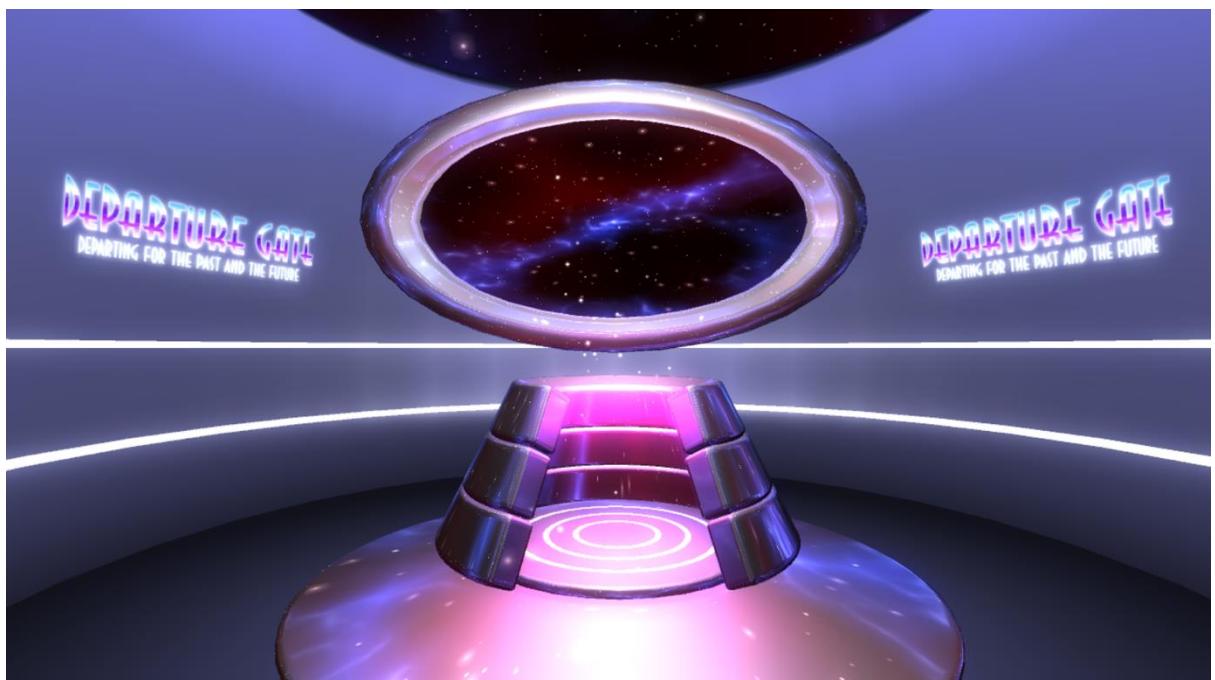


Figure 16 Introduction to Moneyworld, Time machine chamber.

After the time travelling, the participant is transferred to a corner store in the 1960s where the main interaction takes place. The virtual shop designed in this research is based on a typical 1960s corner shop with the items displayed behind the counter. Figure 17 shows the shop-keeper in the corner shop. The interaction starts with a tutorial by the same unembodied voice, introducing

the old money system to the participant. The same voice asks the user to identify three coins from the set and state the value in pence of each of them. After the review, the voice demonstrates how to use the coins in order to buy items.



Figure 17 Corner store layout with shopkeeper ECA.

After the introduction the application starts with a small tutorial on the gameplay delivered by another agent, Alex (instructor). Alex provides background information to the user on the currency and assistance when needed. In the 1960s, the currency used in Britain was an old monetary system based on pence, pounds and shillings. There were 12 pence in a shilling and 20 shillings in a pound. After the description, Alex asks the participant to review the coins via an understanding exercise through speech. Associated error recovery dialogue was included for instances where the user

was silent or answered with an incorrect response. Alex also tells the user which item to purchase in the shop. Figure 18 depicts Alex in the virtual portal within the shop.



Figure 18 Alex shown in the virtual portal.

After Alex's tutorial, she introduces the multimodality of the application, that of the coin submission tray. For the user to pay for the products in the shop, a virtual wallet is presented on the bottom of the screen with all the coins (see Figure 19).



Figure 19 Coin Tray.

During this part of the tutorial, Alex directs the user on the way to submit the coins prior to the shopping task:

"You'll see at the bottom of your screen that you've been given some of the old coins to use. When the shop-keeper tells you how much it is for each item, click on the coins, one at a time, to make up the required amount.

Please click on some coins now to move them to the tray." The game play is straightforward, the player is asked by Alex to buy a list of items one at a time from the shopkeeper. Once Alex dictates which item is to be bought, the virtual portal detracts, and the shopkeeper enters the shop greeting the user and asking how he can help to which the user is expected to respond verbally.

An example of this interaction:

Shop-keeper: "Hello. How can I help you?"

User: "I'd like a box of cornflakes please"

Shop-keeper: "Cornflakes. A nice, healthy start to the day. That's 1 and 9 please."

The user is then expected to submit the relevant coins with the most efficient combination of coins. The shopkeeper only accepts the correct amount of money asked for the item (Figure 20).



Figure 20 Coin submission.

In total, the user is given four items on their shopping list to 'buy' at the virtual shop and is given feedback after each item for correct payment made, efficiency of payment (payment made with the fewest number of coins), and efficiency of task (whether any additional help was required for each item on the shopping list).

3.1.2 Technology used

The application used throughout this thesis is an executable program developed in Unity. Unity is a game engine platform for developing interactive applications such as games. Unity can handle graphics, animation, sound, gameplay, scripting and the user interface. The manipulation of the assets is usually done through scripting, in this case C#. After the development stage the interactive application is exported as a standalone executable.

The assets used in this application such as the 3D furniture and items were developed using the open-source 3D authoring tool Blender3D. These assets are then exported as Autodesk FBX files that are then imported into Unity. The textures and images were prepared using Photoshop. The characters used in Moneyworld were bought from Rocketbox³⁰. The models came already rigged with skeletal bones that were then manipulated in Autodesk 3DSMax in order to create the animations.

The opensource library PocketSphinx is used to handle speech recognition. Pocket Sphinx is described as “a lightweight speech recognition engine, specifically tuned for handheld and mobile devices, though it works equally well on the desktop”³¹. Pocket Sphinx has been developed at Carnegie Mellon University over 20 years. In this research, Pocket Sphinx was compiled as a Windows DLL that is accessed during runtime. Another in-house DLL was developed as a plug-in with the purpose of being the bridge that connects Unity to the external Pocket Sphinx DLL.

³⁰ Rocketbox is an art studio that is specialized in 3D characters and animations

³¹ As found in: <https://github.com/cmusphinx/pocketsphinx>

Lip-synching is rather difficult to be executed in order to emulate humans as close as possible unless the development team has access to the multimillion pounds of equipment used by big game studios such as Rockstar. There are though affordable software packages that provide lip-synching and face expressions by mapping the lip-movements and face expressions of a human to the 3D model. Unfortunately, the results are not seamless. The software used for this research is called Faceshift. Faceshift uses a depth-sensitive Microsoft Kinect camera that “recognises” the features and face movements of the actor that are then mapped onto the 3D character in the form of blend shapes. This method, although the most fitting in terms of budget, does not provide the most detailed facial animations.

For the quantitative collection along with the technographic survey, Survey Monkey was used.

3.2 Experimental design and experimental procedure

3.2.1 Experimental design

For the purposes of this research an experimental mixed methods approach

was chosen. The advantages of experimental research over observational

research is the level of control over variables which makes it easier to draw

conclusions about causal relationships from the data (Jack, et al., 2005).

Quantitative research requires the collection and analysis of numerical data,

whilst qualitative research involves experiential or narrative data (Hayes, et al.,

2013). A combination of quantitative and qualitative methods within the

same study is referred to as mixed methods analysis (Wisdom, et al., 2012;

Creswell and Plano Clark, 2006). Researchers noticed that the strengths of

each single method used in combination could address the biases of other methods since all approaches have limitations.

More specifically the mixed methods procedure that was followed is a concurrent procedure which refers to the collection and analysis of both quantitative and qualitative data and the merging of the information during the interpretation of the overall results (Creswell, 2011). Quantitative data were collected by administering standardised questionnaires to the participants (participant responses to a measure) and qualitative data were collected by a short exit interview after the completion of each task (semi-structured interview with both open and close ended questions from which themes among the participants are derived).

The reason behind choosing this method is that while employing the practices of both quantitative and qualitative research, the mixed methods approach allows for detailed exploration of a complex phenomenon. Through the triangulation of data sources, researchers can expand understandings or confirm findings from one method to another (Creswell, 2011).

3.2.2 Data collection method

The experiment approach followed throughout this research consists of a contrastive study where two versions of the application are experienced by the participants. The two versions differ from each other in a design characteristic.

In usability experiments of this nature, a repeated-measures design is preferable due to advantages over between-subjects. Repeated-measures design, also known as within subjects' design is a method in which the

researcher manipulates the independent variable by using the same participants for all conditions. The advantages of repeated measure design are that differences between conditions can only be caused by two situations 1) the way the participants are handled 2) any other factor that affects the participants' performance from one time to another with the latter factor being minor compared to the first (Field, 2013). Other advantages are that this method allows comparisons to be made for each participant (Landauer, 1988) and that it requires fewer participants compared to between subject designs.

The order of experience in this design is balanced across the cohort in order to avoid biases (habituation or fatigue effects) introduced by the order in which the participants are experiencing the designs (Preece, et al., 2002). Another bias in experimental design that needs to be tackled is that of researcher bias. To avoid researcher biases the whole procedure is standardised with the researcher giving minimum input and following the same scripted procedure for every participant. See Appendix A for the full scripted procedure followed by the researcher. That allows for the data to be used for statistical analysis (Whiteside, et al., 1988; Coolican, 1994).

Participants' attitudes towards the ECAs and the system were measured using questionnaires completed after experiencing each version of the service. Also, subjective attitudes to the experiences were collected through exit interviews after each version while the researcher is making direct observations regarding the participants' behaviours during the experiment.

A 2x2 factorial experimental design was adopted for the usability evaluations as the applications had two different factors each constituted by two levels. Based on the experimental design, the participants were divided into equal and

balanced groups with all group subjects experiencing both design options as shown in Table 2.

2x2 design		Factor 2	
		1	2
Factor 1	A	A1	A2
	B	B1	B2

Subject 1	➤	A1	➤	Standardised questionnaires	➤	B2	➤	Standardised questionnaires	➤	Exit interview
Subject 2	➤	A2	➤	Standardised questionnaires	➤	B1	➤	Standardised questionnaires	➤	Exit interview
Subject 3	➤	B1	➤	Standardised questionnaires	➤	A2	➤	Standardised questionnaires	➤	Exit interview
Subject 4	➤	B2	➤	Standardised questionnaires	➤	A1	➤	Standardised questionnaires	➤	Exit interview

Table 2 Within subject design (repeated measures) based on a 2x2 factorial design.

3.3 Evaluation Metrics

3.3.1 Quantitative data collection

All the questionnaires employed in this research employ a Likert format (Likert, 1932). In Likert scales the participants are presented with a stimulus statement, which is the attribute to be measured, followed by an agree-

disagree scale. Coolican, (1990) has described the advantages of this format as follows:

Participants prefer the Likert scaling technique because it maintains their direct involvement in the process and is "more natural" to complete.

The Likert scale has been shown to be effective in measuring changes over time.

The Likert technique has been proven to have a high degree of reliability and validity.

3.3.1.1 **Usability Metrics**

Even though there are many metrics for usability measurement, part of usability engineering is to find the right metric among them that fits the specific aims of a research (Landauer, 1988).

Previous research (Dutton, et al., 1993; Jack, et al., 1993; Love, et al., 1992) has identified salient attributes of the perceived usability of interactive systems. The result of this research is the CCIR MINERVA usability questionnaire that was chosen for this research which has been developed and tested as a tool for assessing users' attitudes (McBreen, 2002; Gunson, et al., 2011). The validity of the questionnaire was confirmed by experimental work (Jack et al., 1993). In order to verify that the questionnaire predicted different user satisfaction and usability for different levels of speech recognition accuracy, a large-scale experiment was conducted; 256 participants were divided into four conditions (group 1: 85% accuracy; group 2: 90% accuracy; group 3: 95% accuracy; group 4: 100% accuracy). Analysis of variance showed that a statistically significant ($p < 0.001$) effect was detected between different

groups as seen in Figure 21. Also, statistically significant differences ($p < 0.001$) were found between group 1 and group 4, group 2 and group 4, group 3 and group 4.

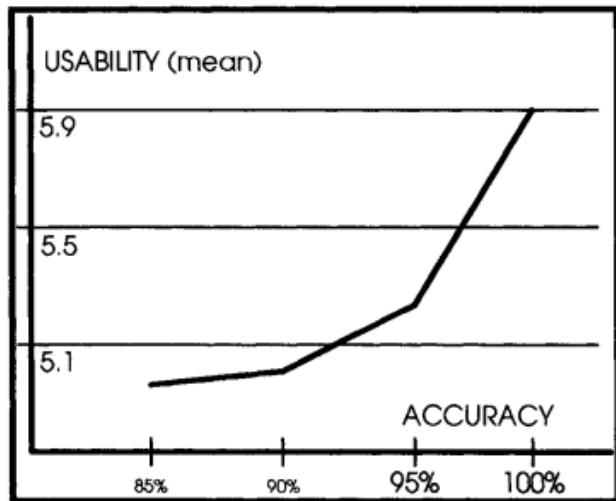


Figure 21 Usability versus accuracy of speech recognition.

The original set of questionnaire attributes was developed in order to assess automated telephone services, since then it has been adapted and has gone through a rigorous testing process through which has been proven to be a robust and reliable measure of usability for spoken dialogue systems and ECAs (Doolin, 2014; McBreen, 2002; Morton, et al., 2004).

The metric that has been used is widely accepted as a reliable tool for measuring usability and has been used in a large number of research experiments (Davidson, et al., 2004; Foster, et al., 1998; Larsen, 1999; Larsen, 2003; Morton, et al., 2004; Sturm and Boves, 2005; Weir, et al., 2009; Doolin, 2014). The questionnaire contains statements on cognitive issues (e.g.

concentration level required by users, stress levels while using the application), transparency and clarity of the system (e.g. ease of use and degree of complication), friendliness (e.g. enjoyment of use and perceived friendliness) and system performance (e.g. the efficiency of the application and users' preferences for a human agent).

The questionnaire is comprised of a series of 18 attitude statements, each statement is followed by a set of tick-boxes on a seven-point Likert scale (Likert, 1932; Rossi, et al., 1983) ranging from "strongly agree" through neutral to "strongly disagree" as seen in Figure 22.

I thought this service was too complicated	Strongly Agree	Agree	Slightly Agree	Neither Agree nor Disagree	Slightly Disagree	Disagree	Strongly Disagree
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 22 Sample Attitude Statement and 7-point Likert Scale.

In order to tackle the problem of response acquiescence (the general tendency for respondents to agree with the statement offered) and adding to the robustness of the questionnaire, the statements are balanced, positive and negative. For the purposes of analysis, the responses are converted into numerical values ranging from 1 (most unfavourable) to 7 (most favourable) allowing for the polarity of the statements. As an example, a "strongly agree" response to a negative statement is converted to a value of 1. Normalised scores over 4 exhibit positive attitudes while scores below 4 negative attitudes, with 4 representing neutral (Table 3). Another action taken to

ensure reliability of the acquired data is the randomisation of the order of questions (Goodhue and Loiacono, 2002).

Attitude Statement Type	7-Point Likert Scale Categories						
	Strongly Agree	Agree	Slightly Agree	Neither Agree nor Disagree	Slightly Disagree	Disagree	Strongly Disagree
	Numerical Values Assigned						
Positive	7	6	5	4	3	2	1
Negative	1	2	3	4	5	6	7

Table 3 Summary of Numerical Values Assigned to each of the 7-Point Likert Scale Categories.

Once the polarity of the responses is normalised, a mean score of these numbers across all the Likert items is calculated for each participant to measure the overall attitude towards the application. A measure of the overall attitude towards the application can then be acquired by averaging all the participants' questionnaire result. The mean scores for individual statements can also be investigated to emphasise any aspects of the design that stands apart as successful, or aspects that require improvement. Finally, the results can also be analysed according to demographic groupings of participants (age, gender etc.) and any significant differences between groups can then be identified.

The CCIR MINERVA usability questionnaire has been adapted in order to foster the needs of these experiments by substituting the name of the application (Moneyworld) into each question as listed in Table 4.

Usability Questionnaire Statements
1. I found Moneyworld confusing to use
2. I had to concentrate hard to use Moneyworld
3. I felt flustered when using Moneyworld
4. I felt under stress when using Moneyworld
5. I felt relaxed when using Moneyworld
6. I felt nervous when using Moneyworld
7. I found Moneyworld frustrating to use
8. I felt embarrassed while using Moneyworld
9. While I was using Moneyworld I always knew what I was expected to do
10. I felt in control while using Moneyworld
11. I would be happy to use Moneyworld again
12. I felt Moneyworld needs a lot of improvement
13. I enjoyed using Moneyworld
14. I thought Moneyworld was fun
15. I felt part of Moneyworld
16. I found the use of Moneyworld stimulating
17. Moneyworld was easy to use
18. I thought Moneyworld was too complicated

Table 4-Usability attributes.

3.3.1.2 Attitudes towards ECAs

The second questionnaire that has been used for this research is also a validated metric for assessing the agent's persona called Agent Persona Instrument (API) (Baylor and Ryu, 2003). The API is a validated instrument for measuring pedagogical agent persona as perceived by the user in applications with educational context. The original instrument is comprised of 25 items with a 5-point Likert scale, with 1 for "strongly disagree", 2 for "disagree", 3 for "neutral", 4 for "agree", and 5 for "strongly agree" as seen in Table 5.

Attitude Statement Type	5-Point Likert Scale Categories				
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
	Numerical Values Assigned				
Positive	1	2	3	4	5
Negative	5	4	3	2	1

Table 5 Summary of Numerical Values Assigned to each of the 5-Point Likert Scale Categories.

Similarly, to the usability questionnaire, the statements were randomised, and the same procedure is followed thereafter.

This questionnaire presents four key factors for agents to be perceived as person-like (Table 6): credibility, engaging, human-like, and the capacity to facilitate learning as listed below:

- Facilitating Learning: These 10 items are related to how well the agent helps the student through the learning process.
- Credible: This factor consists of 5 questions related to the credibility and believability of the agent and its advice for helping the learner understand the learning content.
- Human-like: These 5 items address the agent's behaviour and emotional expression in terms of its naturalness and personality.
- Engaging: This factor consists of 5 questions that relate to how entertaining and enjoyable it is for the learner to work with the agent.

The four key factors are further categorised into two latent variables; Informational Usefulness (facilitating learning and credible) and Affective Interaction (human-like and engaging). Arguably, due to the focus of the research, the predictors that fall in the Affective Interaction category are of bigger importance since learning is not the focus of this work and was consequently not assessed. From all 24 items, one was excluded since it was not consistent for both experimental conditions ("The agent's movement was natural").

<i>Facilitating Learning</i>	<i>Credible</i>	<i>Engaging</i>
<p>1. The agent led me to think more deeply about the presentation.</p> <p>2. The agent made the instruction interesting.</p> <p>3. The agent encouraged me to reflect what I was learning.</p> <p>4. The agent kept my attention.</p> <p>5. The agent presented the material effectively.</p> <p>6. The agent helped me to concentrate on the presentation.</p> <p>7. The agent focused me on the relevant information.</p> <p>8. The agent improved my knowledge of the content.</p> <p>9. The agent was interesting.</p> <p>10. The agent was enjoyable.</p>	<p>1. The agent was knowledgeable.</p> <p>2. The agent was intelligent.</p> <p>3. The agent was useful.</p> <p>4. The agent was helpful.</p> <p>5. The agent was instructor-like.</p> <p><i>Human-like</i></p> <p>1. The agent has a personality.</p> <p>2. The agent's emotion was natural.</p> <p>3. The agent was human-like.</p> <p>4. The agent's movement was natural.</p> <p>5. The agent showed emotion.</p>	<p>1. The agent was expressive.</p> <p>2. The agent was enthusiastic.</p> <p>3. The agent was entertaining.</p> <p>4. The agent was motivating.</p> <p>5. The agent was friendly.</p>

Table 6 The API (Agent Persona Instrument).

There are not many metrics specifically designed for ECAs. Although other metrics were considered for evaluating attitudes towards ECAs with closest being the conversational agents scale (CAS) (Weiss, et al., 2015), and Attitude Toward Agent Scale (ATAS) (Van Eck and Adcock, 2003). The first metric, although fitting with the nature of the experiment, was not validated in English by the time the experiment was conducted; the second metric is focused more on the pedagogical aspect of the agent. On the other hand, API is a standardised metric with assessed Cronbach's alpha of items for each factor that indicated that the items showed very reliable consistency within the factors (Baylor, 2005).

3.3.2 Qualitative data collection

Qualitative data are important as they provide an insight on participants' subjective attitudes to the experiences of using the different versions of the application. This practice allows the researcher to further understand the user's attitude towards the system and justify any statistical differences that arise from the Usability Questionnaire responses.

For the collection of qualitative data, carefully designed post-experience interviews were employed. The researcher needs to carefully word the questions included in the questionnaire in order to extract exactly how a user feels towards the system.

Also, researchers made direct informal observations about the behavior of the participants during the experiment, that provide information on non-verbal reactions to the system.

3.4 Statistical Analysis of Experiment Data

3.4.1 Hypothesis testing

In order to conduct a statistical analysis, the null hypothesis (H_0) needs to be defined by the researcher. The null hypothesis means that there is no difference in the dependent variable among the conditions of the independent variable or that the mean difference will be 0 when testing for differences between conditions. The null hypothesis is rejected when the result becomes statistically significant by running statistical tests. The definition of how strong the result must be in order to be defined as statistically significant depends on the significance level (α) adopted for the test. For a specified value of α , the test is:

If $p < \alpha$, reject H_0 ; otherwise do not reject H_0

Where p is defined by:

$$P = P(\text{data at least as extreme as the observed data} | H_0)$$

The smaller the value of p is, the stronger the evidence is to reject H_0 (Jack, et al., 2005).

Conventionally, within HCI, the value of α is set to 0.05 with p values lower than that described as statistically significant.

Another important factor that must be considered is whether the hypothesis is one-tailed or two-tailed. In a one-tailed hypothesis, apart from the null hypothesis H_0 a second hypothesis is introduced H_1 . When null hypothesis is rejected it means that there is evidence to support H_1 . One tailed hypothesis is employed when there is strong evidence that the departure will be towards only one direction.

In the case where strong theoretical evidence for a directional test does not exist, it is better to opt for a two tailed test where significance lies at the tail of a distribution curve. Since no strong theoretical evidence exists suggesting a shift towards a specific direction, two-tailed tests were used for all the experiments described in this thesis. The type of data collected for the analysis also determine the statistical tests that must be used. Data are categorised as nominal, ordinal and interval. Definitions and examples of each can be found in Table 7.

Nominal:	Values indicate different named categories; one category is not higher or better than another Examples: Country of birth, sex, eye color, marital status, primary mode of transportation
Ordinal:	Finite number of categories with ordering Examples: Response to treatment (much improved, somewhat improved, same, worse, or dead), socioeconomic status, cancer staging, different doses of a drug as a treatment variable
Interval:	Variable with ordering but also a meaningful measure of the distance between categories Examples: Temperature in Celsius or Fahrenheit, number of days stayed in the hospital, score on an IQ test, score on a quality of life measure
Ratio:	Interval scale with a true zero Examples: Temperature in Kelvin, height of a person, serum cholesterol, drug concentrations, most laboratory test values

Source: Yang Y, West-Strum D: *Understanding Pharmacoepidemiology*
www.accesspharmacy.com
 Copyright © The McGraw-Hill Companies, Inc. All rights reserved.

Table 7 Data categories.

The Likert scales used throughout this thesis were regarded as interval data as they were intended by the original authors of the scale to be used in this way, and those authors performed analysis as if it were interval data.

When it comes to Likert scales and Likert type data, there is no clear answer as to if they should be analysed parametrically or non-parametrically. The answer can be different for Likert scales than it is for Likert items and that adds to the confusion. One side maintains that Likert scale data can be analysed as interval data. According to Altman (1991), "parametric methods require the observations within each group to have an approximately normal distribution ... if the raw data do not satisfy these conditions ... a non-parametric method should be used" Parametric analysis of ordinary averages of Likert scale data is justifiable by the Central Limit Theorem if the sample size is large enough "for reasonably large samples (say, 30 or more observations in each sample) ... the t-test may be computed on almost any set of continuous data" (Jekel, et al., 2001). Thus, analysis such as t-test, ANOVA and regression procedures can be applied (Capod, 2017). This is also

supported by Lubke and Muthén (2004) who found that true parametric values can be found in factor analysis when using Likert scale data if the assumptions are met (skewness, number of categories, etc). On the same note, the fact that the accurate p-values can be returned from the F tests in ANOVA on Likert items under certain conditions was supported by Glass et al. (1972).

The other camp maintains that the intervals between values are not equal since they are ordered categories, thus any parametric operation applied to them is invalid. Consequently, only non-parametric statistics should be used on Likert scale data (Jamieson, 2004; Grace-Martin, 2017).

Nonetheless parametric statistical tests, such as t-tests and ANOVA's were used to analyse the data. Ordinal data do not always guarantee a normal distribution but if the assumption of normal distribution (which is the primary assumption of parametric tests) is correct, the use of parametric tests on ordinal data is possible. If the data do not depart substantially from normal, the sampling distribution remains almost the same. The preference of parametric tests over non-parametric derives from the fact that they are more versatile and powerful (McBreen, 2002).

Furthermore, regarding the multiple regression since the dependent variable is an aggregated score of all the Likert items in the Likert scale it can be treated as continuous and analysed as such with linear regression. To support the decision to analyse the data parametrically, a further exploration of the data was conducted. With the purpose of determining if the data are normal, the following tools were used:

- Histograms
- Stem and Leaf plots

- Box plots
- P-P plots
- Q-Q plots
- Skewness and kurtosis

As mentioned earlier a repeated measure or within subject design was used in this thesis and the data were analysed as interval and parametrically. One of the most popular tests on interval data is t-test. A related samples t-test is a fitting method when comparing two interval dependent variables evaluated by the same group of people for both conditions and is used regularly in this thesis.

Another method for comparing sets of data that is often used in this thesis is an analysis of variance (ANOVA). ANOVA is a common method when more than two sets need to be compared although when used for only two sets, it gives essentially the same results as a t-test. ANOVA is ideal where multiple variables are present and for exploring interactions between these variables. A repeated measures ANOVA which is used in repeated measures designs, centres on the F statistic which measures the deviation from uniformity.

When an effect is present, the value of the F statistic is usually higher (McBreen, 2002).

This thesis explored some of the data using repeated measures ANOVAs first and compared these results with figures retrieved from t- tests. Repeated measures ANOVAs for comparing two groups gives essentially the same results as t-tests and both were used in this thesis although for the main experiment t-tests were preferred.

3.4.2 Statistical analysis for multiple linear regression

In order to answer the third research question “Which factors relating to the HECA’s persona attributes account for variability in usability, and to what extent?”, a multiple linear regression is used. Multiple linear regression analysis estimates the coefficients of a linear equation, involving multiple independent variables (IVs), that best predict the value of the dependent variable (DV).

Building a complex regression model with multiple predictors/variables/features can be a daunting task as it must be decided which predictors should be included in the models and which ones should be discarded. The way the predictors are selected and entered the model is of great importance due to the impact they have on the regression coefficients. Those coefficients depend on the variables in the model. When the predictors are uncorrelated the order in which the variables enter the model is of minor importance, but uncorrelated variables are rare in this type of research. The general rule regarding regression models is that the sparser the model the better. Therefore, one must be selective and have a decent sample size (Field, 2013).

There are multiple ways to do multiple linear regression such as cross validation, penalized methods or choosing variables based upon past research and/or theory (which is the ideal). In this research, there was no prior knowledge to select some variables based on previous research as no prior research looked on the relationship of the agent’s persona and usability. The predictors used in this research were informed by the nature and theoretical base of the experiment.

In this research, the chosen model was the OLS (ordinary least squares) full model with 9 items of the API questionnaire as predictors and the usability mean value for the shopkeeper agent and instructor agent respectively as the DV. The predictors chosen for the multiple regression belonged to the affective interaction category and were: "The agent has a personality", "The agents emotion was natural", "The agent was human-like", "The agent showed emotion", "The agent was expressive", "The agent was enthusiastic", "The agent was entertaining", "The agent was motivating", "The agent was friendly".

From the literature (Tibshirani and Hastie, 2016), it is known that sparser statistical models perform better and tackle the problem of overfitting. Thus, a reduction of complexity was achieved by selecting the IVs based on theory rather than using all 24 predictors. Another reason for not using all 24 items as IVs is for model interpretability; by removing irrelevant features a model is more easily interpreted.

3.5 Sample Size Justification

3.5.1 Sample size for t-test

Sample sizes are often dictated by the limitations in financial and technical resources as well as time but in order to get robust data a power analysis is favoured. Discount usability testing³² which is an alternative to higher cost

³² According to Nielsen "The "discount usability engineering" method is based on the use of the following three techniques: scenarios, simplified thinking aloud and heuristic evaluation" (Nielsen Norman Group, 1994).

usability testing does not provide rigorous data. According to Ghanam, (2007) "Discount methods tend to find superficial problems; thus, they usually are not suitable for in-depth usability studies. Discount usability testing is not a replacement of traditional usability testing but can be very advantageous compared to doing nothing." Even Nielsen who pioneered discounted usability argued that "When it comes to selecting usability methods, there are many parameters to consider, and many different scenarios. That's why both expensive and cheap usability methods make sense under the appropriate circumstances." (Nielsen Norman Group, 2007). Discount methods work well on pilot studies though.

Statistical power analysis takes advantage of the interdependent relationship among the variables in statistical inference (Significance level (α), power, sample size (N), effect size (ES)). When planning a research study, the sample size needs to be determined beforehand and for that reason a specified power for given α and ES needs to be determined (Cohen, 1992).

3.5.1.1 **Significance level (α)**

According to Cohen (1992), α represents the maximum risk taken so that the null hypothesis is not mistakenly rejected, also known as Type I error. Unless stated otherwise the typical value is equal to 0.5. When multiple hypotheses are tested, a more conservative value is recommended ($\alpha = 0.01$) in order to minimise the risk. For non-directional tests where the parameters can be either negative or positive, the α may be defined as two sided or one sided (Cohen, 1992).

3.5.1.2 **Power**

Power refers to the statistical power of a statistical test and is defined as the probability of rejecting a false null hypothesis. When the value of the effect size (ES) is different than zero, then the null hypothesis is false and the failure

to reject it also sustains what is known as a Type II error. For any given ES, N and α the probability of Type II error occurring is β thus power is $1-\beta$. A common value for power is .80 ($\beta = .20$), any value lower than .80 would impose too big of a risk for Type II errors to occur. A value over .80 would demand a larger sample size which could exceed the resources of the researcher (Cohen, 1992).

3.5.1.3 **Sample size**

The researcher needs to know the sample size needed in order to acquire the desired power for the defined α and hypothesised ES before conducting the experiment. The value of N increases proportionally to the power desired and inversely proportionally to ES and α (Cohen, 1992).

3.5.1.4 **Effect size**

The number of participants is dictated by the size of effect wished to be detected (measuring the strength of a phenomenon). There are two strategies available to know the effect size before conducting the study. One way is to find a reasonable value for the effect size from previous studies. Another way is to use researcher judgment and heuristics to estimate a likely effect size for the study.

For Cohen's d an effect size of 0.2 to 0.3 might be a "small" effect, around 0.5 a "medium" effect and 0.8 to infinity, a "large" effect (Cohen, 1988).

3.5.2 Sample size in regression

It was assessed that 90 samples would be enough for regression analysis based on the rule of thumb that you need 10 to 15 samples per predictor (in case the number of predictors was 9 or less).

Given 9 independent variables were selected to be included in the regression analysis, the sample size of 90 was reckoned as sufficient for the analysis (Tabachnick and Fidell, 2001).

3.5.3 Sample size for technographic survey (Study 2)

For the technographic survey, a sample size calculator was used³³. More details in chapter 4.

3.6 Ethical Procedure

For this research, the University of Edinburgh School of Informatics Ethical Review Procedure³⁴ was followed, according to the schools' ethics code and practice. The level that was required for the experiments in this thesis was Level 1 indicating low risk.

The researcher's checklist for compliance with the Data Protection Act, 1998 and Procedure for Ethical were reviewed so that I understood the necessities

³³ <http://www.raosoft.com/samplesize.html>

³⁴ Details can be found in: <https://www.ed.ac.uk/informatics/research/ethics/procedure>

in order to conduct the research experiments detailed in this thesis. These documents have been reviewed and approved by myself and my supervisor Professor Austin Tate and submitted to the School of Informatics as required in the School's ethics processes.

3.7 Summary

This chapter provides a description of the methods that were used all through this thesis for the empirical evaluation of ECAs. Firstly, details on the experimental design used for these evaluations, were described taken from methods commonly used in evaluations of this nature. The evaluation metrics for both the usability and agents as well as qualitative methods were discussed in order to assure their reliability and suitability.

Firstly, the importance of using effective evaluation methods for both the experiment interfaces and the actual experiment design was described. The complementary nature of several evaluation strategies was documented together with their impact on the empirical research that was conducted in this thesis.

A discussion on the statistical analysis techniques was then followed, explaining the different types of data and the suitable statistical approaches required to analyse them to report the research findings.

Another important factor in research of this nature, that of sample size justification, is then discussed with an introduction of the four variables involved in statistical inference. The ethical considerations of research involving human participants are also mentioned.

Lastly, a description of the experimental system used in the thesis is provided along with the technology used to develop it.

In the following chapters, a series of empirical evaluations on the assessment of the effectiveness of ECAs are presented. By utilising the experimental methods described in this chapter the author draws conclusions regarding the representations and functionality of ECAs in the context of mobile serious game applications.

Chapter 4 Preliminary work

4.1 Pilot study 1

4.1.1 Introduction

As discussed in the background chapter, the popularity of serious games (SGs) is undisputed with an increasing number of big companies showing interest in investing in SGs. With the increasing interest in SGs in the corporate pipeline, the importance of usability analysis of these systems increases accordingly.

Companies often use SGs during corporate training and learning with behavioural change as the objective (Donovan,L, 2012). Multiple studies verify the effectiveness of agents in behavioural change as well as the transfer of this change to the real world (Hershfield et al., 2011; Yee et al., 2009).

The importance of the use of agents along with games has been expressed by Preben Wik (2011): "Task-based, interactive exercises and the use of sound, pictures, agents and games will not only enrich learning by making it a more worthwhile experience to learn. By presenting content to be learned in a rich multimodal environment, a more robust memory trace is also created and thus the retention will be increased. Motivational and cognitive factors may hence fuse during learning activities and influence the outcome of the skill building".

In SGs, Baylor and Kim (2005) reported that students were significantly more motivated and learned significantly more when the agents' functions were separate (e.g. instructor, collaborator) instead of one with combined

functions; hence two separate agents are used for this experiment. However, the value of ECAs in the context of SGs needs to be addressed further.

To date, little research has been carried out to determine whether the implicit or explicit presentation of feedback affects the users' perceptions of usability in SGs with multiple ECAs; the same is true for the presentation of an application as a learning software or a game.

Aims

The main aim of this evaluation is to act as a methodological sand box which will help decide the methodology approach adopted for the main experiment. Also, it aims to establish that a serious game is a suitable environment for the main experiment.

The experiment presented in this chapter explores user perceptions towards two interaction conditions of an application. The research investigates the overall usability of the application, and a user preference for either the gaming (implicit feedback) or learning mode (explicit feedback) in this casual game.

More specifically the aim of this research is to investigate the users' subjective attitudes in relation to the two conditions:

1. The game version, where the application is presented as a game and the feedback after the task is implicit, in the form of stars and points.
2. The learning version, where the application is presented as a learning tool and the feedback after the task is explicit in the form of text.

Objectives

- Examine the extent to which the addition of game elements during feedback affects users' performance. This will allow future applications to incorporate virtual agents along with game mechanics in order to achieve maximum engagement and efficiency.
- Explore methodological approaches.
- Examine the extent to which the presentation of the application as a SG improves the quality of the interaction for the given domain and task.
- Explain the results obtained in terms of existing theories.

4.1.2 Experiment Interface Design³⁵

In both versions, the user experienced a tutorial on the pre-decimalised currency. After the tutorial, the user was introduced to the main game where Alex, the virtual instructor introduced the way the user could use and submit the coins during the shopping task. After the introduction, the user engaged in a shopping task at the virtual shop with the shop-keeper. The way the user interacted in the game was multimodal where speech was used to interact with the agent and a mouse to handle and submit the coins.

In the learning version, Alex referred to herself as the 'learning assistant' and repeatedly within her tutorial used language which referred to learning. Also, the feedback was offered explicitly in written form and on the left side an

³⁵ Special thanks to Hazel Morton and Nancie Gunson for their contribution on the development and analysis of this experiment.

inventory of the items purchased was presented with no other information.

After each item was purchased, a screen containing text feedback on the user's performance appeared (see Figure 23).

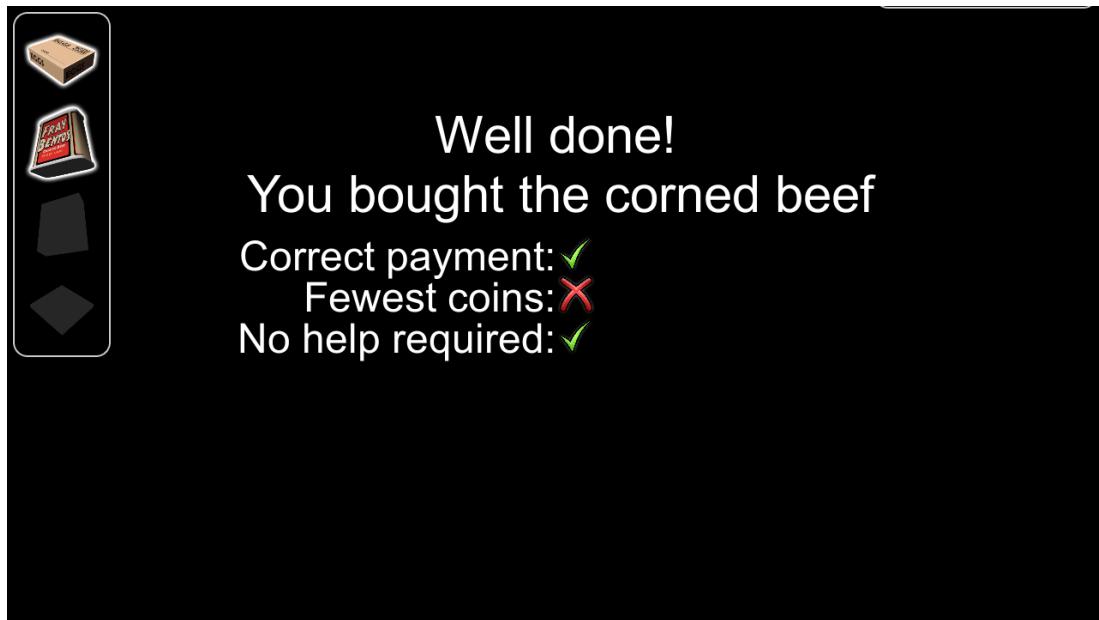


Figure 23-Screen shot of learn mode feedback.

In the gaming version, Alex referred to herself as simply the 'assistant' and within her tutorial used language which avoided specifying any reference to learning. The feedback provided to the user was more implicit in nature compared to the learning version and was presented in the form of stars and points which is a common depiction among casual games as an indicator of progress. The inventory on the left side of the screen contained apart from the items, 3-star outlines that indicated 3 levels of efficiency (time, less coins used, no help needed). Following this, large points appeared in the centre of the screen in red, then the ticker at the bottom of the inventory box turned adding in these points (see Figure 24).



Figure 24-Screen shot of gaming mode feedback.

A script of utterances was written for the application to ensure that both versions provided identical responses to the participant input regardless of which version was experienced. Two professional voice actors (one male, one female) were employed to record the utterances.

The interface was presented on a standard 24" PC monitor, with NVIDIA Quadro K2000 graphics card. The technical characteristics of the experimental setup can be found in Table 8.

Graphics Card Description	NVIDIA Quadro K2000 3D Vision™ Pro
Workstation model	Dell Precision T3600
Processor	Intel® Xeon® Processor E5-1603, 10M Cache, 2.80 GHz, 0.0 GT/s Intel® QPI
Monitor	SA450 Series Screen size 24"
Resolution	1920 x 1200

Table 8-Technical information of experimental setup.

4.1.3 Experiment Design

For this experiment a 2x2 factorial within-subjects design was used. The experiment was conducted in a dedicated laboratory setting within the university.

A cohort of 65 participants (33 females, 32 males) took part. Equal numbers of male and female participants were recruited for this experiment to minimise the effect of participant gender as a confounding variable. The participants were balanced for version and shopping list order with an age under 40 years old. It was decided that only those participants who would have had no prior experience with the currency would be recruited for this experiment; therefore, all participants were under the age of 40 at the time of the experiment. The age limit was calculated based on the context of the game (on pre-decimalised currency) since the old sterling coins that were used for the game were in circulation till 15 February 1971. Therefore, it is highly unlikely for someone under 40 years old to have knowledge about the old money system.

A contrastive study was undertaken in this experiment; two or more versions of the system with a design characteristic altered were experienced by the participants (Table 9). The results acquired from this method are considered to approximate the responses the system would generate in a real-world context of use.

Feedback mode	
Version 1	Version 2
Presented as a learning tool	Presented as a game
Explicit text-based feedback	Implicit visual rewards feedback in the form of stars and points

Table 9-Experiment versions.

A repeated-measures design was widely used in this method as it allows for maximum control compared to between-subject designs. Also, an abundance of data was collected based on subjective attitudes to experiencing different versions of the system.

Participants' attitudes were measured using a dedicated usability questionnaire completed after experiencing each version of the service (information for the metric can be found in chapter 3).

Overall usability scores were obtained and analysed. The mean scores for individual statements can also be examined in order to identify design aspects that were either particularly successful or needed to be improved. Finally, the results can also be analysed based on the participant demographics such as age or gender, so that important differences between the groups be identified and analysed.

In addition to the quantitative data, the approach allows for qualitative data to be collected by using structured interviews with participants after the completion of the task. Data gathered from these interviews can be very

useful in providing an insight into why participants responded in the ways they did.

Title	Usability Evaluation:	
Design		Repeated measures
Null Hypothesis		There is no difference in usability ratings between software version
Dependent Variables		Usability Questionnaire Responses (1-7 Likert scale)
Other Data		Exit Interview Answers
(Experiment) Independent Variables:	1	Feedback presentation (2 levels)
Other Variables:	Presentation Order	Feedback presentation order randomised.
Other Variables: Location	Shopping list Order	Shopping list presentation order randomised.
	Researcher Differences	Controlled by following a prepared procedure and script.
		CCIR, University of Edinburgh
Cohort		N = 65
Remuneration		£30
Duration:		45-60 minutes

Table 10 Summary table of usability evaluation: Implicit – Explicit feedback.

4.1.4 Experiment Procedure

A total number of 65 participants (33 females, 32 males) aged from 18 to 40 were recruited from a customer list and social media to take part in the

feedback experiment. Participants were welcomed and informed about the purpose of the experiments. The experiment was conducted on a desktop PC with a 24-inch monitor, and participants were also provided with a microphone for communicating with the ECAs within the application.

Upon arrival, participants were allocated randomly one of the two treatment groups (list order and feedback) and then they were assigned randomly one of the two orders of experience.

As mentioned in chapter 3, a standardised script was used to assure maximum control. The script was informing the participant that they were about to try two versions of an application and after each one they would be asked to complete a questionnaire and go through a short exit interview. Participants were informed in the beginning of the session that no personal details were to be collected.

The participants experienced each session individually and the two versions were balanced for order across the cohort. Subjects experienced both versions in a repeated measured design. After each version, the participants were asked to complete an attitude questionnaire on a laptop computer (separate from the experiment machine). After each session, the researcher performed an exit interview where the participant was asked about the versions and the overall experience. The sessions lasted approximately 45 minutes for each participant.

- **Participant Induction**

[5 minutes]

- Researcher greets participant
- Gives consent form to participant (MUST BE SIGNED)
- Outlines research session – participants will be trying 2 versions of the application.
- Informs participant they can end the session at any time and that all data are kept confidential and anonymous.

- **First (randomised) design experienced**

[10 minutes]

- Researcher introduces application.
- Informs participant to complete the application.
- **Tutorial experienced**
- Participant experiences first design of game.
- Participant completes the predefined metrics. **[Metrics for quantitative data collection of first experience]**

- **Second (randomised) design experienced**

[10 minutes]

- Researcher introduces second version of the application.
- Informs participant to complete the application.
- Participant experiences the second design of game.
- Participant completes the predefined metrics. **[Metrics for quantitative data collection of second experience]**

- **Exit Interview**

[10 minutes]

- Researcher asks preference between designs,
- general comments and suggestions **[Exit interview]**

- plus, technographic and demographic questionnaire
- Researcher thanks the participant and provides incentive along with a receipt slip (MUST BE SIGNED).

4.1.5 Results

4.1.5.1 Quantitative

An overall mean usability score was calculated from the 18 usability attributes scores for each of the two treatment groups. Overall mean scores for the questionnaire taken did not differ between the two versions. The learning version scored an overall mean score of 5.30, and the gaming version 5.46. Both versions scored A favourably on the 7-point scale but a repeated measures ANOVA on the overall mean scores found no significant differences between the versions. To examine any differences for each of the individual attributes on the questionnaire between the versions, a repeated measures ANOVA was run on the mean scores; version was the within-participants factor and gender and order of participation were the between-participants factors. The descriptive statistics are presented in Table 11.

Descriptive Statistics

	<i>Shopping list</i>	<i>Order of experience</i>	<i>Mean</i>	<i>Std. Deviation</i>	<i>N</i>
<i>Learn_version</i>	<i>Shopping list 1</i>	<i>Game first</i>	5.14	.69	16
		<i>Learn first</i>	4.88	.77	17
		<i>Total</i>	5.01	.73	33
	<i>Shopping list 2</i>	<i>Game first</i>	5.51	.84	15
		<i>Learn first</i>	5.65	.49	17
		<i>Total</i>	5.58	.67	32
	<i>Total</i>	<i>Game first</i>	5.32	.78	31
		<i>Learn first</i>	5.26	.74	34
		<i>Total</i>	5.29	.75	65
<i>Game_version</i>	<i>Shopping list 1</i>	<i>Game first</i>	5.28	.57	16
		<i>Learn first</i>	5.39	.55	17
		<i>Total</i>	5.34	.55	33
	<i>Shopping list 2</i>	<i>Game first</i>	5.25	.74	15
		<i>Learn first</i>	5.87	.47	17
		<i>Total</i>	5.58	.68	32
	<i>Total</i>	<i>Game first</i>	5.27	.65	31
		<i>Learn first</i>	5.63	.56	34
		<i>Total</i>	5.46	.62	65

Table 11-Descriptive statistics.

In order to determine if the difference in the overall mean usability scores for each treatment group was statistically significant, further statistical analysis was required.

Hypothesis testing

Question:

Is there a statistically significant difference between Learn version mean and Game version mean usability scores?

H_0 : There is not a statistically significant difference between Learn version mean and Game version mean scores.

H_a : There is a statistically significant difference between Learn version mean and Game version mean scores.

Data analysis

User Attitude Results

Overall mean scores for the questionnaire taken did not differ between the two versions. The learning version scored an overall mean score of 5.30, and the gaming version 5.46. Both versions scored quite favourably on the 7-point scale; however, a repeated measures ANOVA on the overall mean scores found no significant differences between the versions $F(1,64) = 3.11, p = 0.082$. Thus, the null hypothesis was not rejected.

To examine any differences between the two versions for each of the individual attributes on the questionnaire, a repeated measures ANOVA was run on the mean scores; version was the within-participants factor and order of participation was the between-participants factors. Although both versions were perceived as highly usable, a repeated measures ANOVA on the overall

mean scores found the interaction between version and version order (Table 12) as the only significant difference between the versions.

Significant Differences

Within-Subjects Effects of overall means	versions * order (df = 1; F = 5.671; p = 0.021)
--	---

Table 12-Significant differences of overall means.

One-way ANOVA

Question:

Is there a statistically significant difference on Learn version mean and Game version mean by order?

H_0 : There is not a statistically significant difference on Learn version mean and Game version mean by order.

H_a : There is a statistically significant difference on Learn version mean and Game version mean by order.

Data Analysis

To examine the research question, an Analysis of Variance (one-way ANOVA) was conducted to determine if there is a significant difference on the Learn version mean and Game version by order. The dependent variables in this analysis were the Learn version mean and Game version mean, and the independent variable was the order of experience (Learn version first, Game version first). The assumptions of normality and homogeneity of variance

were assessed. Normality was assessed using the One-Sample Kolmogorov-Smirnov test on both mean scores. As seen in Table 13, there was no significant difference on the Learn version mean score by order ($df=1$; $F=0.083$; $p.=0.774$) but there was a significant difference on the Text mean score by order ($df=1$; $F=5.866$; $p.=0.018$).

ANOVA

		df	F	Sig.
Learn version mean	Between Groups	1	.083	.774
	Within Groups	63		
	Total	64		
Game version mean	Between Groups	1	5.866	.018
	Within Groups	63		
	Total	64		

Table 13 One-way ANOVA.

The results from the one-way ANOVA show that there was a statistically significant difference between the mean scores of the game version depending on the order with which participants experienced it. As seen in the descriptive statistics (Table 14), when the game version was experienced second, participants tended to rate it more favourably. Since the analysis showed a statistical significance between the two versions by order, the null hypothesis was refuted for the Game version.

		N. of participants	Mean score	St. dev.
Game version	1	31	5.270609	.6529045
mean	2	34	5.635621	.5617818
	Total	65	5.461538	.6295301

Table 14-Descriptive Statistics-Game version by order of experience.

Paired samples t-test

	Learn	Game	Sig.		
			T	df	(2-tailed)
Pair 13	Learn13 - Game13-I enjoyed using Moneyworld.	5.44	5.55	-2.14	64 .036
Pair 14	Learn14 -Game14-I thought Moneyworld was fun.	4.44	5.44	-3.10	64 .003
Pair 16	Learn16 - Game16-I found the use of Moneyworld stimulating.	4.9	5.21	-2.05	64 .045

Table 15-Significant differences in individual attributes.

Usability attribute: Enjoyment and fun

"I enjoyed using Moneyworld" and "I thought Moneyworld was fun": both attributes are related to the feeling of entertainment that the user gets by using the application. The presentation of the application as a game and the

rewards systems which is familiar to the player made the application more enjoyable to the participants and, consequently, they had more fun playing it compared to the version that was introduced as a learning application with the explicit feedback.

Usability attribute: Stimulating

"I found the use of Moneyworld stimulating": participants found the Game version more stimulating than the learn version. This can be attributed to the fact that the collection of stars and scores as well as the display of a dummy high score on the screen made them want to try harder, so that they would take the first place. This motivational effect can make the Game version more stimulating and provide evidence that the implicit feedback can affect the way users respond to the application.

Significant Effects in Usability Attributes by order

In order to test the significance of these results, the scores for each individual attribute were analysed in a similar way with the testing of significant effects between the mean scores by order by using the same set of factors as on the overall mean scores.

There was significant interaction between the versions and the order that the versions were experienced for 4 of the 18 attributes (*concentration, nervous, felt in control, embarrassed*); the significance for 3 of them was $p < 0.001$. The two-factor analysis of variance showed a significant main effect of the order factor for 3 of the 18 attributes (*happy to use again, enjoyment, stimulation*).

Order of experience

The significant difference between version and order led to a further analysis by comparing the first experiences of users. The two versions were rated fairly similar by those who experienced them in the order Game-Learn—although the Learn version was rated slightly higher, this was not significant when tested using a repeated-measures ANOVA. In contrast, those who experienced the Learn version first subsequently rated the Game version significantly higher (repeated-measures ANOVA, $p=0.003$). This was a statistically significant result and indicated a contrastive effect. In this preliminary study on the Moneyworld system, the tutorial was incorporated as part of the first experience and not included on the second experience. Subsequently the tutorial was moved as a standalone in the beginning of the session for the main experiment.

The relationship is visible in the estimated marginal means plot (see Figure 25).

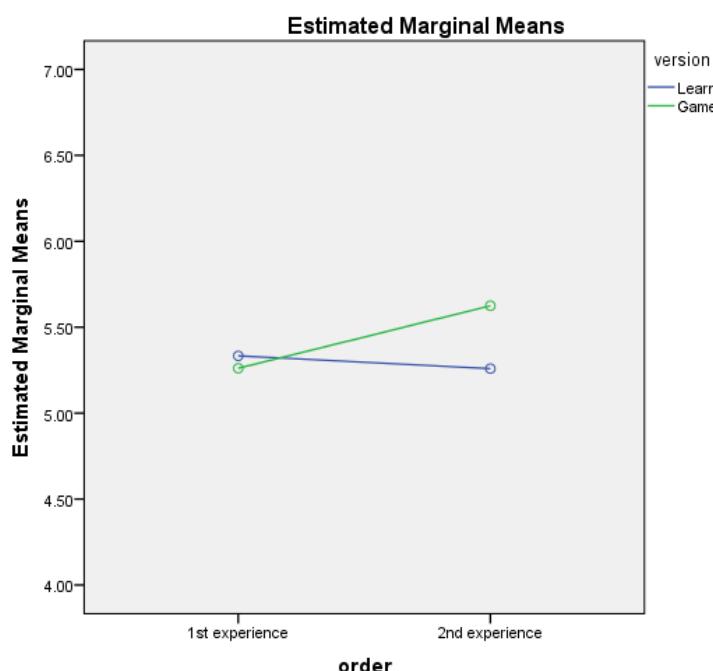


Figure 25- Estimated marginal means Version by Order.

The mean attitude score for the participants' first experience of MoneyWorld was 5.26. Following their second experience, it was 5.48. A repeated-measures ANOVA was carried out on the mean attitude score with *experience* as the within-subjects factor and *version order* as between-subjects' factors. There was a (0.021) significant effect of experience (Table 16), which means that there was a tendency for participants to rate the second version higher, regardless of what it was.

Significant Differences

Within-Subjects Effects of overall means based on experience	Experience (df=1; F=5.671; p = 0.021)
---	---------------------------------------

Table 16-Significant differences of overall means based on first experiences.

4.1.5.2 Qualitative

After experiencing each version, participants were asked to comment on their experience with the application and then specifically on each version they experienced.

The first question participants were asked was: "What did you like about the MoneyWorld experience?". Even though this was an open question, the answers were organised and analysed for recurring themes. In terms of what participants liked in this experience, 22 responded they learned about the old

money system; 20 mentioned the clarity of instructions, ease of use and transparency; 10 liked the visual elements of the application and the graphics; 9 responded the ECAs, especially the shop-keeper; 6 liked that they were able to interact with the system using voice input and 6 referred to the interactivity of the game.

In the question: "What did you dislike about the MoneyWorld experience?", 12 participants responded that they did not enjoy that there were some issues with the speech recognition; 7 responded that they were embarrassed when having to speak; 6 said that it was not clear when they were expected to speak; 6 responded that the application was slow; 4 that the tutorial was too quick; 4 replied that the application felt clunky; 3 that it was not interesting and 3 responded that they did not like Alex's voice.

In the question "Before doing the shopping task, you had a tutorial with the assistant Alex. What did you think about the tutorial?" which focused more on the tutorial, most participants replied positively with 41 stating that it was good, clear and easy to understand. "The tutorial was thorough and informative" replied 16 participants, while 2 responded that it was well paced. There were also a few negative comments about the tutorial, specifically that it was slow and long according to 8 participants; that it was rushed according to one and that it was not clear according to one participant.

When asked if they enjoyed the shopping task, only one stated that they did not as they found it repetitive. The rest responded that it was a good learning experience, fun and interesting.

In the question "Do you feel you understood the old money", one person stated that they did not without giving further explanation. The rest of the participants replied that it was well explained.

Participants were asked what they thought about the way they controlled the coins in MoneyWorld. The majority (60) responded that it felt intuitive or straightforward and easy to control. Other participants commented that it was useful that you could remove the coins from the tray; that they liked having a trial run before the main task; that they preferred clicking on the coins instead of having to speak; and that it was clear when the coins were on the tray.

In the question "What did you think of the shop-keeper", opinions were mixed. Some participants (12) thought that the shop-keeper was fine or ok. Some participants (12) commented on the human characteristics of the agent by saying that he was polite, friendly and kind or that he seemed to be nervous at times. Other participants (10) commented that they liked that he was interactive and conversational and 6 said that the animations and facial expressions were good. Some participants (14) made negative comments about the agent like that he was robotic, creepy, strange looking, stilted and in general they did not like the way he looked; this can be explained by the uncanny valley theory.

Participants were also asked their opinion about Alex. Most comments on Alex were positive with 18 participants reporting that they felt Alex was helpful and gave good explanations; 16 thought she was fine or good and 12 commented that her voice or accent was easy to understand. A few comments (4) were made on the lip synching that was lacking or that she was robotic, her face did not add to the experience or was off-putting.

Inventory

Participants were asked after each version of MoneyWorld they experienced what they thought about the various feedback and reward features

implemented. These data have been combined and the following section details the opinions given for the feedback, irrespective of the order.

Before asked about the inventory, participants were presented with a laminated picture of the inventory as it appeared on screen during the shopping task and they were asked if it was clear what the shopping items indicated. Almost all participants stated that it was clear what the inventory indicated, and some provided additional feedback. In the feedback given, participants commented that the inventory gave them an idea of the progress or an idea on what the next item would be, while others commented that they did not pay much attention to it. Only 2 participants stated that it was not clear what the inventory was for but realised its functionality when the shopping task started.

Gaming Version

While focusing specifically on the gaming version, participants were asked if it was clear what the stars were for. Most participants (54) stated that it was clear and elaborated on their response by saying that the stars were awarded for using the least amount of coins and getting the correct value of the item. Some associated the three stars with the three items offered on the Learn version feedback. Only nine participants stated that it was not clear what the stars were for and one that they did not notice them. Some participants commented that it was not clear when the task session started but it became clear during the research session.

In a similar way, participants were asked if it was clear what the points were for; 56 of them stated that it was clear, and they gave reasons ranging from general overall points awarded for buying the product to detailed points breakdown. Only 7 participants stated that it was not clear what the points

were for; they commented that it was not clear how the points were calculated even if there was a general assumption that the points were allocated for completing each shopping task correctly.

Finally, participants were informed about the purpose of the stars and points and were asked about their opinion. Many comments suggested that participants liked this reward system with comments like:

- “It was great, the competition element gave an edge and a sense of achievement.”
- “It’s more stimulating having a reward.”
- “It’s a good way of keeping you interested.”
- “Without it, it would be quite boring. It’s a good way to provide an incentive to people to think more.”
- “It made it more like a game and made it more competitive.”
- “It made me want to get everything correct.”
- “If it’s a game, it is good to get rewards if you get the answers right.”
- “I’m very much responsive to that. You want someone to say ‘good’ even if the task is simple.”
- “Good way to indicate progress. It would be nice to do something with the points that you earn.”
- “Quite good way of showing you how well you have done or how you could have done. Do you get to do anything with them later?”

The last two comments indicate a level of expectation from the user for the ‘game’ to progress further, or that the user would be able to use the points gained in some way.

Learning Version

When focusing on the learning version of the application, participants were asked if it was clear what each item of the feedback meant. Most participants (58) stated that it was clear. However in the comments, it appears that although the “correct payment” and “fewest coins” feedback was clear, the “no help required” feedback was not clearly understood.

Finally, participants were asked what they thought about the feedback. Comments were mixed. Some example comments follow:

- “Great but I was quite nervous waiting for it, there was an element of suspense.”
- “Very good because I could see my progress.”
- “Clear, explicit.”
- “The feedback explains the task, it could explain how you could do it right if you made a mistake.”
- “I liked the idea. It’s good, but it pulls you out of the experience by blacking screen.”
- “The feedback stayed on the screen for a while, but it was clear.”
- “It confirmed my understanding of the system. No indication of speed you’d solved it though, and not a very engaging presentation.”
- “It was informative, although imprecise. I am not sure what help was though.”
- “It was very long in terms of duration. Should click to bypass.”

- “It feels more of an achievement getting the ‘Congratulations’ or ‘Well done’.”
- “It was a bit dull in the way it was presented. Dated looking.”
- “It was informative, but I preferred the visual than the textual feedback.”
- “Compared to the points system, if you were getting things wrong, this would be more demoralising to have it spelled out to you, rather than just getting 2 stars instead of 3.”
- “I thought it was good, but I missed stars.”
- “I didn’t like that as much. Disappeared but stars were there constantly.”

Some of the comments indicate that participants were prone to making comparisons with the stars and points style of reward offered in the game version, while others thought it was patronising having things spelled out.

Explicit Preference

Finally, participants were asked which version of MoneyWorld they preferred. Participants were asked to give their answer in terms of their first or second version experienced, and the answers were re-ordered for each version.

Fifty-one participants (78.5%) stated that they preferred the gaming version, 11 participants (16.9%) stated that they preferred the learning version, and 3 participants (4.6%) had no stated preference.

Participants were asked to elaborate on their answer. The majority of the comments for the Game version referred to the fact that there was a points system (22 participants), the inclusion of stars as a form of reward (10), that the gaming version was more challenging (7), that it was more game-like (4) and that it appeared to be quick (3).

Some sample comments made by participants follow:

- "The visual stimulation of stars and points was more rewarding."
- "Much clearer, I'm competitive so I liked the points system, felt pacier."
- "Stars and points were better than the feedback screen that lasted ages."
- "The reward system was quite motivating. Get something for your effort. Yellow stars and points provide motivation."
- "I felt it had more point to it. It linked cause and effect. Also I had more incentive to get things right."

For the Learn version, participants commented that they did not like the stars rewards system in the gaming version (3), or that they preferred a learning application (2). Example comments made are:

- "It [Learning version] is clearer. The first one gives you points but you're not sure why. In the second it's clear what you've done is right."
- "It was more visually pleasing. The second was more for children with stars and things."
- "It wasn't obvious what the rewards were for, so not as clear."

Finally, participants were asked if they had any further comments to make.

Some comments offered suggestions, but most were in favour of the application:

- "I felt frustrated because I didn't get the high score, the graphics were very nice, and I liked the interaction through voice because it is not something that I come across very often."

- “It would have been nice to pick my own items, rather than having them specified. It was dragging on towards the end, and it would have been nice to just get the things, rather than having to be prompted what to buy on each item.”
- “It was fun to use, quite clear, quite enjoyable and well laid out.”
- “It was a bit flashy to start with (space scene). Quite educational.”
- “It might have been useful to have a visual of the price, rather than just being told of it. It was fine in a quiet room like this, but in a classroom, it would be trickier.”
- “The shopkeeper’s twitching was alarming.”
- “It was a really interesting learning game.”
- “I think the shop keeper grows on you after the first version.”
- “It was a good tool for learning. I was not just being taught on something, but I was also getting a chance to test what I learned at the same time.”
- “Alex didn’t seem as friendly the second time.”

4.1.6 Discussion and conclusions

The aim of the pilot study was to act as a methodological sand box which will help make methodological decisions for the main experiment. Also, it aims to establish that a serious game is a suitable environment for the main experiment.

The evaluation focuses on investigating users' subjective attitudes towards two versions of a virtual shopping task which used ECAs; the focus on one version was to be a learning application with an explicit feedback and the focus on the other was to be a game with an implicit feedback in the form of stars and points.

The mean overall usability scores across the two versions were positive, suggesting that Moneyworld was generally well received by participants. The positive usability scores for both versions were supported by the qualitative data collected during the exit interview; most of the participant comments for the application were positive.

Taking the mean score results, no overall statistical significance for version was found. The overall mean for the Learn version was 5.30 indicating a positive attitude towards the application. The mean for the Game version was slightly higher at 5.46 that is also positive but the difference between the means was not statistically significant.

A repeated-measures ANOVA with version as the within-subjects factor, and gender, version order and shopping list order as between-subjects factors showed that the difference in the mean-attitude score was not significant.

Looking at the individual attributes on the questionnaire, "I enjoyed using Moneyworld", "I thought Moneyworld was fun", "I found the use of Moneyworld stimulating" were found to be statistically significant between the two versions with the Game version scoring higher in all the cases. This phenomenon can be explained by the theory discussed in Chapter 2 where participants wanted clear and quick feedback. From the qualitative analysis, it is prominent that most participants were familiar with an implicit reward system of stars and scores and they associated it more with games. The

association of such a rewards system with games made the Game version more appealing to participants with many finding the rewards motivating and the application enjoyable. Also, in some comments, participants found stars and points a good way to track their progress throughout the game because they liked to check their achievement collections. This result is consistent with the work of Formanek (1994) and Wang and Sun (2012) "Reviewing rewards provides entertainment, a sense of accomplishment, and memories linking play events to specific rewards."

Furthermore, there was a statistically significant interaction between version and version order ($p=0.021$). Participants who experienced the Learn version first went on to rate the Game version significantly higher. This is a contrastive effect. Given their expectations of the application, the gaming version was then scored significantly higher than their initial experience. Such a difference was not found the other way. For those participants who experienced the Gaming version first, there is no significant difference between versions indicating that the learning version did not improve their experience after trying the gaming version.

Additionally, as the second version experienced by the user did not include the tutorial, a between-subjects comparison of versions based only on the second experience found that the Learn version scored 5.32, and the gaming version 5.64; this was not a significant difference.

There was a tendency for the version experienced second to be rated more positively. This can be attributed to the fact that participants knew what to expect and were familiar with the application which is also mirrored in the comments during the exit interview. After the analysis, results showed the effect of order experience was moderately significant ($p=0.021$) indicating a

possible learning effect and/or a positive effect of removing the tutorial in the second experience. Subsequently, the application is adjusted for future designs by removing the tutorial from the first experience and run it as a standalone in the beginning of the session.

Lastly, during the exit interview participants stated their preferences: 51 participants (78.5%) stated that they preferred the gaming version, 11 participants (16.9%) stated that they preferred the learning version, and 3 participants (4.6%) had no stated preference. When asked to justify their choice, most participants referred to the reward system using either stars or points. Others found the game version more challenging and gamer like and some found it quicker.

Text feedback is assumed to interrupt the game flow especially in an application where the interaction is multimodal, social and no text is used up until then.

When focused on the agents most participants liked interacting with them, especially the shopkeeper who had the collaborator role in the interaction. Many of the comments were positive referring to his funny comments and quirky personality. The negative comments had mostly to do with the lip synching or the face animations which can be explained by the uncanny valley theory.

There were three effects in play. One is that the usability of the Game version was rated higher than that of the Learn version, although not being statistically significant. The second is that the individual attributes of finding the application fun, enjoyable and stimulating were significantly higher for the Game version because participants associated the implicit rewards with games that are usually regarded as a fun activity. The last one is that the

version experienced second tended to be rated more positively. When the two versions were experienced in the order Learn-Game, these two effects were combined to create a statistically significant difference in attitude towards the two versions. When the two versions were experienced in the order Game-Learn, on the other hand, the two effects were in opposition to each other, and effectively cancelled each other out. The fact that the tutorial was experienced within the first only version was found to be problematic; thus, in the main experiment the tutorial was experienced once in the beginning of the session and the actual experiment versions were experienced afterwards.

The empirical data alongside with the qualitative data collected during the post-experience interview, provided an insight into the effect of the use of implicit feedback in the form of points and stars to the overall usability explaining why this was the case.

In terms of methodology many changes were made for the main experiment starting by using t-tests instead of F-tests. Even though, F-tests and t-tests provide almost identical results when there are only two groups for comparison, ANOVAs are computationally expensive without providing more information than t-tests. The second change is performing power analysis before conducting the experiment with a higher cost usability testing and not discounted usability as it provides more robust data. It was assessed that for the purpose of an in-depth academic research, discounted usability would not suffice. The third change is calculating and reporting effect sizes as effect sizes can reveal how meaningful the measured effect is in real life.

4.2 Study 2: Survey on the use of Mobile Devices and game playing

4.2.1 Introduction

The analysis in this report is based on online surveys conducted between February and September of 2016. Adults living in the UK and overseas were eligible to participate so long as they were 18 years of age or older. The survey was conducted in English.

The survey included questions about the frequency that users played games, the platform on which they played games, the type of games they played, how many hours they spent playing, what kind of technology they own, how much time they spent using this technology, the type of activities they used their mobile devices for and the screen size of their devices. The full questionnaire used in this survey can be found in Appendix B.

4.2.2 Purpose of the research

Descriptive research, as the one presented in this section, is used in order to describe associations (e.g. the association between user age and gaming habits) and estimate specific parameters in a population (e.g. the time users spend on their mobile devices) (Kelley, et al., 2003).

The author developed this technographic survey in order to identify technology patterns and collect insights on the users' digital habits and the way they use their mobile devices (tablet and smartphones).

The research results are intended to be used as a basis to update the existing application in order to reflect the real world and make it more relevant to today's users. Through this research, empirical data based on real-world observations can be collected and due to the breadth of coverage the results can be generalisable to a population (Kelley, et al., 2003).

The questions tried to be answered through this survey are:

- Is gaming part of the user's digital habits?
- How do users use their devices?
- Which is their device of choice when it comes to playing games?

4.2.3 Questionnaire Design

The questionnaire was designed in accordance with similar academic questionnaires (Steinkuehler and Squire, 2013) and industrial market research³⁶.

In order to ensure clarity and that the survey was understood – as intended – by the participants, three pilot sessions with cognitive testing elements were conducted. In these sessions, the respondents were presented with the questionnaire items and they were asked to think aloud justifying their response. In some cases, follow-up questions were asked by the author in order to achieve clarity. The questionnaire was deemed to be suitable for data collection.

³⁶ An example of such a market research is: http://www.theesa.com/wp-content/uploads/2014/10/ESA_EF_2014.pdf

As this interview schedule was structured, closed questions were used followed by a range of pre-coded responses. Attention was paid to the wording of the questionnaire, so that no bias would be introduced. The questionnaire was accompanied by a covering letter in the introduction page informing the participant about the purpose of the research, how much time the questionnaire would take to be completed and contact information of the researcher.

4.2.4 Survey Methodology

The survey targeted individuals from the general public. The survey was delivered in a period of six months between February and September of 2016 with 226 responses in total collected by an online survey through a web link publicised in social media and mail lists of the School of Informatics. The sampling for this survey was random allowing for the result to be generalised and statistical analysis to be performed (Kelley, et al., 2003).

The number of participants needed for the survey is typically determined by how confident the researcher wants to be in the results. In this research, a sample size calculator³⁷ was used to determine the sample size based on the margin of error that the researcher can tolerate, the desired confidence level, the population size and the response distribution. The input values for each were based on the most commonly used ones with 5% margin of error, 90% confidence level, 20000 population size and 50% response distribution as there was no hypothesis suggesting skew of the results. The sample size was calculated to be 267 with 226 finally obtained which gave a slightly higher margin of error of 5.44%.

³⁷ Raosoft sample size calculator: <http://www.raosoft.com/samplesize.html>

The participants were informed in the first page of the survey about the purpose of this research, the amount of time needed in order to complete it and contact information for enquiries (see Appendix B).

4.2.5 Results

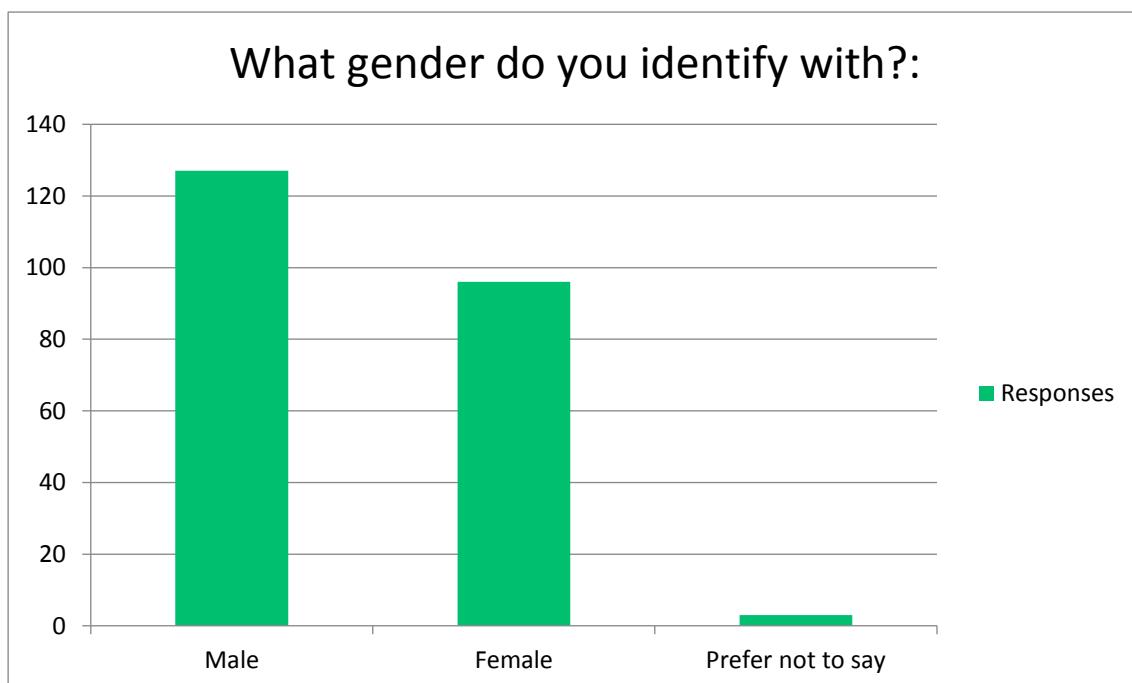
Question 1:



Answer Choices	Responses	
19-25	24.34%	55
26-30	38.94%	88
31-35	19.91%	45
36-40	5.75%	13
41+	10.62%	24
Prefer not to say	0.44%	1
	Answered	226

The participants were asked to select the age group they belonged to. The age group 26-30 had most of the answers from the participants with 38.94% followed by the age group 19-25 with 24.34%, 31-35 with 19.91%, over 41 with 10.62% 36-40 with 5.75% and 0.44% of the participants who did not wish to disclose their age. The majority (83.19%) of participants were under 35 years old with more than half (63.28%) under the age of 30.

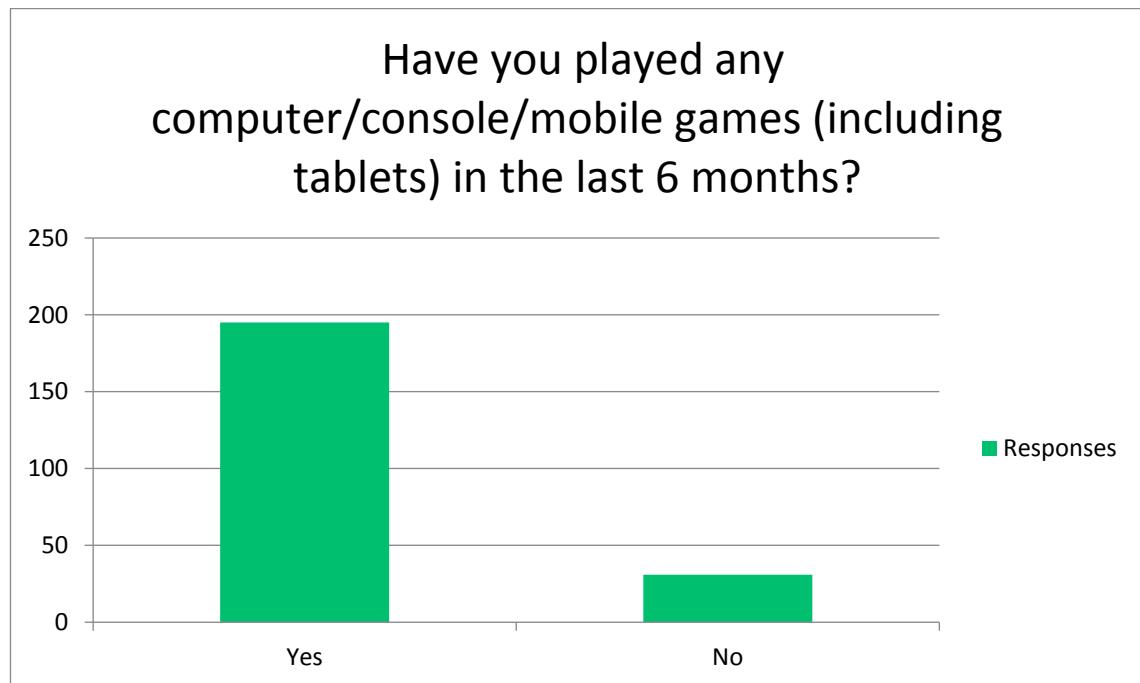
Question 2:



Answer Choices	Responses	
Male	56.19%	127
Female	42.48%	96
Prefer not to say	1.33%	3
	Answered	226

The cohort of the participants consisted of 127 males, 96 females and 3 persons who preferred not to say.

Question 3:



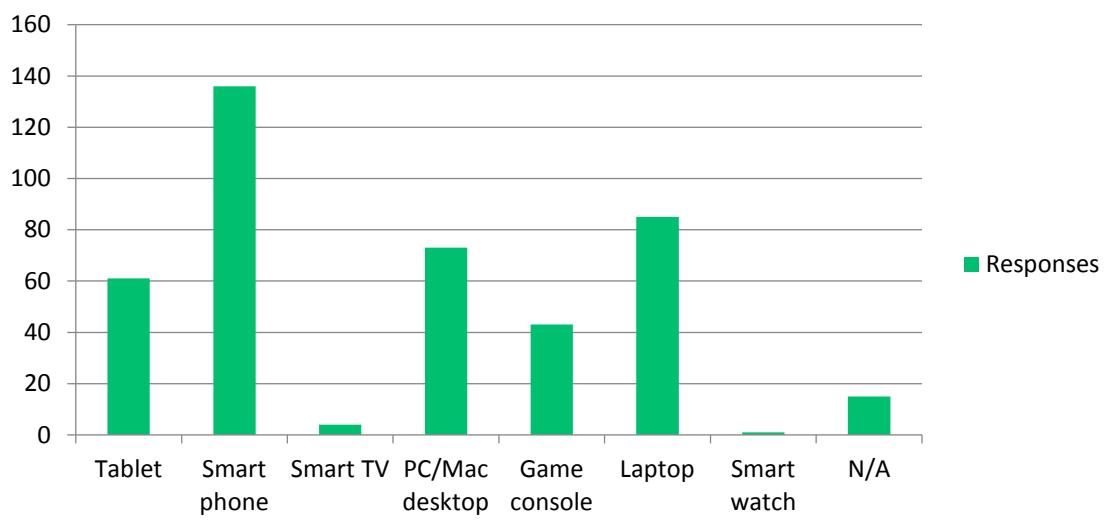
Basic Statistics				
Minimum	Maximum	Median	Mean	Standard Deviation
1.00	2.00	1.00	1.14	0.34

In the question "Have you played any computer/console/mobile games in the last 6 months", most participants (86.28%) answered "Yes" which is

statistically significant compared to those who have not played any games ($t(224) = 8.828$, $p < .001$).

Question 4:

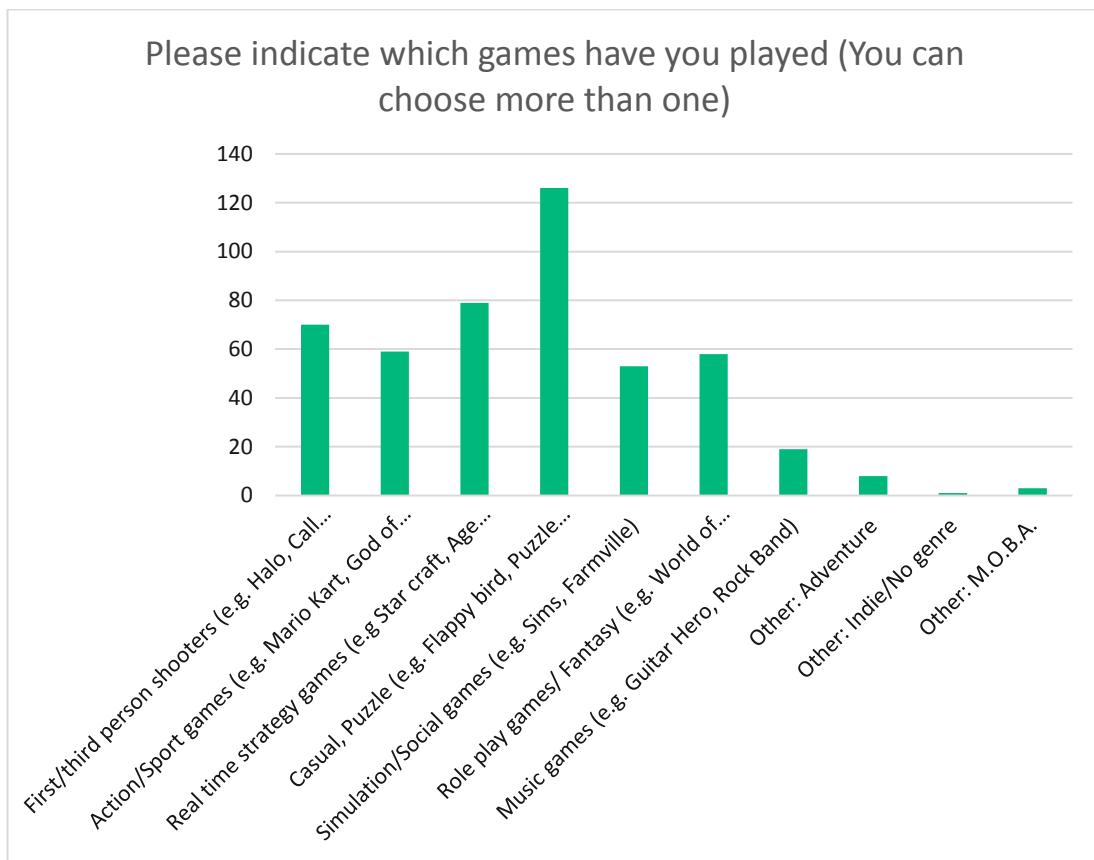
On what device do you usually play games ?(You can choose more than one)



Answer Choices	Responses	
Tablet	26.99%	61
Smart phone	60.18%	136
Smart TV	1.77%	4
PC/Mac desktop	32.30%	73
Game console	19.03%	43
Laptop	37.61%	85
Smart watch	0.44%	1
N/A	6.64%	15
	Answered	226

In the question "On what device do you usually play games?", 60.18% of the participants answered that they play games in their smartphone, the second most frequent answer was laptop with 37.61%, followed by PC/Mac desktop with 32.30%, tablet with 26.99%, game console with 19.03%, smart TV with 1.77% and finally smartwatch with 0.44%.

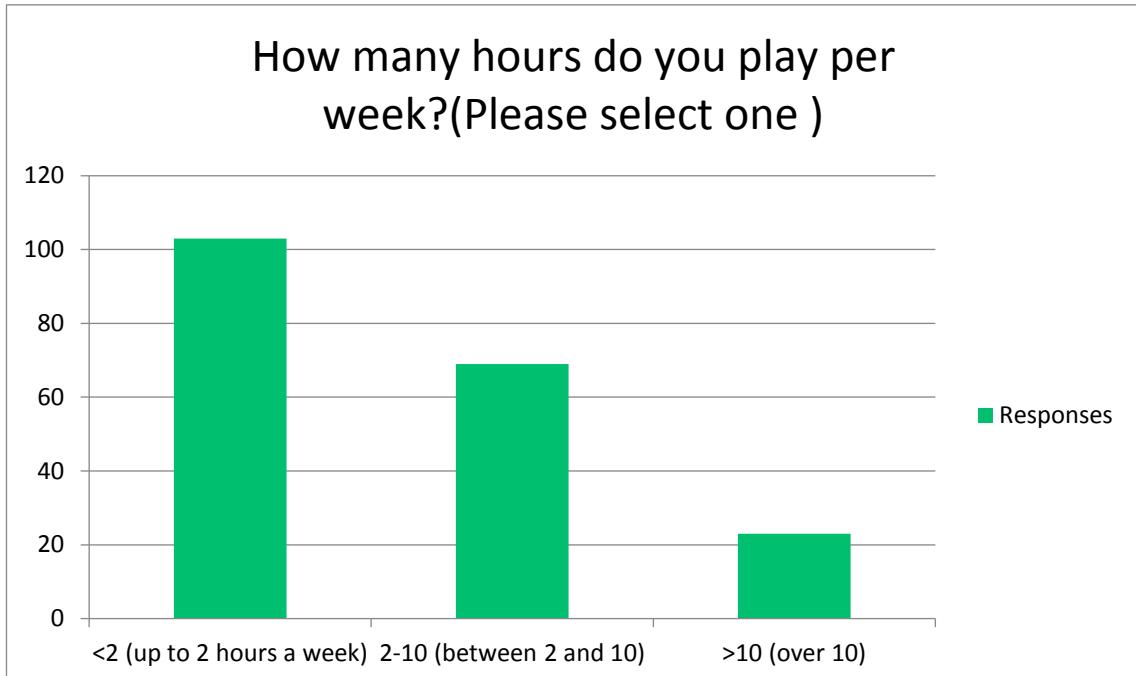
Question 5:



Answer Choices	Responses	
First/third person shooters (e.g. Halo, Call of duty)	36.08%	70
Action/Sport games (e.g. Mario Kart, God of war, FIFA)	30.41%	59
Real time strategy games (e.g. Star craft, Age of Empires, Civilization)	40.72%	79
Casual, Puzzle (e.g. Flappy bird, Puzzle Quest, Bejewelled, Solitaire)	64.94%	126
Simulation/Social games (e.g. Sims, Farmville)	27.31%	53
Role play games/ Fantasy (e.g. World of warcraft, Final Fantasy)	29.89%	58
Music games (e.g. Guitar Hero, Rock Band)	9.79%	19
Other: Adventure	4.12%	8
Other: Indie/No genre	0.51%	1
Other: M.O.B.A.	1.54%	3
	Answered	194

In the question "Please indicate which games have you played" the most popular answer was "casual games" with 64.94% followed by "real time strategy games" with 40.72%, "first/third person shooters" with 36.08%, "action/sport games" with 30.41%, "role play games" with 29.89%, "simulation/social games" with 27.31%, "music games" with 9.79% while participants also mentioned "adventure games" with 4.12%, "multiplayer online battle arena (MOBA)" with 1.54% and "indie/no specific genre" with 0.51%.

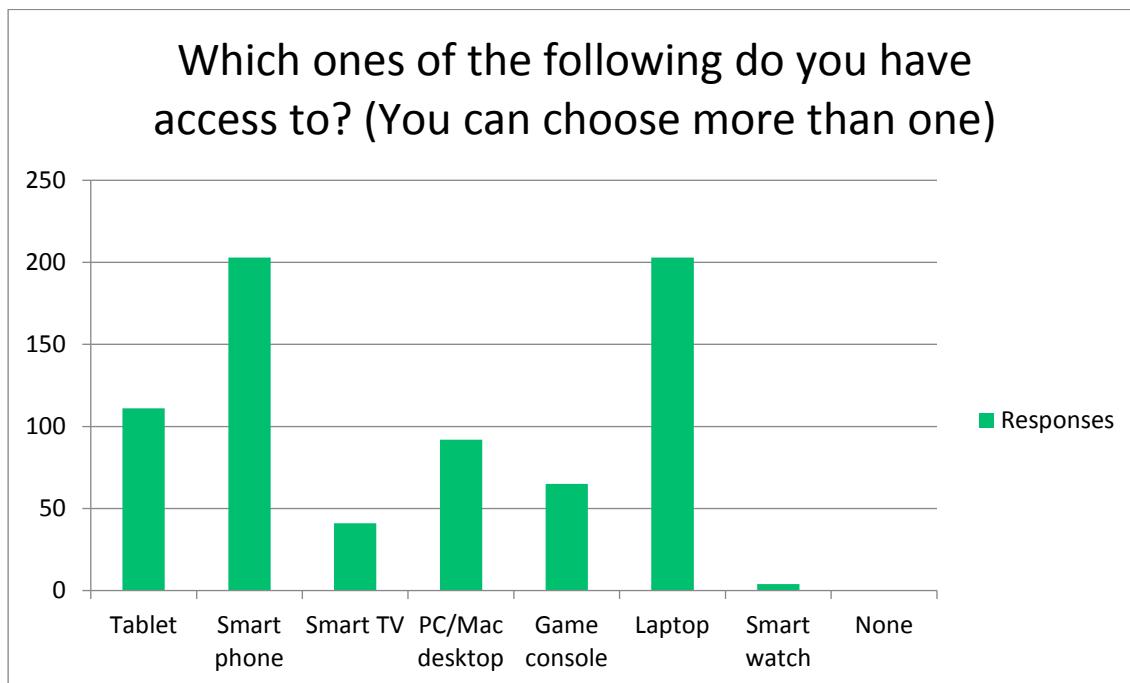
Question 6:



Answer Choices	Responses			
<2 (up to 2 hours a week)	52.82%	103		
2-10 (between 2 and 10)	35.38%	69		
>10 (over 10)	11.79%	23		
Answered		195		
Basic Statistics				
Minimum	Maximum	Median	Mean	Standard Deviation
1.00	3.00	1.00	1.59	0.69

In the question "How many hours do you play per week?", a percentage of 52.82% answered that they play games "up to 2 hours per week", 35.38% answered "between 2 and 10" and 11.79% "over 10".

Question 7:



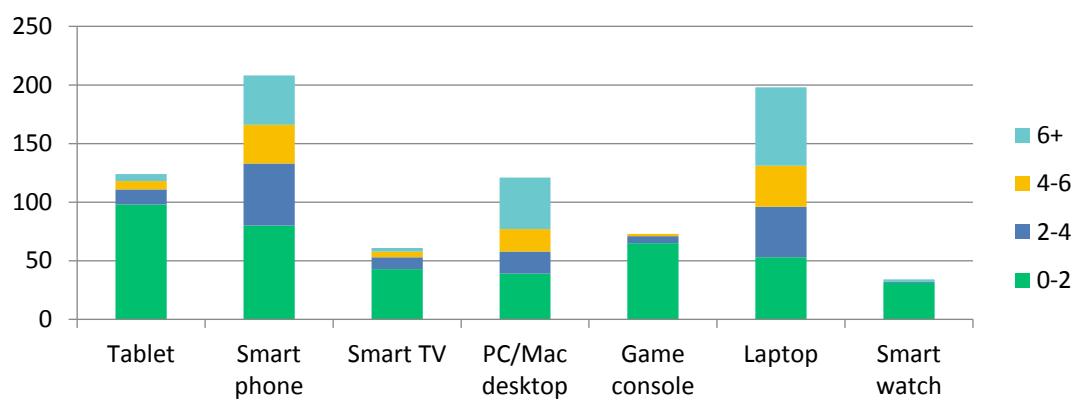
Answer Choices	Responses	
Tablet	49.12%	111
Smart phone	89.82%	203
Smart TV	18.14%	41
PC/Mac desktop	40.71%	92
Game console	28.76%	65
Laptop	89.82%	203
Smart watch	1.77%	4
None	0.00%	0
Other (please specify)		4
	Answered	226

Other: iPod, server

In the question "Which ones of the following do you have access to?", an equal number of people answered that they own a "smartphone" and a "laptop" which were the most frequent answers with 89.82% each, the next most popular answer was "tablet" with 49.12%, followed by "PC/Mac desktop" with 40.71%, "game console" with 28.76%, "smart TV" with 18.14% and finally "smart watch" with 1.77%.

Question 8:

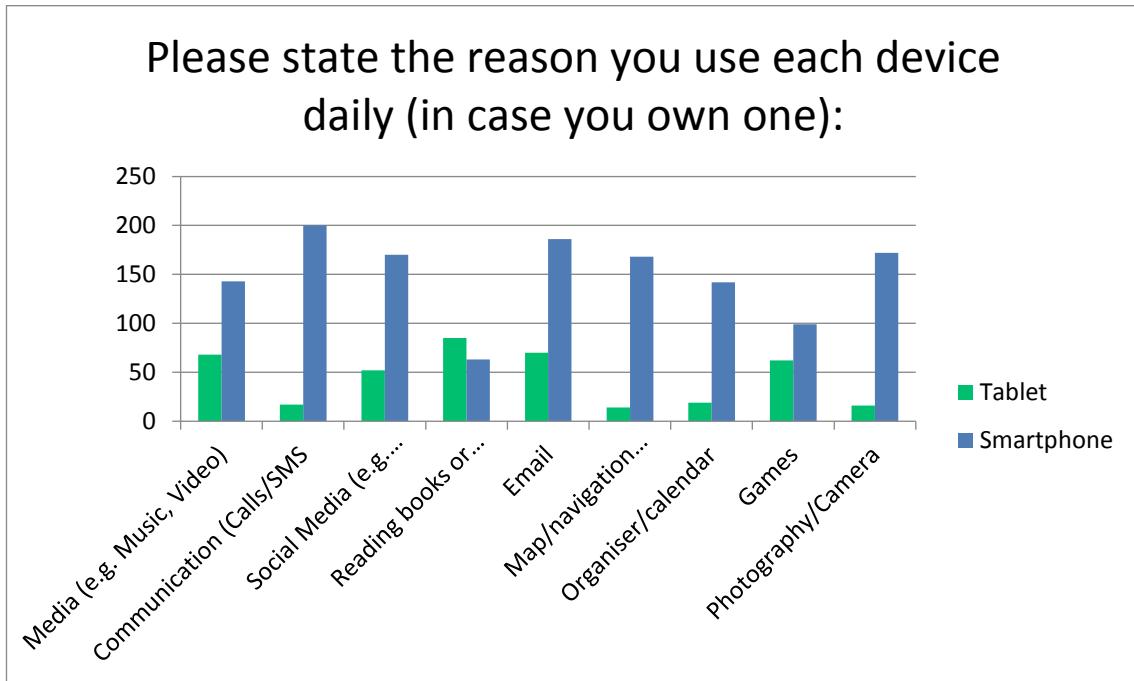
Please state how many hours you use every device you own, daily (on average). Please select one group for each device. (work and leisure)



	0-2		2-4		4-6		6+		Total
Tablet	79.03%	98	10.48%	13	5.65%	7	4.84%	6	124
Smart phone	38.46%	80	25.48%	53	15.87%	33	20.19%	42	208
Smart TV	70.49%	43	16.39%	10	8.20%	5	4.92%	3	61
PC/Mac desktop	32.23%	39	15.70%	19	15.70%	19	36.36%	44	121
Game console	89.04%	65	8.22%	6	2.74%	2	0.00%	0	73
Laptop	26.90%	53	21.83%	43	17.26%	35	34.01%	67	197
Smart watch	91.18%	31	2.94%	1	0.00%	0	5.88%	2	34
								Answered	226

Out of respondents who answered that they have access to a tablet, 79.03% use it for up to 2 hours a day while only 10.49% are using it for more than 4 hours. Similarly, those who have access to a smart TV, a game console and/or a smart watch use them for up to 2 hours with percentages of 70.49%, 89.04% and 91.18%, respectively. Notably, in the case of smartphones, PC/Mac desktops and laptops, the usage time was distributed differently. The participants who have access to smartphones, laptops and/or PC/Mac desktops use them for more than 6 hours per day with 20.19%, 34.01% and 36.36%, respectively.

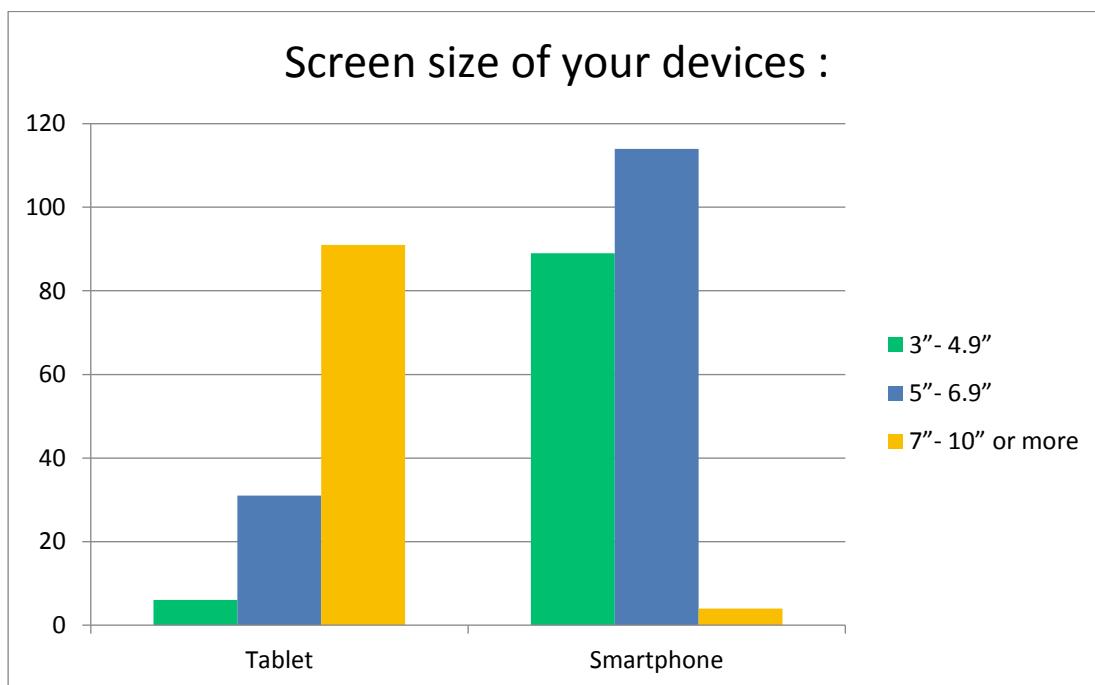
Question 9:



	Tablet	Smartphone	Total
Media (e.g. Music, Video)	38.20%	68	80.34% 143 178
Communication (Calls/SMS)	8.25%	17	97.09% 200 206
Social Media (e.g. Facebook, LinkedIn)	28.57%	52	93.41% 170 182
Reading books or documents (e.g. PDF, WORD)	66.93%	85	49.61% 63 127
Email	34.31%	70	91.18% 186 204
Map/navigation applications	8.05%	14	96.55% 168 174
Organiser/calendar	12.84%	19	95.95% 142 148
Games	50.41%	62	80.49% 99 123
Photography/Camera	9.09%	16	97.73% 172 176
Other (please specify)			9
		Answered	219

Out of 226 respondents who took part in this survey, 111 had access to a tablet and 203 owned a smartphone (some in conjunction with a tablet). Those who had access to a tablet, and/or a smartphone were asked to state the reason they use their devices every day. Over 80% of smartphone users stated that they use the device for playing games and media, 91.18% use it for email, 93.41% for social media, over 95% use it for communication, the map/navigation, the organiser/calendar and photography/camera, while almost half of the participants (49.61%) use it for reading. On the other side, 66.93% of people who have access to a tablet use it for reading, 50.41% use it for games, 38.20% for media, 34.31% for emails, 28.57% for social media, 12.84% as an organiser/calendar while fewer than 10% use it for communication, photography and map/navigation.

Question 10:



	3"- 4.9"		5"- 6.9"		7"- 10" or more		Total
Tablet	4.69%	6	24.22%	31	71.09%	91	128
Smartphone	43.00%	89	55.07%	114	1.93%	4	207
Other (please specify)							6
						Answered	219

In terms of the screen size of the devices, 71.09% of tablet owners had a device of over 7", while 55.07% of smartphone owners had a device that was between 5" and 6.9". A percentage of 43% owned a device with a screen size ranging from 3" to 4.9".

4.2.6 Discussion

From this survey a few interesting facts were identified.

Age

The majority (83.19%) of participants who answered this survey were between 19 and 35 years old.

Games and gender

The majority (86.28%) stated that they have played games in the last 6 months. Even though there was no hypothesis regarding the gender, an interesting observation was that males had different preferences in game genres than females. Also, males have spent more time playing games compared to their female counterparts; 59.46% of males have been playing for over 2 hours per week compared to 29.63% of females. Casual games and puzzles were the genre of choice for 90% of the female participants, followed by simulation/social games with 23.75%, role play games, real time strategy

games, first/third person shooter, action/sport games and music games. For the male population, 56.76% preferred real-time strategy games, followed by first/third person shooter with 54.05%, action/sport games, casual games/puzzles, role play games, simulation/social games and music games as shown in Table 17.

-	FIRST/THIRD PERSON SHOOTERS	ACTION/SPORT GAMES	REAL TIME STRATEGY GAMES	CASUAL, PUZZLE	SIMULATION/SOCIAL GAMES	ROLE PLAY GAMES/FANTASY	MUSIC GAMES	TOTAL
Q2: Male	54.05% 60	44.14% 49	56.76% 63	43.24% 48	27.93% 31	37.84% 42	10.81% 12	159.69% 305
Q2: Female	12.50% 10	10.00% 8	15.00% 12	90.00% 72	23.75% 19	16.25% 13	10.00% 8	74.35% 142
Total Respondents	70	57	75	120	50	55	20	191

Table 17 - Game genre preference by gender.

Device

In terms of the type of device where participants play games, the most frequent answer was the smartphone with 60.18%, followed by the laptop with 37.61% and the PC/Mac desktop with 32.30%. Compared with the fact that 89.82% of the participants stated that they own a smartphone and/or a laptop and the fact that smartphones were the device that have been used most during the day closely followed by the laptop, it can be assumed that users are keen on playing games on their mobile devices. Even though tablets are also mobile devices, the participants preferred using smartphones for most activities (Figure 26) apart from reading which can be attributed to the fact that most tablets have bigger screens than most smartphones and reading on a small screen can be tedious.

Q9 Please state the reason you use every device daily (in case you own one):

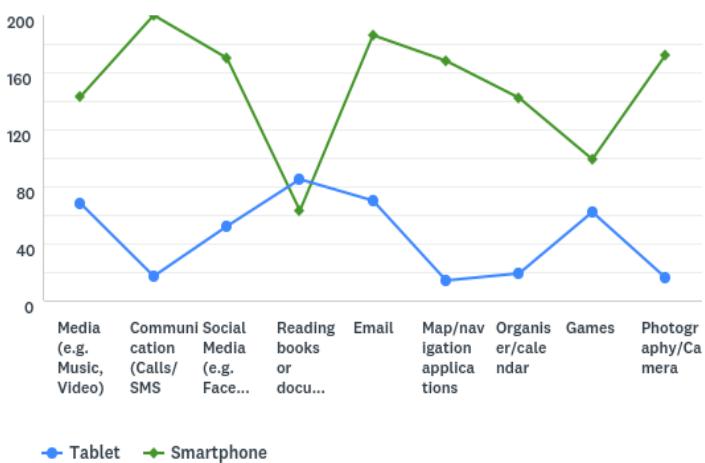


Figure 26- Activities on mobile devices daily.

Thus, based on this survey it was inferred that using a smartphone for the main experiment would be more relevant to today's users.

4.2.7 Summary

This chapter presented two preliminary studies, one usability evaluation and one technographic survey, the findings of which were used as the basis of the main experiment.

The first part of the chapter consisted of a pilot evaluation (study 1) that aimed to assess the effectiveness of a serious game as a platform for the main experiment and act as a sand box for methodological exploration. After introducing the aims and objectives of the study, the application used was presented along with a series of pictures. The two contrasting versions were introduced along with the procedure followed.

Although both versions were perceived as highly usable (with the game version scoring higher), the difference between the two versions in terms of usability was not statistically significant. When examining the individual attributes, quantitative data showed that users found the game version statistically significantly better in three cases: enjoyment of use, fun and stimulating. According to the qualitative data, 78.5% of the participants stated that they preferred the game version with only 16.9% preferring the learn version and 4.6% that had no preference. Hence the version that was presented as a serious game was selected as the basis of the main experiment.

The second part of the chapter presents the technographic survey which seek to provide an insight on the use of mobile devices and computers along with game -playing habits.

The results showed the mobile devices are the devices of choice for many everyday tasks such as checking email, media and photography. When it comes to gaming smartphones are used by 61% of the participants, followed by laptop and PC/Mac desktop. Also, most of the participants who own smartphones answered that their device has a screen size between 5" and 6.9". These results informed the decision to use a smartphone for the main experiment.

Chapter 5 Main Experiment and Evaluation

5.1 Introduction

Aims

This experiment investigates user attitudes to two versions of a mobile serious game (MSG) involving speech recognition and conversational agents (CAs). The objective of this experiment is to examine the extent to which the illusion of humanness evoked by a conversational agent affects the usability of the MSG application and the users' attitudes towards agents with different roles. This empirical evaluation is hoped to provide empirical evidence on the use of ECAs within MSGs as a lack of said evidence was identified in the background chapter and contribute to the research community. Also, the data from this study could be the basis for future designs of similar applications. This experiment would also be very useful from an investigative standpoint, as experiments outside a formal laboratory environment are rare in the literature.

Objectives

- Examine the impact on usability of a humanoid ECA to a mobile serious game.
- Examine the extent to which the presence of a humanoid ECA affects the quality of the interaction for the given domain and task.

- Identify which attributes of the humanoid ECAs contribute to the overall usability, and in what way.
- Explain the results obtained in terms of existing theories, particularly the “illusion of humanness”.

5.2 Experimental Interface Design

As discussed in Chapter 2, Isbister and Doyle, (2002) claim that an agent with physical appearance, sound and animation can cause a powerful visceral reaction on the user – evoke the “illusion of life”. By enhancing realism in movement, creating natural sounding speech and creating the right visual style that fits the application, user’s reaction to the agent can be amplified. Based on the assumptions that human-like realism can evoke an illusion of life and subsequently an illusion of humanness, two versions of agent representation are put to the test based on the spectrum of application interface design in relation to human likeness introduced in Chapter 2 (Figure 27).

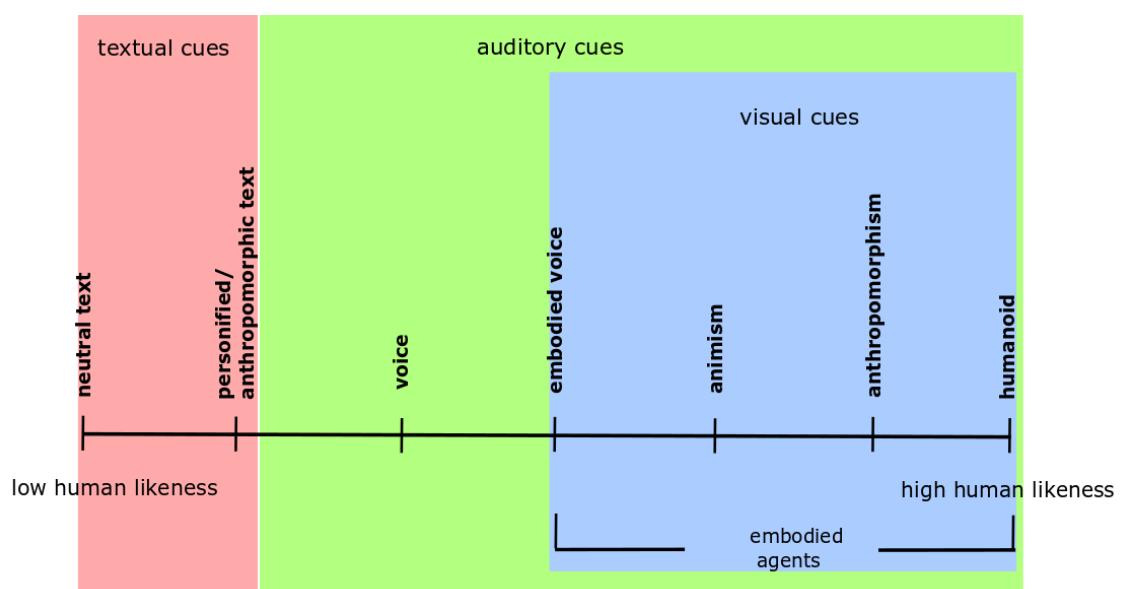


Figure 27-Spectrum of application interface design in relation to human likeness.

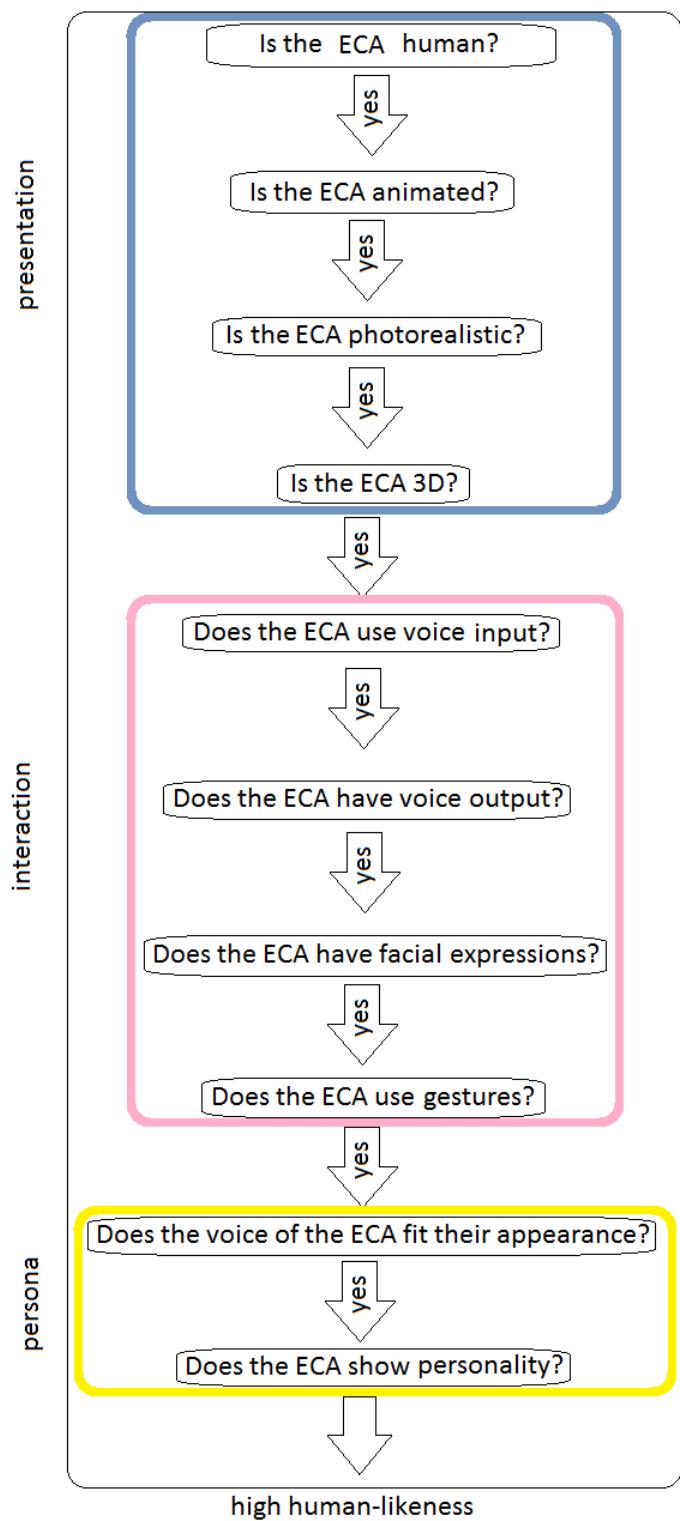


Figure 28-ECA design decisions that result in high human-likeness

In order to achieve high-human likeness, a series of design decisions were made by following the ECADM (Figure 28). The choices were based on the literature which suggested that realism in all levels evokes the illusion of life. For the purposes of this research, two versions of a finance-related SG were compared, the high human-likeness version where the agents were represented by a humanoid ECA and a low human-likeness version where the agents are represented by neutral text conversational agents.

5.2.1 Materials

In the neutral text version, both the instructor agent (Alex) and the collaborator agent (shopkeeper) were presented in the form of a neutral text, as shown in Figure 29.



Figure 29 - Neutral text instructor.

In the ECA version, the instructor agent (Alex) was presented in the form of a female head at the right-top corner and the collaborator agent (shopkeeper) as a contextually-relevant full-body character as shown in Figures 30 and 31. In the ECA version, the agents were embodied with facial expressions and voice and can make gestures.



Figure 30 - ECA instructor.

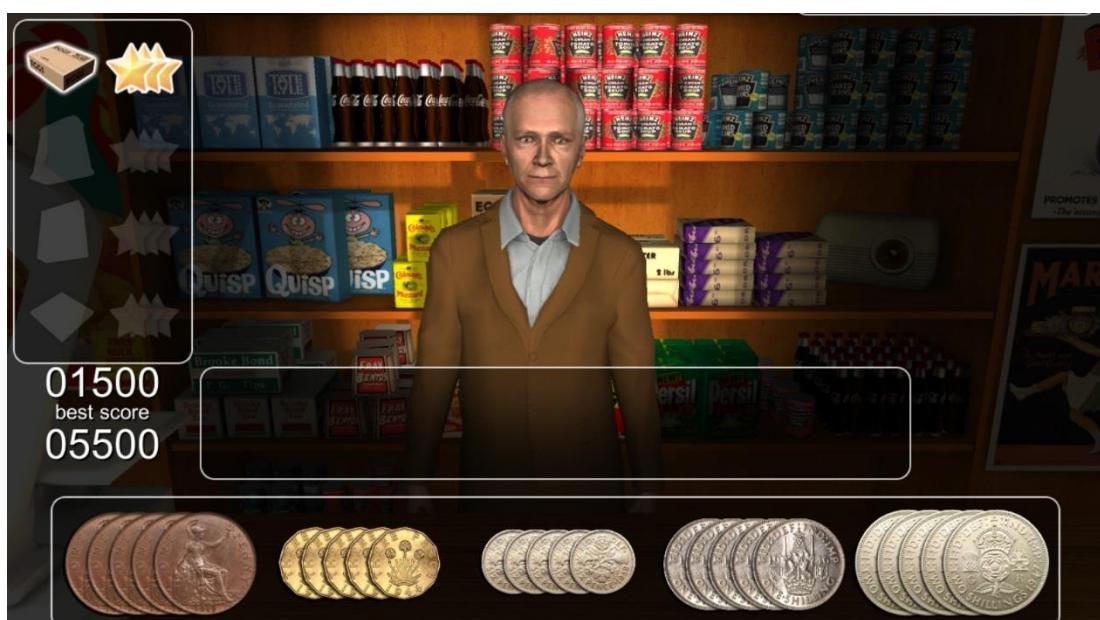


Figure 31 ECA collaborator.

5.3 Experimental Design

The experimental approach involved a contrastive study where two or more versions of the system, differing in some design characteristic, were experienced by the participants. Participants were asked to perform tasks by using the dialogue system. The results obtained from this procedure were considered to approximate the responses the system would generate in a real-world context of use.

A 2x2 factorial experimental design was adopted for the main experiment as the application had two different factors each constituted by two levels as it is shown in table 18. The columns of the table represent the two shopping lists used to avoid overexposure between designs, and the rows represent the level of humanness of the agents used (text-low humanness level, HECA-high humanness level). There was no hypothesis for the shopping lists.

2x2 design		Shopping list	
		1	2
Agent	HECA	V1	V3
	Text	V2	V4

Table 18 - 2x2 factorial design table for the main experiment.

In this approach, a repeated-measures design was largely used to ensure maximum control over between-subject variability and a rich set of data were collected based on both performance measurements and subjective attitudes to the experiences of using the different versions of the system.

5.3.1 Hypothesis testing

For the research presented in this chapter, a two-tailed approach was adopted. Even though the experiment was controlled, it took place outside a formal lab in an environment simulating the real world (open space workstation within the Informatics Forum communal space) thus avoiding limitations of controlled experiments in labs (McInnes, 2005). The setup allowed for observation under circumstances where ambient noise - and in some cases people - are present. In contradiction to lab environments, it is more likely for people to use their mobile devices in public spaces where ambient noise is present.

5.3.2 Sample size

Since the number of descriptive statistics from previous experiments was limited to nil, calculating the effect size was not possible because the standard deviation (*sd*) was not reported (those studies that reported the *sd* used a low number of participants). Thus, the strategy of power analysis was preferred.

Considering that even small effects needed to be detected, the effect size of 0.3 was chosen. For Cohen's *d*, an effect size of 0.2 to 0.3 might be a "small" effect, around 0.5 a "medium" effect and from 0.8 to infinity, a "large" effect.

(Cohen, 1988). A t-test was chosen as the inferential test since there was only one variable of difference and it was necessary to determine if there was a significant difference between the two data sets regarding this variable.

Finally, the number of tails chosen for this power analysis was two as the hypotheses had no direction.

For the calculation of the number of participants needed in order to detect even small effects, G*Power was used with the input parameters detailed in Table 19.

Test family	t-test
Sample groups	Same subjects (repeated measures)
Number of tails	Two
Effect size	0.3
Significance level (α)	0.05
Power	0.8

Table 19-Input parameters for power analysis.

The output parameters given as a result are detailed in Table 20.

Noncentrality parameter δ	2.846050
Critical t	1.986979
Df	89
Total sample size	90
Actual power	0.803794

Table 20-Output parameters for power analysis.

A total of 90 participants were recruited for this experiment. The participants were balanced for version and shopping list order as indicated by the 2x2 design, with age between 18 and 40 years old and a median age of 25.

5.3.3 Participants

The participants were balanced for version and shopping-list order with an age of under 40 years old. The age limit was calculated based on the context of the game, since the old sterling coins that were used for the game were in circulation till 15 February 1971. Therefore, it was highly unlikely for someone under 40 years old to have knowledge about the old money system. The experiment was also within-subjects and balanced.

Subject 1	➤	V1	➤	Standardised questionnaires	➤	V4	➤	Standardised questionnaires	➤	Exit interview
Subject 2	➤	V3	➤	Standardised questionnaires	➤	V2	➤	Standardised questionnaires	➤	Exit interview
Subject 3	➤	V4	➤	Standardised questionnaires	➤	V1	➤	Standardised questionnaires	➤	Exit interview
Subject 4	➤	V2	➤	Standardised questionnaires	➤	V3	➤	Standardised questionnaires	➤	Exit interview

Table 21 Within subject design (repeated measures).

Data were collected from a cohort of 90 participants (47 males, 43 females) with an average age of 25.6 years old. Most participants were international students and professionals (38 native language English, 7 Chinese, 13 Greek,

3 Russian-Ukrainian, 1 Bulgarian, 2 French, 2 German, 3 Hindi, 3 Italian, 1 Indonesian, 1 Japanese, 2 Lithuanian, 3 Romanian, 6 Spanish, 1 Malay, 1 Polish, 1 Telugu, 1 Palestinian Arabic; some were bilingual). The participants were divided into equal and balanced groups with all group subjects experiencing both design options as shown in Table 21.

Title	Usability Evaluation: Presence of Humanoid Animated Agents on Mobile Serious Game	
Design		Repeated measures
Null Hypothesis		There is no difference in usability ratings between software version There is no difference in API ratings between software version
Dependent Variables		Usability Questionnaire Responses (1-7 Likert scale) Agent Persona Instrument (1-5 Likert Scale)
Other Data		Exit Interview Answers
(Experiment) Independent Variables:	1	Agent Embodiment (2 levels)
Other Variables:	Presentation Order	Agent presentation order randomised.
Other Variables:	Shopping list Order	Shopping list presentation order randomised.
	Researcher Differences	Controlled by following a prepared procedure and script.
	Location	Informatics Forum, Edinburgh
Cohort		N = 90 power.t.test (power=0.8, d=0.3, sig.level=0.05, type="paired")
Remuneration		£10
Duration:		45-60 minutes

Table 22 Summary Table of Usability Evaluation: Presence of Humanoid Animated Agents in Mobile Serious Game.

5.3.4 Materials

For this research, two validated questionnaires were used: one to assess the usability of the application and two identical questionnaires (API), one for each agent. The questionnaires were modified to fit the context of the application; therefore, irrelevant Likert items were removed and more specifically the item "The agent's movement was natural".

Responses for the usability questionnaire were on a Likert-type scale, ranging from 1 = "Strongly agree", 2 = "Agree", 3 = "Slightly agree", 4 = "Neutral", 5 = "Slightly disagree", 6 = "Disagree", 7 = "Strongly disagree". Responses for the API were on a Likert-type scale, ranging from 1 = "Strongly disagree", 2 = "Disagree", 3 = "Neutral", 4 = "Agree", 5 = "Strongly agree".

5.4 Experimental Procedure

First, the participants were informed about the purpose of the experiment and then they started the tutorial. In the tutorial, just like in the pilot study, a female unembodied voice welcomed the participants and introduced the concept of the game. The user went through the teleporter and the time/space channel and arrived at the 1960s corner shop in order to play the game. In the corner store, the same voice introduced the old coins to the participant followed by a coin review dialogue.

The same voice then asked the user to identify three coins from the set and to state the value of each of them in pence. After the review, the voice demonstrated how to use the coins in order to buy items.

The tutorial was the same for both versions and was experienced once at the beginning of the session. A different voice than that of Alex, the assistant/instructor, was used for the tutorial in order to avoid overexposure of one style over the other. After each participant interacted with the tutorial, they were asked to answer some relevant questions to the tutorial.

After finishing with the tutorial's questionnaire, the user played Version 1 of Money World, where they were asked to buy 4 items by Alex, the assistant/instructor, who appeared on the right-top corner window, followed by Version 2. The scene comprised the corner store; the shopkeeper/collaborator that the player interacted with in order to buy items as dictated by Alex; and on the left side there was an inventory of the items purchased and the rewards system.

5.4.1 Questionnaires

5.4.1.1 Usability questionnaire

The usability questionnaire used in this evaluation is a standardised and validated metric for assessing usability. Details on the usability questionnaire can be found in Chapter 3, section 3.3.1.1.

Usability Questionnaire Statements

1. I found Moneyworld confusing to use
2. I had to concentrate hard to use Moneyworld
3. I felt flustered when using Moneyworld
4. I felt under stress when using Moneyworld
5. I felt relaxed when using Moneyworld
6. I felt nervous when using Moneyworld
7. I found Moneyworld frustrating to use
8. I felt embarrassed while using Moneyworld
9. While I was using Moneyworld I always knew what I was expected to do
10. I felt in control while using Moneyworld
11. I would be happy to use Moneyworld again
12. I felt Moneyworld needs a lot of improvement
13. I enjoyed using Moneyworld
14. I thought Moneyworld was fun
15. I felt part of Moneyworld
16. I found the use of Moneyworld stimulating
17. Moneyworld was easy to use

Usability Questionnaire Statements

18. I liked the voices in Moneyworld.

19. I thought the voices in Moneyworld were very clear.

20. I thought Moneyworld was too complicated

Table 23 Usability attributes.

5.4.1.2 API questionnaire

The second questionnaire that has been used for this research was also a validated metric for assessing the agent's persona called agent persona instrument (API) (Baylor and Ryu, 2003) as shown in Table 24. Details on the usability questionnaire can be found in Chapter 3, section 3.3.1.2.

<i>Facilitating Learning</i>	<i>Credible</i>	<i>Engaging</i>
<p>1. The agent led me to think more deeply about the presentation. 2. The agent made the instruction interesting. 3. The agent encouraged me to reflect what I was learning. 4. The agent kept my attention. 5. The agent presented the material effectively. 6. The agent helped me to concentrate on the presentation. 7. The agent focused me on the relevant information. 8. The agent improved my knowledge of the content. 9. The agent was interesting. 10. The agent was enjoyable.</p>	<p>1. The agent was knowledgeable. 2. The agent was intelligent. 3. The agent was useful. 4. The agent was helpful. 5. The agent was instructor-like.</p> <p><i>Human-like</i></p> <p>1. The agent has a personality 2. The agent's emotion was natural. 3. The agent was human-like. 4. The agent's movement was natural. 5. The agent showed emotion.</p>	<p>1. The agent was expressive. 2. The agent was enthusiastic. 3. The agent was entertaining. 4. The agent was motivating. 5. The agent was friendly.</p>

Table 24 The API (Agent Persona Instrument) attributes (Baylor and Ryu, 2003).

The dependent variables in the evaluation were the usability and API questionnaire responses and the responses given during an exit interview.

The exit interview was designed in order to retrieve information on the following topics:

- Participant's view of the use of spoken HECA and text CA in a MSG.
- The effective deployment of spoken HECA and text CA in the interface.

To summarise, the evaluation of two types of conversational agents was undertaken in the context of an MSG application. Participants in this evaluation completed usability and API questionnaires related to each conversational agent followed by an exit interview.

The following page details the researcher procedure that was followed for each session.

Money world: Mobile – Embodied CA versus Disembodied CA

The experiments took place in a laboratory setting that simulates aspects of the real-world environment in the School of Informatics, Informatics Forum.

1. Participant Induction

[5 minutes]

- Researcher greets participant.
- Outlines research session – participants will be trying 2 versions of a smartphone-based game called Money world.
- Informs participant they can end session at any time and that all data are kept confidential and anonymous.

2. Tutorial

[5 minutes]

- Researcher introduces tutorial.
- Informs participant they are about to experience the tutorial.

3. First (randomised) design experienced

[10 minutes]

- Researcher introduces Money world.
- Informs participant to complete the game.
- Participant experiences first design of game.
- Participant completes game usability and API questionnaires.

[GameUsab1]

[API instructor]

[API collaborator]

4.Second (randomised) design experienced

[10 minutes]

- Researcher introduces second version of Money world.
- Informs participant to complete the game.
- Participant experiences second design of game.
- Participant completes game usability and API questionnaires.

[GameUsab2]

[API instructor]

[API collaborator]

5.Exit Interview

[10 minutes]

- Researcher asks preference between designs, general comments and suggestions

[InterviewQ here]

5.5 Results

The results presented in this section answer the three research questions:

R1: To what extent do HECA s affect the usability of a mobile serious game (MSG)?

R2: To what extent do users perceive a difference in agent persona between ECA and neutral text presentation as measured by the agent persona instrument (API)?

R3: Which factors relating to the HECA's persona attributes account for variability in usability, and to what extent?

Research question one is answered by a paired t-test analysis on the Usability questionnaire data; research question two is answered by paired t-test analysis on the API questionnaire data and research question 3 is answered by performing a multiple regression analysis with data from both the usability and the API questionnaires.

5.5.1 Quantitative analysis

Identifying influential cases and data correction

The data were analysed parametrically as discussed in the Methodology chapter. In order to support this choice, a further exploration of the data was conducted. With the purpose of determining if the data were normal, the following tools were used:

- Histograms
- Stem and Leaf plots
- Box plots
- P-P plots
- Q-Q plots
- Skewness and kurtosis
- Z-scores

Case 77 was deemed to be an outlier. The outlier was not removed, instead the mean score for the ECA version was corrected from 2.50 to 4.72; this was

the next highest score plus one unit as suggested by Field (2013). The regression runs with the new value.

5.5.1.1 **Research question 1: Usability Questionnaire Results**

R1: To what extent do HECA affect the usability of a mobile serious game (MSG)?

- Identify the extent to which HECA (based on the usability questionnaire) affect usability.
- This research question will be answered by performing paired t-test analysis on the usability questionnaire data.

An overall mean usability score was calculated from the 18 usability attributes (see Chapter 3, section 3.3.1.1.) scores for each of the two treatment groups.

The overall mean scores for the questionnaire taken differed between the two versions. The ECA version received the highest overall mean score of 5.32 (which translates to slightly agree on overall usability), while the Text version received a score of 4.40 (which translates to Neutral on overall usability).

Table 25 details the descriptive statistics for the mean scores of the two versions.

Descriptive Statistics

Order of experience		Mean	Std. Deviation	N
	ECA first	5.21	.72	45
ECA MEAN	Text first	5.43	.80	45
	Total	5.32	.76	90
	ECA first	4.20	1.06	45
TEXT MEAN	Text first	4.60	.95	45
	Total	4.40	1.02	90

Table 25-Descriptive statistics.

Although there were two between-subjects' factors, order of experience and list order, only results by order were reported. This is because different lists were used to balance the versions and to avoid overexposure. Also, there was no hypothesis connected to it.

In order to determine if the difference in the overall mean usability scores for each treatment group was statistically significant, further statistical analysis was required.

Hypothesis testing

Paired T-test

Hypothesis Question:

Is there a statistically significant difference between HECA mean and Text mean.

H_0 : There is not a statistically significant difference between HECA mean and Text mean.

H_a : There is a statistically significant difference between HECA mean and Text mean.

Data Analysis

To examine the hypothesis question, a dependent sample *t*-test was conducted to examine if mean differences existed on the HECA overall mean and Text overall mean.

The dependent samples test of correlated mean differences assumes a normal distribution or a curve that is bell-shaped and symmetrical. The assumption of normality was examined using a one-sample Kolmogorov-Smirnov (KS) test and both were normally distributed. See Appendix C for analysis. (Statistics Solutions. (2013)).

A dependent sample *t*-test for paired means is an appropriate statistical analysis if each of the two samples can be matched on a characteristic. As seen in Table 26, there is a statistically significant difference between the two mean scores ($t=9.45$; $df=89$; $p.=0.000$) and therefore we rejected the null hypothesis.

Paired samples test

		t	df	Sig. (2-tailed)
Pair 1	ECA_MEAN - TEXT_MEAN	9.45	89	.000

Table 26-Paired samples test

Effect size

It is important to advise the effect size to see whether the effect is substantive regardless of its significance.

There are many ways to calculate the effect size but since the t-test was used due to only one variable of difference, Cohen's d was used. Cohen's d is an objective and free from the measuring scale (standardised) measure for determining the magnitude of an effect. It is essentially a measure of whether a statistically significant result has practical significance or not.

The formula for Cohen's d is given by:

$$\text{Cohen's } d = (M_2 - M_1) / SD_{\text{pooled}}$$

$$\text{where: } SD_{\text{pooled}} = \sqrt{(SD_1^2 + SD_2^2) / 2}$$

For this experiment:

$$\text{Cohen's } d = (4.405556 - 5.317259) / 0.903305 = 1.01.$$

Cohen (1988) reports the following intervals for the d values: from .1 to .3 there is a small effect; from .3 to .5 there is an intermediate effect; and from .5 and higher there is a strong effect. Therefore, Cohen's effect size value ($d = 1.01$) suggested a high practical significance which means that the inclusion of an HECA in the MSG has a meaningful real-life impact on the usability.

A summary of the rules of thumb on magnitudes of effect sizes for Cohen's d is given in Table 27 along with the values of effect size for this experiment.

Effect Size	Use	Small	Medium	Large	Effect size for this experiment
<i>Cohen's d</i>	<i>t-tests</i>	0.2	0.5	0.8	1.01

Table 27-Cohen's d and omega-squared rules of thumb and reported effect sizes for this experiment.

Type I error

In order to avoid a Type I error for multiple t-tests (for all 18 statements) a Bonferroni Correction and Holm-Bonferroni Sequential Correction were calculated.

Post-hoc Bonferroni Correction and a Holm-Bonferroni Sequential Correction tests showed that all ECA statements' scores were found to be statistically significant compared to the Text statements' scores.

This analysis is only needed when the difference in the overall mean usability scores is found to be statistically significant in the first paired t-test. This decreases the likelihood of reporting results that are in fact erroneous

(as a result of Type I error). With no correction the chance of finding one or more significant differences in 18 tests = 0.6028 (60.28%). After applying the Bonferroni correction, the alpha value equals 0.0027778.

Type II error

Since there are only two conditions, sphericity is not an issue in this experiment therefore there is a decreased probability of Type II error.

Individual statements

Although the main test compares the overall means of each version (ECA-TEXT), it does not reveal which attributes were significant or not as it is an omnibus statistical test. To examine any differences for each of the individual attributes on the questionnaire between the versions, a paired t-test was run on the mean scores of each question. The Confidence Interval of the Difference was increased to 99.9972% as dictated by the Bonferroni correction details which can be found in the Type I error subsection. The results of these tests are reported in Table 28.

Paired T-test

Hypothesis Question:

Is there a statistically significant difference for each of the individual attributes on the questionnaire between the versions ECA and Text?

H_0 : There is not a statistically significant difference for each of the individual attributes on the questionnaire between the versions ECA and Text.

H_a : There is a statistically significant difference for each of the individual attributes on the questionnaire between the versions ECA and Text.

Data Analysis

Based on the data from the paired samples t-test summarised in Table 27, the null hypothesis can be rejected for all the attributes meaning that the difference between versions was statistically significant for all the 18 attributes. After Bonferroni correction the new alpha value is alpha=0.0027.

Paired samples test*

	version	Mean ECA	Mean Text	Sig.		
				T	df	(2-tailed)
Pair 1	ECA1 - TEXT1-I found Moneyworld confusing to use.	5.93	5.01	5.485	89	.000
Pair 2	ECA2 - TEXT2-I had to concentrate hard to use Moneyworld.	5.18	4.28	5.135	89	.000
Pair 3	ECA3 - TEXT3-I felt flustered when using Moneyworld.	5.44	4.42	6.169	89	.000
Pair 4	ECA4 - TEXT4-I felt under stress when using Moneyworld.	5.83	4.76	6.618	89	.000
Pair 5	ECA5 - TEXT5-I thought Moneyworld was too complicated.	6.16	5.84	3.209	89	.002
Pair 6	ECA6 - TEXT6-I felt nervous when using Moneyworld.	5.49	4.92	3.567	89	.001
Pair 7	ECA7 - TEXT7-I found Moneyworld frustrating to use.	5.29	3.77	7.913	89	.000

Paired samples test*

		Mean ECA version	Mean Text version	T	df	Sig. (2-tailed)
Pair 8	ECA8 - TEXT8-I felt embarrassed while using Moneyworld.	5.32	4.70	3.567	89	.001
Pair 9	ECA9 - TEXT9-I felt Moneyworld needs a lot of improvement.	4.17	2.99	7.913	89	.000
Pair 10	ECA10 - TEXT10-I felt in control while using Moneyworld.	5.20	4.23	5.356	89	.000
Pair 11	ECA11 - TEXT11-I would be happy to use Moneyworld again.	5.18	4.24	5.913	89	.000
Pair 12	ECA12 - TEXT12-I felt I relaxed when using Moneyworld.	5.04	4.32	4.481	89	.000
Pair 13	ECA13 - TEXT13-I enjoyed using Moneyworld.	5.26	4.30	5.433	89	.000
Pair 14	ECA14 -TEXT14-I thought Moneyworld was fun.	5.22	4.30	6.144	89	.000
Pair 15	ECA15 -TEXT15-I felt part of Moneyworld.	4.64	3.51	7.060	89	.000
Pair 16	ECA16 - TEXT16-I found the use of Moneyworld stimulating.	4.76	4.26	3.554	89	.001
Pair 17	ECA17 - TEXT17-Moneyworld was easy to use.	5.81	4.98	4.891	89	.000
Pair 18	ECA18 - TEXT18-While I was using Moneyworld I always knew what I was expected to do.	5.34	4.47	3.990	89	.000

*99.9972% Confidence Interval of the Difference alpha=0.0027

Table 28 Sample t-test summary after Bonferroni correction.

As seen in Table 28, all usability attributes came back to be statistically significant. The ECA version scored higher on all questions. This difference support that the illusion of humanness effect theory holds in participants' perceptions of the software usability. The difference between the two versions can be seen in the error plot below (Figure 32).

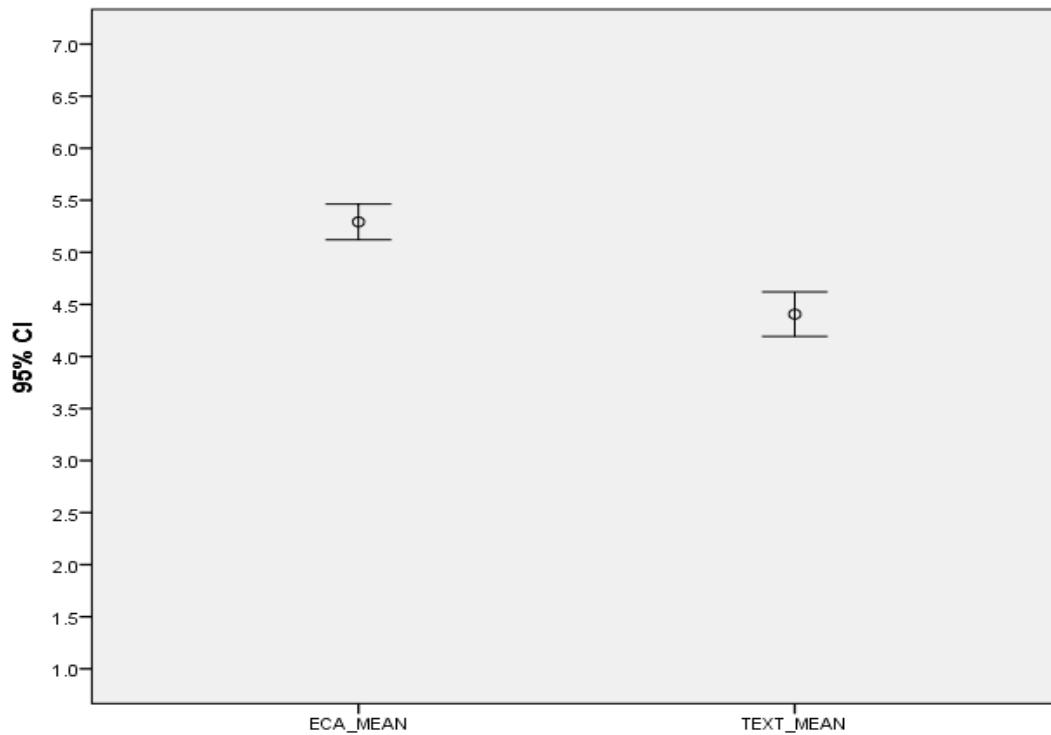


Figure 32-Error plot.

As seen in Table 28, Text version scored below neutral in 3 attributes (frustration, needs a lot of improvement and immersion) and over slightly agree in only 2 (confusing to use and too complicated). The ECA version scored overall above average and was perceived to be usable. It scored between neutral and slightly agree in 3 attributes (needs improvement, stimulation and immersion), and over agree in all the rest except one where it was scored as strongly agree; that translates to participants feeling that the version was not too complicated.

The difference in the overall mean usability scores of the two versions of the game could be attributed to all the attributes. The most relevant attributes have been discussed individually below.

Usability Attribute: Concentration

"I had to concentrate hard to use Money world." The data in Table 28 show that users reported that they had to concentrate harder when using the Text version compared to the ECA version. A possible explanation is the increased cognitive load from having to read from a mobile screen. The t-test confirmed the difference in these mean usability scores to be statistically significant ($T=5.13$, $df=89$, $p<0.001$).

Usability Attribute: Frustrating

"I found Money world frustrating to use." The data in Table 28 show that users reported feeling more frustrated while using the Text version of the game compared to the ECA version. Participants also commented that they felt the Text version was less responsive. The t-test confirmed the difference in these mean usability scores to be statistically significant ($T=7.91$, $df=89$, $p<0.001$).

Usability Attribute: Embarrassed

"I felt embarrassed when using Money world". What is interesting in this case is that although participants reported quite often that they would feel embarrassed using a speech recognition system in public, both versions were rated relatively high. The t-test confirmed the difference in these mean usability scores to be statistically significant ($T=3.43$, $df=89$, $p<0.001$) which confirms that they felt less embarrassed playing the game with the ECA.

Usability Attribute: Enjoyed using

"I enjoyed using Money world." The data in Table 28 show that users reported that they enjoyed more the version with the ECA compared to the Text version. T-test confirmed the difference in these mean usability scores to be statistically significant ($T=5.43$, $df=89$, $p<0.001$).

Usability Attribute: Fun

"I thought Money world was fun." According to the data in Table 28, users reported that the ECA version was more fun than the Text version. According to participants' comments the addition of the HECA made the application to feel more like a game and consequently the experience more fun. The t-test confirmed the difference in these mean usability scores to be statistically significant ($T=6.14$, $df=89$, $p<0.001$).

Usability Attribute: Felt part

"I felt part of Money world." In terms of immersion, participants reported that the ECA version felt more immersive and like a real transaction. The t- test confirmed the difference in these mean usability scores to be statistically significant ($T=7.06$, $df=89$, $p<0.001$).

Usability Attribute: Knew what to do

"When I was using Money world, I always knew what I was expected to do." The data in Table 28 show that users reported feeling like they had a better understanding on what they were expected to do while using the ECA version of the game compared to the Text version. Participants figured out that they had to respond verbally in ECA version, due to the visual and auditory cues.

The t-test confirmed the difference in these mean usability scores to be statistically significant ($T=3.99$, $df=89$, $p<0.001$).

5.5.1.2 **Research question 2: Agent Persona Instrument Analysis**

R2: To what extent do users perceive a difference in agent persona between ECA and neutral text presentation as measured by the agent persona instrument (API)?

- Identify the extent to which the attributes of the ECA (based on API for each agent) differ from that of the text agent.
- This research question will be answered by performing paired t-test analysis on the API questionnaire data for each agent (instructor, collaborator).

The Agent Persona Instrument (API) is a validated instrument for measuring pedagogical agent persona as perceived by the learner. More information in Chapter 3, section 3.1.1.2.

In this experiment we had two agents, an instructor agent that gives instructions on how the coins should be used and says which items should be purchased next and a collaborator agent which interacts with the user during the transaction.

An overall mean score was calculated from the 24 agent questions scores for each of the two treatment groups and each of the agents.

Finally, although there were two agents in each version, they were assessed and analysed separately as they serve different purposes in the interaction

and they are different on many levels; therefore, the two agents cannot be aggregated.

Collaborator agent

The overall mean scores for the collaborator agent questionnaire did differ between the two versions. The ECA agent received the highest overall mean score of 3.67 which translates to between neutral and agree and that participants reacted positively to the agent. The Text agent received a score of 2.81 which translates to between disagree and neutral about their reaction towards the agent. Table 29 details the descriptive statistics for the mean scores of the two versions.

Descriptive Statistics

		Mean	Std. Deviation	N
Collaborator Mean ECA version	Total	3.6713	.58255	90
Collaborator Mean TEXT version	Total	2.8153	.68948	90

Table 29-Descriptive statistics for the collaborator agent.

Instructor agent

The overall mean scores for the instructor agent questionnaire taken differed between the two versions. The ECA version received the highest overall mean score of 3.54 which translates to between neutral and agree and, thus, participants reacted positively to the agent. The Text version received a score

of 2.91 which translates to between disagree and neutral on their reaction towards the agent. Table 30 details the descriptive statistics for the mean scores of the two versions.

Descriptive Statistics				
		Mean	Std. Deviation	N
Instructor agent ECA Mean	Total	3.5407	.57995	90
Instructor agent TEXT Mean	Total	2.9116	.68379	90

Table 30-Descriptive statistics for the instructor agent.

Although there were two between subjects' factors, analysis showed no effect for order, see Appendix D for descriptive statistics.

In order to determine if the difference in the overall mean usability scores for each treatment group was statistically significant, further statistical analysis was required.

Hypothesis testing

Paired T-test

Hypothesis Question:

H_0 : There is not a statistically significant difference between ECA mean and Text mean.

H_a : There is a statistically significant difference between ECA mean and Text mean.

Data Analysis

To examine the research question, a dependent sample *t*-test was conducted to examine if mean differences exist on the ECA overall mean and Text overall mean.

Collaborator agent

The assumption of normality was examined using a one-sample Kolmogorov-Smirnov (KS) test and both were normally distributed (see Appendix D).

As seen in Table 31, there is a statistically significant difference between the two mean scores ($t=13.068$; $df=89$; $p.=0.000$); therefore, the null hypothesis was rejected.

Paired samples t- test

		<i>t</i>	<i>df</i>	Sig. (2-tailed)
Pair 1	ECA_MEAN - TEXT_MEAN	13.068	89	.000

Table 31-Paired samples t-test for collaborator agent version means.

Instructor agent

Again, the assumption of normality was examined using a one-sample Kolmogorov-Smirnov (KS) test and both were normally distributed as none of the means were statistically significant. See Appendix D for analysis.

As seen in Table 32, there is a statistically significant difference between the two mean scores ($t=8.428$; $df=89$; $p.=0.000$); therefore, it is assumed that there is a statistically significant difference between the ECA mean and Text mean for the instructor-agent persona questionnaire.

Paired samples t- test

		<i>t</i>	<i>df</i>	Sig. (2-tailed)
Pair 1	ECA_MEAN - TEXT_MEAN	8.428	89	.000

Table 32-Paired samples t-test for instructor agent version means.

Effect size

Collaborator agent

First, the effect size was calculated for the collaborator agent persona questionnaire. A t-test was used for this experiment because there was only a single variable of difference; therefore, Cohen's d was selected.

For this experiment:

Cohen's $d = 1.34$

A summary of the rules of thumb on magnitudes of effect sizes for Cohen's d are shown on Table 33 along with the values of effect size for this experiment (collaborator agent).

Effect Size	Use	Small	Medium	Large	Effect size for this experiment
Cohen's d	t-tests	0.2	0.5	0.8	1.341

Table 33-Cohen's d and omega-squared rules of thumb and reported effect sizes for the collaborator agent persona.

Instructor agent

The effect size for the instructor agent persona questionnaire is:

Cohen's $d = 1.34$

A summary of the rules of thumb on magnitudes of effect sizes for Cohen's d and omega-squared are shown on Table 34 along with the values of effect size for this experiment (instructor agent).

Effect Size	Use	Small	Medium	Large	Effect size for this experiment
Cohen's d	t-tests	0.2	0.5	0.8	1.341

Table 34-Cohen's d and omega-squared rules of thumb and reported effect sizes for the instructor agent persona.

Individual statements ANOVAs

Although the main test compared the overall means of each version (ECA-TEXT), it did not inform which items were significant or not as it was an omnibus statistical test. To examine any differences for each of the individual items on the questionnaire between the versions, a paired t-test was run on the mean scores of each question. The results of these tests are given in Table 35.

Type I error

In order to avoid a Type I error for multiple t-tests (for all 24 statements), a Bonferroni Correction was run. More details in Appendix D.

Bonferroni's adjustment:

Lower alpha to **0.0020833**

Type II error

Since we have only two conditions, sphericity is not an issue in this experiment therefore there is a decreased probability of Type II error.

Collaborator agent

Questionnaire statement	ECA (Mean =)	TEXT (Mean =)	t	df	p.
The agent kept my attention. -	4.01	4.28	5.85	89	.000
The agent made the instruction interesting. -	3.79	2.52	12.11	89	.000
The agent presented the material effectively. -	4.09	3.64	3.69	89	.000
The agent helped me to concentrate on the presentation. -	3.73	3.12	5.00	89	.000
The agent was knowledgeable. -	3.68	3.21	4.34	89	.000

Questionnaire statement	ECA (Mean =)	TEXT (Mean =)	t	df	p.
The agent encouraged me to reflect what I was learning. -	3.23	3.00	1.83	89	.070
The agent was enthusiastic. -	3.68	2.29	10.93	89	.000
The agent led me to think more deeply about the presentation. -	3.16	2.68	4.18	89	.000
The agent focused me on the relevant information. -	3.69	3.61	0.69	89	.493
The agent improved my knowledge of the content. -	3.50	3.26	1.87	89	.065
The agent was interesting. -	3.72	2.50	11.97	89	.000
The agent was enjoyable. -	3.78	2.50	11.32	89	.000
The agent was instructor-like. -	2.80	3.53	-5.16	89	.000
The agent was helpful. -	3.86	3.53	2.86	89	.005
The agent was useful. -	3.82	3.59	1.94	89	.056
The agent showed emotion. -	3.76	1.81	17.88	89	.000
The agent has a personality. -	3.96	1.94	18.87	89	.000
The agent's emotion was natural. -	3.29	2.53	4.87	89	.000
The agent was human-like. -	3.78	2.06	13.73	89	.000
The agent was expressive. -	3.81	2.18	12.51	89	.000
The agent was entertaining. -	3.77	2.23	12.50	89	.000
The agent was intelligent. -	3.34	2.86	5.07	89	.000
The agent was motivating. -	3.53	2.69	7.97	89	.000
The agent was friendly. -	4.34	3.06	11.86	89	.000

Table 35-Mean scores and results of paired t- tests on Individual Agent Persona Instrument for version - Collaborator agent.

The difference on the API Likert scale between the two designs for the collaborator agent is illustrated in the Figure 33. The HECA version of the collaborator scored higher than the text version on all cases but one (The agent was instructor-like).

Collaborator agent-API

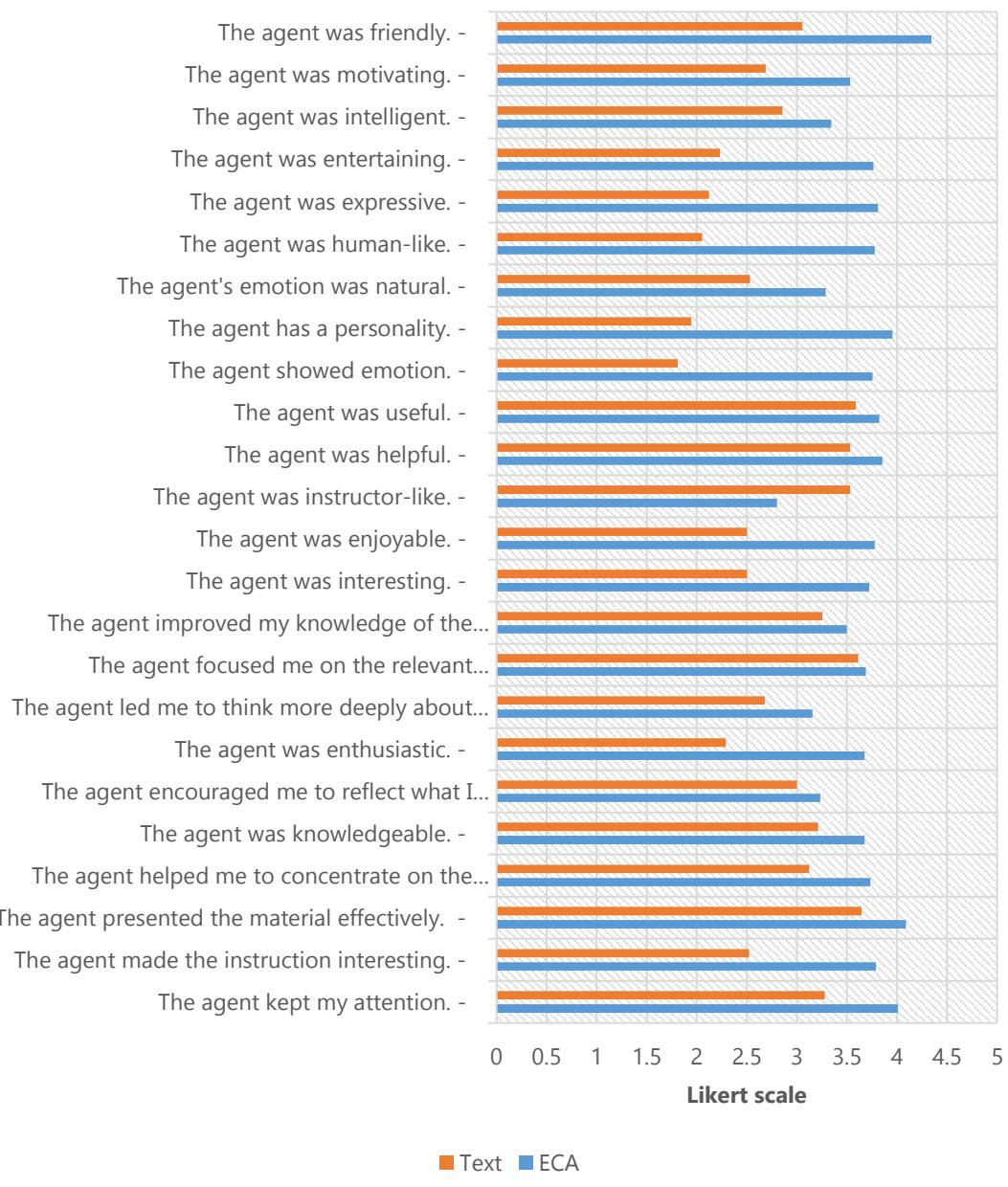


Figure 33-API Profile – API items for the collaborator agent and the difference between designs.

As seen in Table 35, Text agent scored below average in 11 attributes (made the instruction interesting, enthusiastic, made me think more deeply about the presentation, interesting, enjoyable, natural emotion, human-like,

expressive, entertaining, intelligent, motivating, friendly) and over agree in only 1 (kept my attention). ECA agent scored above average in all attributes apart from 1 (the agent is instruction like). It scored over agree in 1 attribute (kept my attention). The difference in the overall mean API scores of the two versions of the game could be attributed to all the items. Based on the literature review, this research focused more on the items of the Human-Like factor of the questionnaire (highlighted in orange) while 6 additional attributes (the ECA: made the instruction interesting, was not instructor-like, was expressive, was entertaining, was friendly and was human-like) were found to have the biggest difference. The attributes of interest have been discussed individually below.

API Attribute: Made the instruction interesting

"The agent made the instruction interesting." The data in Table 35 show that users reported feeling that the ECA agent made the instruction more interesting than the text agent. This can be attributed to the interactive role of the collaborator agent where an embodied agent with auditory output was found more interesting than a neutral text instruction. Also, participants commented that the text version was boring. The paired t-test confirmed the difference in these mean API scores to be statistically significant ($t=12.11$, $df=89$, $p=0.000$).

API Attribute: Human-like

"The agent was human-like." The data in Table 35 show that users reported that the ECA was more human like than the text agent. Participants commented that they felt like the text agent was less responsive. Also, for the purpose of this experiment, neutral language was used for the text agent, since it was evaluated how the personification of an agent changes the

perception of the agent's persona and the usability of the application. This is an indication that people treated the ECA in accordance with the anthropomorphic form theory that further support the illusion of humanness. This theory states that people have the tendency to find humanoid forms and human-like characteristics appealing. The paired t-test confirmed the difference in these mean API scores to be statistically significant ($t=13.73$, $df=89$, $p=0.000$).

API Attribute: Emotion

"The agent showed emotion". Again, participants reported that the ECA was more emotive than the text which can be attributed to the personification of the agent. This result is connected to the media equation theory (participants treated the ECA in a social manner), the persona effect theory and supports the illusion of humanness. The paired t-test confirmed the difference in these mean API scores to be statistically significant ($t=17.88$, $df=89$, $p=0.000$).

API Attribute: Personality

"The agent has a personality." The data in Table 35 show that users found that the ECA had more of a personality compared to the Text version. Again, this result is connected to the media equation theory and the persona effect theory while it further supports the illusion of humanness effect. The paired t-test confirmed the difference in these mean API scores to be statistically significant ($t=18.87$, $df=89$, $p=0.000$).

API Attribute: Natural emotion

"The agent's emotion was natural." According to the data in Table 34, users reported that the ECA had a more natural emotion than the Text version. Again, this is linked to the persona effect and the control factor of this

experiment where the text agent was neutral. The paired t-test confirmed the difference in these mean usability scores to be statistically significant ($t=4.87$, $df=89$, $p=0.000$).

API Attribute: Expressive

"The agent was expressive." In terms of expressiveness, participants reported that the ECA was more expressive and interacted like a real transaction and this can be justified by the animation of the agent which mimicked a real-life person. The paired t-test confirmed the difference in these mean usability scores to be statistically significant ($t=12.51$, $df=89$, $p=0.000$).

Instructor agent

Questionnaire statement	ECA (Mean =)	TEXT (Mean =)	t	df	p.
The agent kept my attention. -	3.87	3.08	6.01	89	.000
The agent made the instruction interesting. -	3.48	2.63	6.40	89	.000
The agent presented the material effectively. -	4.17	3.59	5.01	89	.000
The agent helped me to concentrate on the presentation. -	3.84	3.24	4.83	89	.000
The agent was knowledgeable. -	3.96	3.44	5.70	89	.000
The agent encouraged me to reflect what I was learning. -	3.49	2.99	3.75	89	.000
The agent was enthusiastic. -	3.10	2.44	4.80	89	.000
The agent led me to think more deeply about the presentation. -	3.27	2.73	4.40	89	.000

Questionnaire statement	ECA (Mean =)	TEXT (Mean =)	t	df	p.
The agent focused me on the relevant information.	4.09	3.70	4.06	89	.000
The agent improved my knowledge of the content.	3.83	3.34	3.47	89	.001
The agent was interesting. -	3.24	2.61	5.15	89	.000
The agent was enjoyable. -	3.29	2.62	6.02	89	.000
The agent was instructor-like. -	4.27	3.80	3.90	89	.000
The agent was helpful. -	4.12	3.67	4.67	89	.000
The agent was useful. -	4.00	3.79	2.07	89	.041
The agent showed emotion. -	2.92	2.01	7.95	89	.000
The agent has a personality. -	3.10	2.11	7.67	89	.000
The agent's emotion was natural. -	2.99	2.52	3.40	89	.000
The agent was human-like. -	3.20	2.22	7.30	89	.000
The agent was expressive. -	3.16	2.16	8.04	89	.000
The agent was entertaining. -	2.89	2.46	3.43	89	.001
The agent was intelligent. -	3.54	2.94	6.09	89	.000
The agent was motivating. -	3.44	2.77	5.84	89	.000
The agent was friendly. -	3.72	3.00	5.85	89	.000

**Table 36-Mean scores and results of paired t-tests on Individual Agent Persona
Instrument for version – Instructor agent.**

As shown in Table 36, all API items became again highly statistically significant. Therefore, it is concluded that there was a big difference between versions overall.

The difference on the API Likert scale between the two designs for the instructor agent is illustrated in Figure 34. The HECA version of the collaborator scored higher than the text version on all cases.

Instructor agent-API

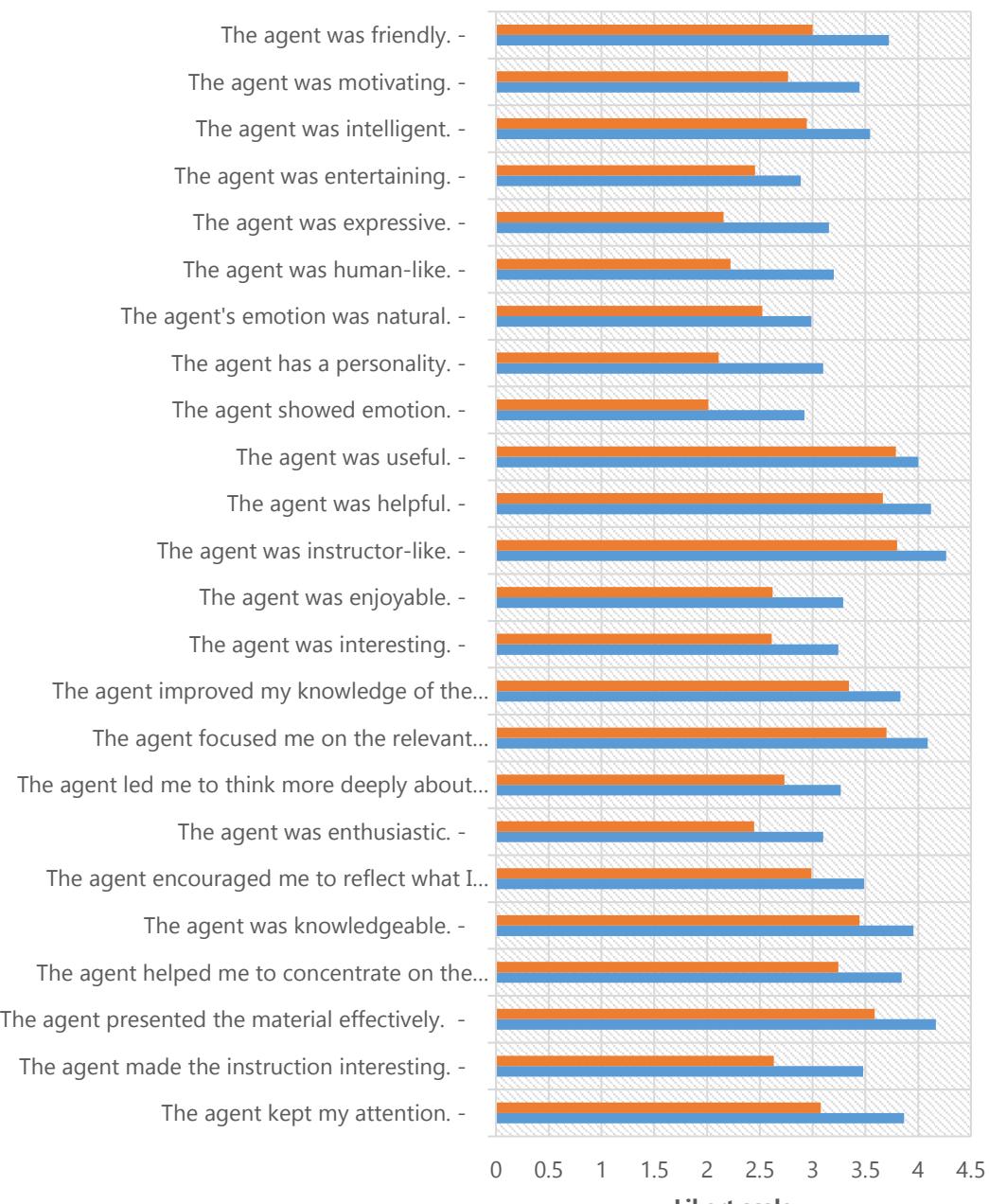


Figure 34-API Profile – API items for instructor agent.

As seen in Table 36, Text version scored below average in 14 attributes (made instruction interesting, encourage to reflect, enthusiastic, think more deeply, interesting, enjoyable, emotional, has personality, natural emotion, human-like, expressive, entertaining, intelligent, motivating, friendly) and over agree in none. The ECA version scored overall above average apart from 3 attributes (emotion, natural emotional, entertaining). It scored above agree in 5 attributes (presented the material effectively, focus on the information, helpful, useful, emotive).

Based on the literature review, there was a focus more on the items of the Human-Like factor of the questionnaire (highlighted in orange). Those attributes have been discussed individually below.

API Attribute: Emotion

"The agent showed emotion." The data in Table 36 show that users reported that the ECA agent showed more emotion than the Text agent. This is connected to the persona effect and is an indication of the illusion of humanness effect. Also, the control factor of the experiment (text agent) lacked any personality. The paired t-test confirmed the difference in these mean usability scores to be statistically significant ($t=7.95$, $df=89$, $p=0.000$).

API Attribute: Personality

"The agent has a personality." The data in Table 36 show that users reported that the Text agent lacked personality compared to the ECA agent. Again, this is connected to the persona effect and further evidence for the illusion of humanness. Also, it connects with the control factor of the experiment where the text agent lacked any personality while the ECA agent mimicked a real

person. The paired t-test confirmed the difference in these mean usability scores to be statistically significant ($t=7.67$, $df=89$, $p=0.000$).

API Attribute: Natural Emotion

"The agent's emotion was natural.". The data again show that the emotion of the ECA agent was more natural than the Text agent. The paired t-test confirmed the difference in these mean usability scores to be statistically significant ($t=3.40$, $df=89$, $p=0.000$). This means they felt less embarrassed playing the game with the ECA.

API Attribute: Human-Like

"The agent was human-like." The data in Table 36 show that users reported that the ECA agent was more human like than the Text agent. This is justified since the ECA could communicate information through linguistic and extralinguistic channels which is associated and is more common with human to human interaction. This result can be attributed to the anthropomorphic form theory and the persona effect which is further evidence of the illusion of humanness effect. The paired t-test confirmed the difference in these mean usability scores to be statistically significant ($t=7.30$, $df=89$, $p=0.000$).

5.5.1.3 **Research question 3: Hierarchical Multiple Regression Analysis**

R3: Which factors relating to the HECA's persona attributes account for variability in usability, and to what extent?

- Identify the extent to which attributes of the ECA (based on API for each agent) contribute to usability (positively or negatively).

- This research question will be answered by identifying the key drivers and examining their coefficients derived from the regression analysis.

Data analysis plan

There are multiple ways to do multiple linear regression such as cross validation, penalised methods or theory based on previous research. No prior research has studied the relationship of the agent's persona and usability; therefore, variable selection could not be based on previous research. It is though considered that best models derive from theory: "It is our experience and strong belief that better models and a better understanding of one's data result from focussed data analysis, guided by substantive theory" (Judd, et al., 2009). Since this research focuses mostly on the affective effect of the HECA using the API instrument, the variables selected for the model belong to the "Emotive interaction" latent variable; this variable is subdivided into the "Human-like" factor and the "Engaging" factor (Figure 35). According to Baylor (Baylor and Ryu, 2003) who developed the instrument, "The characteristics of the Engaging factor represent the social richness of the communication channels (Whitelock et al., 2000) and play an important role to provide 'personality' to the agent and enhance the learning experience", while "the Human-like factor of pedagogical agent persona is what makes it figuratively 'real'. Thus, both the Human-like factor and Engaging factors shape the pedagogical agent's social presence and personality". That limits the number of predictors to 9 ("The agent was human-like", "The agent was entertaining", "The agent was friendly", "The agent has a personality", "The

agent showed emotion", "The agent emotion was natural", "The agent was enthusiastic", "The agent was expressive" and "The agent was motivating").

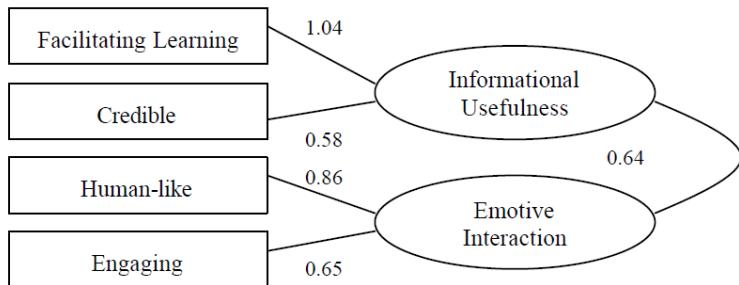


Figure 35-Factors and latent factors as presented by the author of the API questionnaire. (Baylor and Ryu, 2003)

Sample size in regression

An *a priori* sample size calculation for multiple regression was performed. Based on the rule of thumb that 10 to 15 samples are needed per predictor, 90 samples for 9 predictors should suffice (Tabachnick and Fidell, 2001). Input and output data for this research can be found in Table 37. More information on sample sizes in regression can be found in Chapter 3, section 3.5.2.

F tests - Multiple Regression: Omnibus (R^2 deviation from zero)

Analysis: A priori: Compute required sample size

Input:	
Effect size f^2	0.20
α err prob	0.05
Power (1- β err prob)	0.8
Number of predictors	9
Output:	
Noncentrality parameter λ	17.600000
Critical F	2.002245
Numerator df	9
Denominator df	78
Total sample size	88
Actual power	0.805798

Table 37-A priori sample size calculation for regression analysis.

Multiple linear regression

In this research, the ordinary least squares (OLS) full model is used with 9 items as predictors and the usability mean value for the shopkeeper agent.

The method used is the hierarchical multiple linear regression, since from theory the “Human-like” factor is more relevant (Model 1: 4 predictors) and is followed by the “Engaging” factor (5 predictors). Model two is a combination of the “Human-like” and “Engaging” attributes and includes the following variables: “The agent was human-like”, “The agent was entertaining”, “The agent was friendly”, “The agent has a personality”, “The agent showed emotion”, “The agent emotion was natural”, “The agent was enthusiastic”, “The agent was expressive” and “The agent was motivating”.

Results for the shopkeeper- collaborator agent

Descriptive Statistics

The descriptive statistics for the predictors used in the model are presented in Table 38. The skewness and kurtosis for each variable were examined with indices for acceptable limits of ± 2 used [37,38,39,13] one predictor variable was skewed. That is a mere indicator of non-normality though, since skewed data often occur due to lower or upper bounds on the data such as Likert data produce (NIST, 2017).

Upon further investigation, all the predictors were normally distributed apart from item 24 (The agent was friendly), while the box plots of items 17 and 21 were not balanced but the Stem-Leaf plots, Q-Q plots and histograms indicated a normal distribution. Thus, the data were treated as normal and were analysed parametrically.

Descriptive Statistics

	N Statistic	Minimum Statistic	Maximum Statistic	Mean Statistic	Std. Deviation Statistic
ASK_ECA24The agent was friendly. -	90	1	5	4.34	.673
ASK_ECA23The agent was motivating. -	90	2	5	3.53	.837
ASK_ECA21The agent was entertaining. -	90	1	5	3.77	1.028
ASK_ECA20The agent was expressive. -	90	1	5	3.81	.982
ASK_ECA19The agent was human-like. -	90	1	5	3.78	.957
ASK_ECA18The agent's emotion was natural. -	90	1	5	3.29	1.073
ASK_ECA17The agent has a personality. -	90	2	5	3.96	.873
ASK_ECA16The agent showed emotion. -	90	1	5	3.76	.916
ASK_ECA7The agent was enthusiastic. -	90	1	5	3.68	.934
Valid N (listwise)	90				

Table 38-Descriptive statistics.

Multiple Linear Regression

Assessing the regression model I: diagnostics

No outliers and residuals were identified. Also, upon further examination for influential cases, none were detected. See Appendix E for full analysis.

Assessing the regression model II: generalisation

How much of the Usability can be explained by the 9 API attributes?

The relevant assumptions of this analysis were tested prior to the multiple regression analysis.

In a summary, no multivariate outliers existed; the assumption of non-zero variance was met as the predictors vary in value; the assumptions of linearity, homoscedasticity and normality were met; the assumption for independent errors was deemed to be inconclusive; the assumption of multicollinearity has been met; the data were suitably correlated with the dependent variable in order to be examined with multiple linear regression. Details on the tests of assumptions can be found in Appendix E.

ANOVA table

The improvement in prediction that results from fitting the model is statistically significantly greater than the inaccuracy within the model for both models as seen in Table 39.

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.752	4	2.188	4.276	.003 ^b
	Residual	43.493	85	.512		
	Total	52.245	89			
2	Regression	11.954	9	1.328	2.637	.010 ^c
	Residual	40.291	80	.504		
	Total	52.245	89			

a. Dependent Variable: ECA_MEAN

b. Predictors: (Constant), ASK_ECA19The agent was human-like. -, ASK_ECA16The agent showed emotion. -, ASK_ECA18The agent's emotion was natural. -, ASK_ECA17The agent has a personality. -

c. Predictors: (Constant), ASK_ECA19The agent was human-like. -, ASK_ECA16The agent showed emotion. -, ASK_ECA18The agent's emotion was natural. -, ASK_ECA17The agent has a personality. -, ASK_ECA23The agent was motivating. -, ASK_ECA24The agent was friendly. -, ASK_ECA7The agent was enthusiastic. -, ASK_ECA21The agent was entertaining. -, ASK_ECA20The agent was expressive. -

Table 39-ANOVA for shopkeeper- interaction partner agent.

Model parameters

For model 1 (“Human-Like” predictors), the strongest and the only statistically significant ($p. = 0.008$) predictor was “The agent was human-like” ($\beta = .39$). In model 2 (full model with all 9 predictors from both “Human-Like” and “Engaging” factors), two were the most statistically significant predictors, “The agent was human-like” ($\beta = .4$) ($p. = 0.010$), “The agent was entertaining” ($\beta = .03$) ($p. = 0.05$) (see Table 40).

	<i>B</i>	<i>SE B</i>	<i>t</i>
Model 1			
<i>Constant</i>	4.06	0.38	
<i>The agent showed emotion</i>			
	-0.13	0.12	-.15
<i>The agent has a personality</i>			
	0.15	0.13	.18
<i>The agent's emotion was natural</i>			
	-0.01	0.09	-0.02
<i>The agent was human-like</i>	0.31	0.11	.39**
Model 2			
<i>Constant</i>	4.08	.51	
<i>The agent showed emotion</i>			
	-0.15	0.14	.18
<i>The agent has a personality</i>			
	0.84	0.14	.09
<i>The agent's emotion was natural</i>			
	-0.02	0.9	-0.03
<i>The agent was human-like</i>	0.3	0.1	0.4**
<i>The agent was enthusiastic</i>			
	-0.06	0.1	-0.07

	B	SE B	β
<i>The agent was expressive</i>	-0.05	0.1	-0.06
<i>The agent was entertaining</i>	0.2	0.1	0.3**
<i>The agent was motivating</i>	0.13	0.12	0.14
<i>The agent was friendly</i>	-0.1	0.15	-0.09

a. Dependent Variable: Usability mean score for embodied conversational agent version

Note: * $p < .10$, ** $P < .05$, *** $p < .001$. $n=90$

Table 40-Hierarchical Multiple Regression Analyses for the Shopkeeper Agent of the Embodied Conversational Agent Version.

Effect size

For multiple regression the formula to calculate the effect size is:

$$F^2 = R^2 / (1 - R^2)$$

Equation 1-Cohen's formula for calculating effect size in multiple regression (Selya, et al., 2012).

In this case, Cohen's formula gives an effect size $f^2 = 0.297$.

This represents a moderate to large effect according to Cohen's guidelines (Cohen, 1988).

Research question 3: How much of the variability in the usability can be accounted for by the predictors of the shopkeeper/ collaborator agent?

For the first model, the 4 independent variables from the "Human-like" factor produced an effect size R^2 of .17 ($F(4,85) = 4.28, p = .003$) which means that the "Human-like factors" accounted for 17% of the variation in ECA Usability. However, for the final model and all 9 predictors, this value increased to 0.229 ($F(9,80) = 2.64, p = .010$) or 23% of the variation in ECA Usability. Therefore, whatever variable entered the model in block 2 and the "Engaging" factors accounted for an extra 6% of the variance. The adjusted R^2 shows how well the model can be generalised. It was 0.13 for the first model and 0.142 for the second model which implies that the model with all 9 predictors includes some non-important variables that add noise to the model.

Results for the Alex- instructor agent

Descriptive Statistics

The descriptive statistics for the predictors used in the model are presented in Table 41. The skewness and kurtosis for each variable were again examined where indices for acceptable limits of ± 2 were used (Trochim and Donnelly, 2006; Tibshirani and Hastie, 2016; Gravetter and Wallnau, 2014; Field, 2013). No predictors had skewness or kurtosis issues.

Upon further investigation, all the predictors were normally distributed with Stem-Leaf plots, Q-Q plots and histograms verifying a normal distribution. Thus, the data were treated as normal and analysed parametrically. Again, the predictors entered the regression hierarchically with the 4 "Human-like" predictors for model 1 followed by the 5 "Engaging" predictors for model 2.

Descriptive Statistics

	Mean	Std. Deviation	N
ECA_MEAN	5.3173	.76617	90
AA_ECA16The agent showed emotion. -	2.92	1.008	90
AA_ECA17The agent has a personality. -	3.10	.995	90
AA_ECA18The agent's emotion was natural. -	2.99	1.022	90
AA_ECA19The agent was human-like. -	3.20	1.041	90
AA_ECA7The agent was enthusiastic. -	3.10	.972	90
AA_ECA21The agent was entertaining. -	2.89	.929	90
AA_ECA23The agent was motivating. -	3.44	.751	90
AA_ECA24The agent was friendly. -	3.72	.750	90
AA_ECA20The agent was expressive. -	3.16	1.005	90

Table 41-Descriptive statistics

Multiple Linear Regression

Assessing the regression model I: diagnostics

No outliers and residuals were identified. Also, upon further examination for influential cases none were detected. See Appendix E for full analysis.

Assessing the regression model II: generalization

How much of the Usability can be explained by the 9 API attributes?

The relevant assumptions of this analysis were tested prior to the multiple regression analysis.

In a summary, no multivariate outliers existed; the assumption of non-zero variance was met as the predictors vary in value; the assumptions of linearity, homoscedasticity and normality were met; the assumption for independent errors has been met; the assumption of multicollinearity has been met; the data were suitably correlated with the dependent variable in order to be examined with multiple linear regression. Full analysis of assumptions can be found in Appendix E.

ANOVA table

The improvement in prediction that results from fitting the model was statistically significantly greater than the inaccuracy within the model for both models as seen in Table 42.

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10.548	4	2.637	5.375	.001 ^b
	Residual	41.697	85	.491		
	Total	52.245	89			
2	Regression	14.947	9	1.661	3.562	.001 ^c
	Residual	37.298	80	.466		
	Total	52.245	89			

a. Dependent Variable: ECA_MEAN

b. Predictors: (Constant), AA_ECA19The agent was human-like. -, AA_ECA18The agent's emotion was natural. -, AA_ECA16The agent showed emotion. -, AA_ECA17The agent has a personality. -

c. Predictors: (Constant), AA_ECA19The agent was human-like. -, AA_ECA18The agent's emotion was natural. -, AA_ECA16The agent showed emotion. -, AA_ECA17The agent has a personality. -, AA_ECA23The agent was motivating. -, AA_ECA24The agent was friendly. -, AA_ECA21The agent was entertaining. -, AA_ECA7The agent was enthusiastic. -, AA_ECA20The agent was expressive. -

Table 42-ANOVA table for Alex-instructor agent.

Model parameters

For model 1, the strongest predictor that was statistically significant was "The agent was human-like" ($\beta = .47$). For model 2, the strongest predictor was "The agent was entertaining" (see Table 43).

	B	SE B	B
Model 1			
Constant	4.31	0.28	
<i>The agent showed emotion</i>	0.11	0.10	.15
<i>The agent has a personality</i>	-0.19	0.12	-.25
<i>The agent's emotion was natural</i>	0.05	0.1	0.07
<i>The agent was human-like</i>	0.35	0.11	.47**
Model 2			
Constant	4.19	.42	
<i>The agent showed emotion</i>	0.07	0.11	.09
<i>The agent has a personality</i>	-0.23	0.12	-.30
<i>The agent's emotion was natural</i>	-0.01	0.1	-0.11
<i>The agent was human-like</i>	0.11	0.28	0.38**
<i>The agent was enthusiastic</i>	0.08	0.11	0.11

	B	SE B	β
<i>The agent was expressive</i>	-0.05	0.11	-0.07
<i>The agent was entertaining</i>	0.30	0.11	0.36**
<i>The agent was motivating</i>	0.02	0.12	0.02
<i>The agent was friendly</i>	-0.07	0.12	-0.07

a. Dependent Variable: Usability mean score for embodied conversational agent version

Note: * $p < .10$, ** $P < .05$, *** $p < .001$. $n=90$

Table 43-Hierarchical Multiple Regression Analyses for Alex Agent of the Embodied Conversational Agent Version.

Effect size

In this case Cohen's formula yields an effect size $f^2 = 0.4$.

This represents a large effect according to Cohen's guidelines (Cohen, 1988).

Research question 3: How much of the variability in the usability can be accounted for by the predictors of the Alex/instructor agent?

For the first model, the 4 independent variables from the "Human-like" factor produced a R^2 of .20 ($F(4,85) = 5.37, p = .001$) which means that the

"Human-like factors" accounted for 20% of the variation in ECA version Usability. However, for the final model and all the 9 predictors, this value increased to 0.29 ($F(9,80) = 3.56, p = .00$) or 29% of the variation in ECA version Usability. Therefore, whatever variable entered the model in block 2 and the "Engaging" factors accounted for an extra 9% of the variance. The adjusted R^2 was 0.16 for the first model and 0.21 for the second model which implies that not all the predictors contributed to the model significantly.

5.5.2 Qualitative analysis

After interacting with each version, participants were asked to comment on their experience with the application and then specifically on each version. All the answers for the open-ended questions were analysed using thematic analysis³⁸.

For this experiment, the tutorial was experienced once in the beginning of the session as standalone.

The first question of the exit questionnaire was referring to the tutorial. The participants were asked what their thoughts on the tutorial were, if they found it useful and if they understood the money system. In the first question "What did you think of the tutorial?" most participants replied positively with 62 stating that it was good, clear and easy to understand. Twenty-two participants valued the information presented (e.g. "The tutorial was interesting and informative"), while four responded that it was well explained and worked well. There were a few negative comments about the tutorial as

³⁸ Braun and Clarke (2006) define thematic analysis as:
"A method for identifying, analysing and reporting patterns within data."

well, such as that it was slow and long according to nine participants; that the money system was confusing according to two participants and that it was borderline annoying according to one participant. Some of the comments given by the participants were:

- “It was clear and provided the relevant information.”
- “Very informative with nice graphics.”
- “It was informative, straight forward and slow.”
- “It was interesting. Did not know about the old money.”
- “I found it informative.”
- “Borderline annoying.”

In the questions “Did you find the tutorial helpful?” all of the 90 participants answered “yes” confirming the changes made after the pilot study to be positive. In the question “Do you feel you understood the old money” only one person out of 90 stated that they did not without giving further explanation. The rest of the participants replied that it was successfully explained.

5.5.2.1 Explicit Preference

After experiencing both versions, participants were asked which version of MoneyWorld they preferred. They were asked to give their answer in terms of their first or second version experienced, and the answers were re-ordered for each version.

Eighty-one participants (90%) stated that they preferred the ECA version, eight participants (8.9%) stated that they preferred the text version and one participant (1.1%) had no stated preference.

Participants were also asked to give reasons for their answer. The majority of comments about the ECA version mentioned that characters were more fun (20); they preferred interacting with a human/character (16); it was more human-like and natural (14); it was more interactive (27 participants); it was easier (17); and the text version was boring and added cognitive load (15).

Some sample comments made by participants are:

- “The interaction with humans makes the game engaging.”
- “I prefer the human voice over the text.”
- “Having human-like characters makes it more captivating and enjoyable.”
- “It was more interactive, and it had sound.”
- “The ECAs were more engaging and fun. It was easier for me to understand the instructions. The shopkeeper was engaging and had fun, more interactive reactions. The text was boring.”
- “The shopkeeper made me feel relaxed. It was more interactive and enjoyable.”
- “I felt I was very familiar, and it was easy to deal with it. I was interacting with a human, so communication was easy.”
- “It was easier to interact with voice. It was enjoyable. Reading was boring for a game.”

For the Text version, participants commented that it was clearer (2), reading was faster than the ECA (2) or that they preferred the text version (4).

Example comments are:

- "I prefer to read because it is faster."
- "Reading didn't take as long as the ECA version."
- "It was clearer to see the text. In voice sometimes, the accent was confusing."
- "It was clear (I am a non-native speaker)."

5.5.2.2 **Agents**

Participants were asked after each version of MoneyWorld they experienced what they thought about the various agents they came across during the session. The following section details the opinions given for the agents, irrespective of the order and the data combined.

ECA version

Before asked about the shopkeeper, participants were presented with a laminated picture of the shopkeeper as it appeared on the screen during the shopping task and asked "The interface that you interacted with in order to buy the items on the list looked like this (show laminated picture of ECA shopkeeper). What did you think about it?". The comments were overall positive. Even though the question did not refer to the agent as "He" but rather asked what they thought about it, most participants commented on

the human characteristics of the agent. Some participants (23) thought that the shop-keeper was human-like. Several participants (16) characterised the agent as friendly while others as funny or fun to interact with (26). Some participants (16) commented that they liked him or liked interacting with him and five said that having a person to interact with made the experience better. Example comments made are:

- "I could imagine how he would be in real life. It was a realistic, human-like character."
- "It was human-like. I liked interacting with someone and receiving positive feedback."
- "Having human-like characters makes it more captivating and enjoyable."
- "I felt I was very familiar, and it was easy to deal with it. I was interacting with a human, so communication was easy."
- "The shopkeeper was engaging and had fun, more interactive reactions."
- "It was more fun and more like a real-life experience."
- "He added a personality. It was fun and interesting to interact with him. He gave funny comments."

Seven participants made negative comments on the ECA which mainly had to do with the uncanny valley theory and the face animations but were accompanied by some positive comments like he was fun or friendly. A couple of examples would be:

- "He was friendly, but he had a worrying expression."
- "He was interesting, funny and human-like. He was friendly, but he had a worrying expression."

Participants were asked their opinion on Alex where they were also presented with a picture of the agent as it was presented in the game. Most comments on Alex were positive with 18 participants reporting that they liked her voice and they could focus better due to the voice; nineteen participants identified her role in the interaction as the agent that gave instructions and their perception was positive as they felt Alex was helpful; seven thought she was human-like, while eight stated that the addition of character was better as it made it more natural or easier to focus; thirteen commented that the interaction was more interesting and fun; and 12 that it was more clear.

Example comments made are:

- "More interesting, less boring, human-like."
- "Because of the voice I was able to perceive emotion. I think this is a better way to receive instructions."
- "The voice along with the visuals was more effective."
- "It was good. It gave instructions."
- "She was nice and friendly. She was encouraging and gave clear instructions."
- "She was more instructive than the text."
- "I found it fun and I enjoyed it. Had awkward animation though."
- "It was interesting and fun, unlike the text version which was blunt."
- "It was helpful and human-like. It felt like a real interaction."

Fourteen comments that were made were on the negative side. Most had to do with the lip synching that was lacking or that she came across as robotic,

her face was distracting or was emotionless and that she did not add much to the game. A couple of examples would be:

- “Her lip synchronisation was not good and this made her funny. She came across as an emotionless robot.”
- “She was creepy and unnecessary. I do not think that she added anything (any value).”
- “Not very useful. Hearing would be fine. The lip synch was off. Nice voice.”
- “It was clear because you get the information. The agent is distracting. It may be boring, but it is clear.”

Text version

While focusing specifically on the text version, participants were asked their opinion on the agents. Similar to the ECA version, before asked about each agent, participants were presented with laminated pictures of the agents as they were presented in the game.

In the question “The interface you interacted with in order to buy the items on the list looked like this (show text-based shopkeeper). What did you think about it?” only a few comments were positive. Some participants (15) answered that it was clear, straight forward or direct although not human-like or emotional. A few (12) had a lukewarm reaction towards the text shopkeeper by saying that it was fine, good or ok but not engaging. Only six thought that it was easy, five liked it, two thought it was helpful and one said that it helps them focus. Example comments made are:

- "Good but poor compared to the ECA version which was an improvement."
- "It was straightforward, clear but not emotional."
- "It was clear because you get the information. The agent (ECA) is distracting. It may be boring, but it is clear."

Most comments regarding the text version of the shopkeeper were underwhelming and negative. Ten said it was boring and less entertaining.

Other comments suggested that they had to concentrate hard to remember the prices and was stressful (14); the text agent was less engaging (eight); it was frustrating to use (eight); and it was confusing (five). A few examples of comments:

- "It was stressing for me to read it. It was more difficult to remember the prices."
- "It was better to have the ECA. I was not sure when I needed to speak."
- "I found it easier to understand the task but less engaging, less entertaining and unrealistic. I thought I had to type answers; it felt like a chat."
- "It was boring. I did not feel immersed. I was frustrated."
- "It was a bit boring. You had to focus."
- "It was acceptable, although I found it emotionless and not encouraging. I didn't like talking with no sound."
- "I got nervous when the text went away because I had to remember. I was not immersed but I was more concentrated on the task. It was a very mechanical experience like an exam."
- "It was clear, but I was a bit confused when the interface did not pick up my voice. I felt frustrated."

In a similar way, participants were asked their opinion on the text version of Alex the instructor. Participants were presented with a screenshot of the interface as it appeared in the game. In the question "The interface that assist you with the list looked like this (show text-based Alex). What did you think about it?", 14 participants said that it was instructor like and 15 said that for giving instructions it was clear. Only one participant answered that it was engaging. A couple of examples would be:

- "It was clear but not interesting."
- "It was helpful, but it was not clear that I had to speak."
- "It was OK. It had no interaction so no difference."
- "It was very effective for instructions only."
- "It was better than the SK text because it was a non-interactive role."
- "It was boring. I did not feel immersed. I was frustrated."

Agent preference

After having been asked about their thoughts on each agent they faced during the game, participants were asked to explicitly state which agent they preferred in each role. Again, participants were presented with screenshots of all the four agents to choose from.

In the question "Which system did you prefer to interact with on the shop?", 76 participants (84.5%) preferred the ECA version, 13 the Text version (14.5%) and one (1%) had no preference. Participants justified selecting the ECA version of the shopkeeper saying they found him more interactive, entertaining, it made the interaction more natural and real, the addition of voice helped them concentrate better and they could focus better. Some of the comments were:

- "Really liked him. He was polite and funny."
- "Because of the voice. For such a small screen, the voice was better than the text."
- "The characters were likable giving another dimension to the instructions"
- "It was easier to keep my attention."
- "It made it seem natural and interactive. I didn't have to focus as much."
- "It was very entertaining. I felt like I was talking to someone."
- "It felt more like a character, a human."

Those who preferred the text version of shopkeeper gave comments such as that it was straight forward, quicker and less distracting.

- "Not getting in the way. Character was distracting from the task at hand."
- "More efficient."
- "The SK was fun, but he was distracting me from understanding and remembering."
- "You can see the price."
- "Text was quicker."
- "I prefer the text system or the ECA with subtitles. The ECA was slow."

In the question "Which system did you prefer to be assisted from?", 67 preferred the ECA version (74.5%), 18 the text version (20%) and five (5.5%) had no preference. Participants who preferred the ECA version of Alex elaborated on their response by saying that the version with the character was more enjoyable and felt more interactive; it was easier to concentrate and understand the instruction because of the voice; it mimicked human to human interaction and added character; and that unlike the text, the agent made the application feel more like a game. A few of the comments were:

- "Felt more interactive. I felt connected. Versions with characters was more enjoyable and amusing."
- "It had sound and it was easier to understand than reading the text."
- "Person seems more welcoming. Make you want to play the game."
- "The text reading was not natural; it does not give the feeling of a game, while the character does."
- "More interesting, the voice keeps your attention more than the text."
- "I like people better. It was clear, more interactive, more like a game."
- "She was more helpful, easier, looking at a person and listen rather than reading and processing."
- "It was clearer what I had to do and having sound made it more intuitive."

The participants who preferred the text version of Alex justified their choice with comments such as that having a character did not add to the interaction and it was distracting because the role of the agent was to give instructions. A few of the comments referred to the fact that her facial expressions (ECA Alex) were weird and it was distracting. Also, a few commented that reading instructions was easier or quicker and text was enough for instructions. Some example comments follow:

- "The characters had a robotic looking and this was distracting."
- "It was clearer for instructions."
- "The text helps me understand better since there was minimum interaction."
- "The text-only system gave the information you needed. The use of characters felt unnecessary."
- "I liked the text for instructions because it was enough."
- "It was very straightforward. The ECA was not very communicative and did not look natural."

Use of agents

Finally, participants were asked if they used agentassistants on their phone and their opinion on speech interfaces and natural language interaction.

Again, the answers were organised and analysed for recurring themes.

The first question participants were asked was: "Do you use assistants/agents such as Siri/Cortana/Speaktoit on your smartphone in your everyday life?".

The majority (48) stated that they do not use agents on their phone, 30 said that they use agent sometimes, nine answered that they use agent every day and three did not own a smartphone. Out of those who use agents, 22 use Siri, nine Ok Google, four Cortana, two Duolingo, one Google now and one S voice. When asked for what tasks they used agents, 14 answered for fun, 11 for web searching, six for checking the weather, five for calendar and reminders, three for calls, three for setting the alarm, two for texting, two for language learning, two for finding their contacts, two for navigation and two for basic functions.

The next question was "What do you like about this kind of interface?" and "What do you dislike about this kind of interface?". In terms of what participants like, 26 participants responded that speech recognition systems are convenient for hands free situations, 12 said that it is faster than typing, ten answered that it is an easier type of interaction, seven said that it is a fun way to interact and five answered that it is a natural way to interact.

Some example comments follow:

- "I liked the usability, flexibility and hands-free mode. It can be used for emergency."
- "It was easier and hands free."

- "The advantages are that it is hands-free. I find the keys on the phone to be tedious."
- "Efficiency; it is faster than typing."
- "Speaking is faster and more human-like."
- "It can be faster and easier. There was voice output."
- "It was faster, futuristic and modern."
- "When the timing is right, they (agents) are easier to use."
- "It could be fun. You save time from typing."
- "Much more natural. It makes things intuitive."
- "Natural, easy and intuitive"
- "It can be funny, and it could be especially good for people with disabilities."

In the question "What do you dislike about this kind of interface?", 20 participants responded that speech recognition systems still have issues with picking up accents, 18 answered that using it in public would be embarrassing, 16 said that speech recognition systems need improvement as there are still many voice recognition issues that make the interface frustrating to use and 11 responded that they are used to do things manually. The main concerns for speech systems were privacy and that recognition is not optimal yet. A few of the comments were:

- "There are still issues with the accents and it is frustrating."
- "Currently it needs improvement as due to accents it is not very reliable."
- "I would be embarrassed in public. I am shy."
- "I would be embarrassed in public and I do not want to bother other people."
- "The voice recognition does not work well for people with an accent."
- "There were speech recognition issues, but not in native language (Korean)"

- "Sometimes it is easier to do it manually. It does not pick up the voice very well."

General observations

Most comments suggested that participants liked the ECA version system. A few interesting observations emerged from analysis:

- People could distinguish the role of the agents more easily when they experienced the ECA version.
- Even though it was not commented much, sound contributed a lot to the interaction as it was observed that participants responded more quickly while experiencing the ECA version.
- In the text version, participants responded as soon as they read the question. In the text version, the questions were presented, as seen in the screenshots, in a text box and the voice recognition is triggered after the question disappears from the screen; this made the application look as non-responsive.
- In the presence of a graphical interface and text user interface, the participants expected buttons instead of voice input thus trying to tap on the items.

For the text version of the shopkeeper, most participants commented that it did not feel immersive, it was not engaging, non-emotive, blunt and boring. A few positive comments mentioned that it was clear, informative, effective and good for non-native speakers who are used to reading subtitles.

For the text version of Alex, most comments were negative with participants characterising it as boring and not interesting, while a few positive comments were that it was instructor -like, clear and adequate for instructions.

When it came to the text version there were a couple of comments that stood out:

- One dyslexic participant preferred the ECA because text added cognitive load and they had to concentrate more. "It was quicker and easier. I am dyslexic, so I don't like reading. The text was boring and wooden."
- Only one participant commented on the small screen of the mobile phone. Even though having a character in such a small screen could be considered problematic, the comment indicated that having the ECA was better than the text. "Because of the voice. For such a small screen, the voice was better than the text."
- Non-native speakers preferred in some cases text because either the accent was confusing, or text was easier especially when they are used to using subtitles. Some of these comments were: "It was easier for non-native speakers because it is similar with using subtitles. However, it can be boring and outdated.", "It does not work well sometimes for non-native speakers. It uses a more formal language.", "Since I am a non-native speaker, I would have probably selected the text, but Alex was more interactive.",

For the ECA version of Alex, most users commented positively on her voice indicating that they focused more on the voice for the non-interactive agent that had the role of giving instructions.

For the ECA version of shopkeeper, participants gave positive comments on the agent's personality and believed that he added character to the interaction while also being friendly and fun.

5.6 Summary

This chapter presents the findings of a large-scale evaluation on the effectiveness of spoken HECAAs in a mobile serious game.

Results show that perceived usability was statistically significantly higher for the version with the ECAs compared to the neutral text version. The ECA version scored 5.32 while the text version scored 4.40. The effect size was also calculated in order to see if the effect is substantive. According to standard thresholds for Cohen's d, the calculated effect size of 1.01 is considered large thus suggesting a high practical significance.

When exploring the agents' persona as perceived by the user, data showed that the difference between the ECA and the text version was statistically significant for both agents with the ECA version scoring higher in both cases. The individual attributes that were the most significant for the shopkeeper/collaborator were: "The agent made the instruction interesting", "The agent was enthusiastic", "The agent showed emotion", "The agent has a personality", "The agent was human-like", "The agent was expressive", "The agent was entertaining" and "The agent was friendly". For the Alexa/instructor agent the most significant attributes were: "The agent showed emotion", "The agent has a personality", "The agent was human-like" and "The agent was expressive".

Upon further analysis, the multiple regression that was conducted in order to identify how much of the variability in usability can be explained by the API attributes, showed that the agents' entertaining, and human-like qualities contributed most to usability for both agents in the scenario.

Qualitative analysis supports the results obtained by the quantitative data with many participants referring to the ECAs as more fun to interact with, more human-like, more engaging, easier to use and making the transaction feel real.

Chapter 6 Discussion and Conclusions - Research Contributions and Design Implications with Respect to Embodied Conversational Agents in Mobile Serious Games

The following chapter hosts the main discussion on the work presented in this thesis. First, the findings of the preliminary work and the way they were incorporated into the main experiment are discussed. Following this, the findings of the main experiment are tackled. By revisiting the research contributions introduced in Chapter 1, the conclusions on the research questions are discussed by taking into consideration both quantitative and qualitative data. The next section moves on to explain the results in relation to theory. The chapter concludes by addressing the implications, limitations and recommendations for future work.

6.1 Introduction

This thesis provided evidence from two large-scale controlled usability experiments and one large scale technographic survey on the role of embodied conversational agents (ECAs) in desktop and mobile serious games (MSGs).

In the preliminary studies, the first experiment investigated the role of game elements as a means of feedback and the effects of serious gaming on overall usability of an application with ECAs (Chapter 4). The second study in the

preliminary work was a technographic survey that provided insight in the game-playing habits and the use of technological devices by a variety of adult users.

The main large-scale experiment (Chapter 5) was built upon the findings and methodological lessons of the preliminary studies and investigated the users' subjective attitudes towards two versions of a MSG and how spoken humanoid ECAs (HECAs) affect the usability and overall experience.

6.2 Key findings

As described in Chapter 1, the main drive behind this research was to advance the knowledge of ECAs' effectiveness in MSGs and contribute empirically to the area of mobile ECAs. The interface was specifically designed to allow the evaluation of communication efficiency and effectiveness between user and computer via multimodal interaction and especially speech recognition. The research showed that the illusion of humanness evoked by the addition of human-like ECAs had a positive effect on usability. The purpose behind the research strategy employed in this thesis was to provide design guidelines through empirical evidence about the effective inclusion of ECAs in MSGs.

A mixed-methods approach was adopted for the interdisciplinary investigation presented in this thesis. This methodology approach allows for evidence triangulation informed by previous theory, the users and the statistical models. Evidence was collected through a series of progressive evaluations based on the research themes of evaluating users' attitudes towards SGs and HECAs.

Table 44 provides a summary of the evaluations along with the main findings resulting from the analysis of the quantitative and qualitative data.

In Chapter 2, it was made apparent that little research has been conducted on the effectiveness of ECAs with the majority of the literature focussing on their design and implementation (Guo, et al., 2014). Also, it was highlighted that little attention has been paid to empirically evaluating their effectiveness and efficiency in SGs and MSGs (Doumanis, 2015). Moreover, no previous research has been found focussing on how the illusion of humanness evoked by the ECA contributes to usability especially on MSGs.

The first experiment focussed on the SG aspect and the introduction of game-like rewards as a form of feedback in a SG with ECAs. In the second evaluation, that of the technographic survey, game playing, and device-use data were collected in order to inform the design of the main experiment. The pinnacle of this thesis is the main evaluation, described in Chapter 5, where a mobile version of the SG was constructed to serve as a platform for assessing if and how the illusion of humanness affects the usability of the application and the effect of the agents on that experience.

	Type of evaluation	Research method	Medium	Evaluation Topic	Evaluation Findings
Preliminary work	Study 1: Usability evaluation	Mixed	PC	Serious game feedback	Even though not statistically significant, the serious game version with the explicit game-like feedback was preferred by participants and perceived as more fun and entertaining.
	Study 2: Technographic survey	Survey	Online	Game playing and devices survey	People use their mobile phones for most tasks, sometimes in conjunction with other devices. Most have smartphones with screen sizes over 5" and play games mostly on them. In the 6 months before answering the survey, 86,3% played digital games.
Main experiment	Usability and agent persona evaluation, regression analysis	Mixed	Mobile device	3D Embodied Conversational Agents: Usability and the "Illusion of humanness"	Research questions can be found in the section "Main experiment" in this chapter. R1: HECAs were rated statistically significantly higher in terms of usability compared to text agents with a large effect size. R2: HECAs were found to be more human-like and entertaining and less instructor like compared to the text agents. R3: The persona attributes that contributed more to usability for both agents were human-like and entertaining. Many participants attributed human-like cognitive and social skills to the HECAs.

Table 44-Summary of findings

6.3 Preliminary work

The preliminary work included two studies, one usability evaluation (pilot study 1) and one technographic survey (study 2).

The main aim of this evaluation was to act as a methodological sand box which would help decide the methodology approach adopted for the main experiment. Also, it aimed to establish that a SG is a suitable environment for the main experiment.

The first version was presented as a learning application with explicit feedback (learn version) while the second version was presented as a game with implicit feedback (scores and stars) (game version).

The results did not reveal a statistical significance between the two versions in terms of usability although both were rated positively (Game: 5.46/7 - Learn: 5.30/7). There was a tendency from the participants to rate the second version more favourably and the explanation might be that they already knew what to do as the first version included the tutorial. In order to avoid ordering effects, for the main experiment the tutorial was removed as part of the first version and was run once in the beginning of the session as a standalone. By examining the individual attributes, the: "I enjoyed using Moneyworld", "I thought Moneyworld was fun" and "I found the use of Moneyworld stimulating" were found as statistically significant in favour of the game version. From the exit interviews, these results can be attributed to the familiarity participants had with implicit rewards such as stars and scores and their association with games which is regarded as a fun activity. The association of such rewards with games made the Game version more appealing in terms of usability while many found the rewards appealing and

the application enjoyable. Also, participants found the rewards a good way to track their progress during the game. The qualitative data revealed that 78.5% of the participants preferred the Game version, 16.9% the Learn version and 4.6% had no preference. When justifying their choice, most mentioned the reward system of the Game version as more appealing while others found the Game version more stimulating and quicker – even though the two versions lasted the same time. These findings justified the use of the Game version as the basis of the main experiment.

The second study aimed to collect data in order to identify patterns on the participants' digital and game-play habits and insights on the use of mobile devices. The data collected showed that 86.3% of the participants had played games in the last 6 months with the majority (60.2%) playing on their smart phone. Fifty-five percent of the participants replied that their smartphone has a screen size between 5" and 6.9". An interesting observation was that they prefer using their smart phones for a plethora of activities (social media, email, navigation, organiser, photography etc.) over tablets apart from reading books or documents. The smaller screen can be deemed as the reason for that. The results from the second study informed the decision to use smart phones with a screen over 5" for the main experiment as it was the device of choice for playing games along with other activities.

6.4 Main experiment

The aim of the experiment was to investigate the users' subjective attitudes towards two versions of a MSG (Moneyworld) and how spoken HECA's can be used in this context. The objective of this experiment was to examine the extent to which the illusion of humanness evoked by a conversational agent

affects the usability of the MSG application and the users' attitudes towards agents with different roles. The focus on one version was for the conversational agents to be presented in the form of HECAs and the focus on the other was for the conversational agents to be presented in the form of a neutral text.

Through this experiment three research questions are being answered:

R1: To what extent do HECAs affect the usability of a mobile serious game (MSG)?

R2: To what extent do users perceive a difference in agent persona between ECA and neutral text presentation as measured by the agent persona instrument (API)?

R3: Which factors relating to the HECA's persona attributes account for variability in usability, and to what extent?

Research question 1:

To what extent do HECAs affect the usability of a mobile serious game (MSG)?

Among the participants who took part in this evaluation ($N = 90$), an overall statistical significance was found between the two versions, the HECA version

and the Text version, ($t=9.45$; $df=89$; $p.=0.000$). Therefore, the difference in usability scores between the two versions is statistically significant.

The overall mean for the HECA version was 5.32 (out of 7), indicating a positive attitude towards Moneyworld. The mean for the Text version was significantly lower at 4.40 which translates to just above neutral. Further, Cohen's effect size value ($d = 1.01$) suggested a high practical significance which means that the inclusion of an HECA in the MSG has a meaningful real-life impact on the usability.

The empirical evidence was supported by the qualitative data collected during the exit interviews. Eighty-one participants (90%) stated that they preferred the ECA version, eight participants (8.9%) stated that they preferred the text version and one participant (1.1%) had no stated preference.

By performing a paired t-test on the mean scores of each usability attribute in order to identify which attributes contributed to the difference between the versions, all were found to be statistically significant with the HECA version scoring higher in all cases.

All the HECA attributes scored above neutral with 3 scoring between neutral and agree (needs improvement, stimulation and immersion), over agree in 14 and over strongly agree in 1 (not too complicated to use). The text version scored below neutral in 3 attributes (frustration, needs a lot of improvement and immersion) and over agree in only 2 (not confusing to use and not too complicated).

The attributes that are of more interest have been discussed individually below.

Usability Attributes: Concentration and ease of use

"I had to concentrate hard to use Money world." The data show that users reported that they had to concentrate harder when using the Text version compared to the ECA version. The empirical data are supported by the qualitative data with 14 participants suggesting that during the text version they had to concentrate hard to remember the prices and was more stressful. This can be connected to the fact that reading from a screen can increase the extraneous cognitive load, while interacting with an ECA did not require to concentrate as hard as there were auditory and visual cues. The explanation is supported by Wik's (2011) previous work who claimed that through task-based interactive exercises with sound, pictures, agents and games, a more robust memory trace is created. The empirical data also support claims by Doumanis (2013) and Van Mulken (1998) that ECAs can improve cognitive functions and that by using ECAs the user can spend their cognitive resources on the primary task. Also, the results contradict one of the main arguments against ECAs, i.e. ECAs can lead to cognitive overload and distract from the main task because participants have to spend cognitive resources in processing visual and auditory information (Walker et al., 1994). The reduced cognitive load compared to the text version contributes to the ECA version by appearing easier to use and demanding less concentration. This is also a possible explanation why most participants replied that reading is the activity they use their smartphones least for, in the technographic survey.

Usability Attribute: Frustrating

"I found Money world frustrating to use." Users reported feeling more frustrated while using the Text version of the game compared to the ECA version. Participants also commented that they felt the Text version was less

responsive. While both versions were identical apart from the control factor, from the observations this can be explained by using the media equation theory. People responded to the questions at the appropriate time when they had visual and auditory cues from the ECA, while on the Text version people responded to the question as soon as they read it (speech input initiated when the question disappeared from the screen for the Text version and when the audio prompt for the question ended for the ECA version) thus making it look non-responsive. Further explanation is "ethopoeia" where people unconsciously apply social rules when interacting with virtual agents and the "illusion of humanness" which is the user's notion that the system possesses human attributes and/or cognitive functions thus responding to it in a social way. This confirms that people treated the ECA as they treat other people, in a social way. The justification for this is that they waited for the HECA to finish their question before answering and when the system was not responsive, they justified the HECA agent like they would do with someone who did not hear them properly. The qualitative data support the evidence with eight participants claiming that the text version was frustrating to use and five that it was confusing.

Usability Attribute: Embarrassed

"I felt embarrassed when using Money world". An interesting finding was that although participants reported quite often that they would feel embarrassed using a speech recognition system in public, both versions were rated relatively high although they felt less embarrassed playing the game with the HECA. A possible justification might be the "illusion of humanness" since the unconscious reaction is like that of conversing with a human thus making it less embarrassing.

Usability Attributes: Fun and enjoyed using

"I thought Money world was fun." Users rated the ECA version as more fun and enjoyable than the Text version. During the exit interview participants commented that the Text version felt outdated, while the ECA version felt more like a game and the graphics resembled more contemporary game. Also, many users commented that the Text version was more neutral, while the shopkeeper's comments and the more human-like interaction made the game more fun. Mulken et al. (1998), while empirically studying the persona effect, found that the presentation was perceived as less difficult and more entertaining even though the presence of an agent had no effect in comprehension. It must be noted that the sample size they used was only 30 participants. Even though the persona effect focusses more on the effect of agents on learning, the effect of ECAs on entertainment and ease of use is the same as in the empirical work presented in this thesis. Another pair of researchers (Koda and Maes, 1996) supported that the presence of an ECA in a game application may result in increased entertainment, an assumption that can also be confirmed from the empirical data presented in this thesis.

Usability Attribute: Felt part

"I felt part of Money world." Especially in game design, immersion is a rather significant element. Sweetser and Wyeth (2005) list immersion as an element of game flow which is the experience during the act of gaming. The empirical evidence shows that the HECA version scored significantly higher than the Text version in terms of immersion. Also, the qualitative data confirmed that participants felt that the ECA version was more immersive and interacted like in a real transaction. This can be justified by the anthropomorphisation of the

system and the “illusion of humanness” which mimicked a real-life interaction.

Usability Attribute: Knew what to do

“When I was using Money world, I always knew what I was expected to do.” Users reported feeling like they had a better understanding on what they were expected to do while using the ECA version of the game compared to the Text version. This can be partially explained by the theory of affordances (perception drives action). While playing the Text version of the game, most participants tried to tap on the items in the background rather than speaking. In the ECA version, due to the visual and auditory cues, they figured out that they had to respond verbally. Since speech interaction is an integral part of this study, the results support that visual and auditory cues evoke a verbal response.

The qualitative data showed that most comments about the ECA version mentioned that characters were more fun (20); they preferred interacting with a human/character (16); it was more human-like and natural (14); it was more interactive (27 participants); it was easier (17); and the text version was boring and added cognitive load (15).

Shneiderman is one of the biggest critics of ECAs. He argues that humanising the system may induce false mental models (Shneiderman and Maes, 1997). An example is that anthropomorphic agents may lead the user to believe that the system is also human-like in terms of cognitive aspects. That can make the user have expectations from the system that it does not possess and may result in a negative experience (Doumanis, 2013) Even though in the case of this research participants had the “illusion” that ECAs had human-like

cognitive aspects, especially in the case of the shopkeeper, that resulted in a positive experience instead of a negative one.

"Humans depend to a great extent on embodied behaviours to make sense and engage in face-to-face conversations. The same happens with machines: embodied agents help to leverage naturalness and users judge the system's understanding to be worse when it does not have a body (Cassell 2001)."

Research question 2:

To what extent do users perceive a difference in agent persona between ECA and neutral text presentation as measured by the agent persona instrument (API)?

This application had two agents, the shopkeeper/collaborator with whom the participant had to interact actively; and Alex/instructor who introduced the way the coins should be used and gave instructions for the items. One of the research questions was to what extent do users perceive a difference in agent persona between ECA and Text agent presentation as measured by the agent persona instrument (API). In order to answer this research question, the API questionnaire was analysed for each one of the two agents.

For the collaborator agent, the quantitative analysis revealed that the overall mean scores of the API questionnaire did differ between the two versions. The HECA agent received the highest overall mean score of 3.67 (out of 5) which translates to between neutral and agree and that participants reacted positively to the agent. The Text agent received a score of 2.81 and therefore

below average which translates to between disagree and neutral about their reaction towards the agent. The difference between ECA mean and Text mean scores of the API questionnaire was also statistically significant ($t=13.068$; $df=89$; $p.=0.000$). Further, Cohen's effect size value ($d = 1.34$) suggested a high practical significance which means that the inclusion of an HECA in the role of the collaborator has a meaningful real-life impact on the API and how participants perceive the agent.

For the instructor/Alex agent: the quantitative data also revealed a statistically significant difference between the two mean scores of the API questionnaire ($t=8.428$; $df=89$; $p.=0.000$). The HECA agent received the highest overall mean score of 3.54 which translates to between neutral and agree, thus participants reacted positively to the agent. The Text agent received a score of 2.91 which translates to between disagree and neutral on their reaction towards the agent. Furthermore, Cohen's effect size value ($d = 1.34$) suggested a high practical significance which means that the inclusion of an HECA in the role of the instructor has a meaningful real-life impact on the API and how participants perceive the agent.

The qualitative data support the quantitative findings. As it was indicated by the percentages of participants when asked to choose which agents' format they preferred along with their comments for the agents, opinions differed between the two agents. While 84.5% of participants preferred the HECA version of shopkeeper, the corresponding percentage for ECA Alex was 74.5%. At the same time, only 14.5% preferred the text version of the shopkeeper compared to a 20% of participants that chose the text version of Alex and 1% had no preference for the shopkeeper compared to 5.5% for Alex. Some of the comments indicate that participants were prone to making

comparisons between the two agents even though they recognized that the agents had different roles.

According to participants, Alex's facial expressions were not as responsive as the shopkeeper's resulting in a larger effect of the uncanny valley theory.

Also, some identified that this agent gave instructions and were not bothered having text; this is because they did not interact with this agent the same way they did with the shopkeeper thus having less expectations which was further supported by participants' comments. In their comments, participants referred to Alex as the instructor or teacher. Also, it was observed that when participants experienced the text version first, they preferred the text version of Alex. This was not the case for the shopkeeper agent. The facial animation along with the designated role of the agent as the instructor—with whom they did not interact directly—justifies the larger percentage of participants preferring the text version even though the majority preferred the HECA version. A couple of examples would be: "It was good for instructions, but I did not care much for it" and "It was less interacting, and it was more giving instructions. It was educational."

It is rather interesting that participants identified the role of the agent as the one giving instructions (mostly in the HECA version) and implied that they had lower expectations or paid less attention because they did not interact directly with Alex.

For the shopkeeper agent, participants recognised the more interactive role he had. Some users commented on the agent's facial expression although not as much as they did for Alex. The presence of the agent in the shop was welcomed as a few participants commented that having an agent in the shop is natural and expected. Even though the mismatch of some face animations

did evoke the uncanny valley effect, the effect was kept to a minimum and it did not affect the overall usability. Also, Masahiro Mori, the person who coined the term “uncanny valley”, amended his theory to acknowledge that interactive game characters are less likely to evoke this feeling of uneasiness while virtual actors are more likely to. This is a possible explanation why the agent with the most interactive role was preferable by the participants.

Research question 3:

Which factors relating to the HECA's persona attributes account for variability in usability, and to what extent?

The regression analysis attempts to model the relationship between participants' assessment of HECA Usability and ECA attributes for both agents. Through the regression analysis it was attempted to answer how much of the variability in the ECA version usability can be accounted for by the ECA attributes (based on the API).

The results from these models gave an indication on which variables (API attributes) are important and their relationship to the dependent variable (Usability mean score).

In both cases of the collaborator/shopkeeper and instructor/Alexa, two predictors were found as statistically significant. The first was “The agent was human-like”. This is supported by the Illusion of humanness and the Persona effect theory as the fact that the agent was human-like contributed to usability which can be explained possibly by the perception of a more natural

and intuitive interaction. This can also be an indication that Persona effect is not an effect of mere presence such as the social facilitation effect.

The second statistically significant predictor for both agents was "The agent was entertaining". The connection between usability and entertainment is quite interesting as it is an indication that people tend to perceive the agent as entertaining which in turn leads to increased usability.

It is interesting that in both cases, that of the collaborator which was the role of the shopkeeper and that of Alex the instructor in this scenario, the same two attributes out of nine were deemed significant for contributing to usability. The first attribute was "The agent was human-like" which is especially important since the underlying theme of the experiment was the illusion of humanness. The variable belongs to the "Human-like" factor which to quote Baylor "address the agent's behaviour and emotional expression in terms of its naturalness and personality." (Baylor & Ryu, 2003). The other factor belonged to the "Engaging" factor, also according to Baylor and Ryu "pertains to the motivational and entertaining features of the agent".

In the case of the shopkeeper the "The agent was friendly.", "The agent showed emotion", "The agent emotion was natural", "The agent was enthusiastic" and "The agent was expressive" variables, even though not significant, had a negative relationship with the DV which can be justified by the uncanny valley theory since the agents' animation and lip-synching weren't flawless thus producing an uncanny feeling, also some comments referred to the shopkeeper as 'overly friendly' and 'creepy'.

In the case of Alex, the attributes with a negative non-significant correlation were "The agent has a personality", "The agent's emotion was natural", "The agent was expressive" and "The agent was friendly". This again can be

explained partly by the uncanny valley theory since the facial expressions are connected to emotion expression. Even though a software for facial motion capture was used, in order to animate seamless facial expressions, thousands of pounds of equipment needs to be used.

Additionally, an interesting finding is that ECAs with different roles in this application do not seem to affect usability differently. Even though the two agents cannot be compared to each other since they have different roles, it is clear from the results that for both agents the same two predictors were found as statistically significant. Regardless the fact that their role in the interaction was different, for both agents the attributes that contributed more to usability were that they were perceived as human like and as entertaining. This conclusion comes from the model that includes the emotive interaction attributes that is more relevant to the present research.

As mentioned earlier in Chapter 5, the model does not include the whole API instrument rather the 9 attributes from the Emotive Interaction latent variable. There are 15 more attributes in the Information Usefulness variable that were not included due to not being relevant to the focus of the experiment, that can be a possible answer on the question on how the rest of the variance can be explained. Thus, Emotive Interaction predicts 23% of the variance in Usability when it comes to the shopkeeper and 29% when it comes to Alexa but since the agents had different roles these percentages cannot be combined. In the question, if these percentages are substantial, the answer is that it is quite relative to the field and subject of the experiment but given there are no previous studies using the exact same tools and the plethora of elements within the application the estimation is that both are rather substantial.

As an extension of the results obtained from the regression analysis, a further investigation was held focusing more on the "human-like" and "entertaining" comments made during the exit questionnaire about the ECA agents. Overall, 55 comments were made for either agents where they were described as human-like or human and 61 comments where they were described as fun and/or entertaining.

The majority of the comments on the Shopkeeper that were positive had to do with the fact that the agent was humanlike (23), made the interaction feel real or referred to as a "real person" (13), he made the interaction fun or he was funny (26) and he was friendly (16). Similar comments were made about the instructor agent where she was described as friendly (18), human-like or like a real person (12) and fun or enjoyable (14). These comments attribute human characteristics or a human dimension to the agent. In their comments, participants:

- Use of pronouns to refer to the agent when the agent was presented in the ECA form.
- During the interaction with the shopkeeper, participants applied social rules and followed similar social cues as in human to human interaction as they waited for the agent to conclude the question before answering.
- Because of the agent's presence, when the system did not pick up their voice they sympathized with the agent as if he couldn't hear them correctly rather than thinking it was their fault. "I relaxed when the SK said that he did not hear me because it made me feel it was not my fault." And "He was entertaining. The comments made it like it was his fault. He was funny and human like."

6.5 Limitations

Even though the usability questionnaire was developed to measure the participant's subjective impression of efficiency and effectiveness (system performance), performance is usually assessed based on scores (effectiveness) and time (efficiency). During the evaluation however, technical limitations did not allow the recording of such data. This relies heavily on the fact that the author had to work with a legacy code that was not originally developed for mobile devices and access to log files was not possible. However, during the evaluation, the researcher documented the participants' errors, problems or observations that directly reflected issues regarding the effectiveness and efficiency of the application. The observations were also based on the stars and scores achieved by participants that were affected by the efficiency of payment (payment made with the fewest number of coins), and efficiency of task (whether any additional help was required for each item on the shopping list) as well as the time it took for the participant to make the payment (effectiveness). Even though in most of the HECA versions participants scored higher and needed less assistance compared to the Text versions -therefore less time to complete- numerical data would have been an additional indication of higher effectiveness and efficiency achieved during the HECA version. However, the added value of using measures for accuracy and speed remains unclear as sometimes separate analysis of these performance variables can lead to contradictory results (Vandierendonck, 2017) thus the subjective impressions recorded via the questionnaire can aid the analysis of such factors.

Even though Moneyworld is a SG, it was not developed as an educational software. The primary purpose of this evaluation was the usability of the

application and how it is affected by the inclusion of HECAs and not learning effectiveness therefore it was not measured. This was a conscious decision as learning is a complex construct making it difficult to measure (Bellotti et al., 2013) while determining whether a SG is successful at achieving the anticipated learning goals is a time consuming, complex, difficult and expensive process (Hays, 2005; Enfield et al., 2012). Chin et al. (2009) attribute part of this difficulty on the fact that video games are inherently open-ended which makes it difficult to collect data. Moreover, as Bente and Breuer (2009) point out, the researcher cannot be sure that the learner is learning what they should, and the researcher cannot be very confident that he is measuring the correct thing. Thus, measuring learning in one session can be problematic since the researcher cannot be sure that the results are a learning or a memory effect. Usually games are designed to be played more than once with SGs not being an exception. Therefore, measuring for learning would require repeated evaluations over a long period of time in order to investigate the long-term effects of the game. Also, games are voluntary (Bartle, 2004), having to play a game because you were told to by someone, in this case the researcher, takes away some of its appeal. When played repeatedly, even the best games can be deprived of their fun and engagement. Game literacy can also vary from person to person and rely on the exposure that each person has to games and technology. Games are just another medium, a medium that some enjoy while others do not just like reading a book.

Another limitation is that when testing an application with speech recognition software in a non-controlled environment, the ambient noise can affect the experience. Speech recognition software is not yet evolved to a level that can

block ambient noise thus affecting the evaluation to a small degree which did not affect the interaction.

Even though the advertisement for participants in the main experiment stated clearly that only people with proficient knowledge of English should participate, a few had difficulties in understanding the language in either verbal or textual form. As a result, a small number of participants had to be turned away. Relevant to international participants, a few had a strong accent and the speech recognition system could not easily pick up their voice because it was developed using an English vocal dictionary in Pocket Sphinx. A way to tackle this issue for future experiments would be instead of self-evaluation of English proficiency, prospective participants should complete a test.

Another limitation would be the diversity of the population. The participants that were recruited for this research were mainly highly educated, with technical knowledge and under 40 years old due to the context of the game.

A few of the comments focused on the ECAs' facial animation. Animating a character by hand is a time consuming and tedious task that not always guarantees a good outcome. For that purpose, there is software that focuses on creating realistic facial and body animation. The main obstacle in the presented research is the financial limitations that did not allow using top tier facial and body animation software which usually costs a few thousand pounds also in equipment and training. That resulted in using software within our budget which created decent animations but there is surely more room for improvement in this area.

6.6 Future work and suggestions

Identifying which aspects of ECA's level of anthropomorphism have the biggest effect. Future work could include further evaluations in order to identify which aspect of the anthropomorphic interface of ECAs evokes most the illusion of humanness and contributes more to usability. In order to examine that, further evaluations need to be carried out to specify which anthropomorphic elements are the ones evoking an illusion of humanness and affecting usability more (different levels of anthropomorphic agents).

Testing the “illusion of humanness” within other contexts. The medium on which Moneyworld was tested was a SG, but the “illusion of humanness” is not specific to a certain topic or medium and could be used in other contexts. The illusion of humanness becomes less of an academic issue but more of a real-life issue due to the increasing use of virtual agents and smart screens (virtual agents with a screen) such as Amazon Echo and Google Home in our homes. One possible topic for future evaluation would be testing the addition of ECAs in home smart screens like Amazon Echo show. Would it be worth adding and for which purposes? A Greek company called MLS already has an ECA version incorporated in their smart screen called MAIC but no data on its usability are available.

Therefore, spoken HECAs as used in this research should be tested on different applications and devices. Assuming the same design guidelines for ECA development are followed along with a similar methodological process, the generalisability of the effect can be investigated.

Controlling for language and cultural background. A few people with a native language other than English preferred the text over ECA since it reminded them of subtitles available with English speaking films. O' Neil and Brown (1997), suggested that ECA preferences might have a cultural basis. There is some anecdotal evidence from the US market indicating that Japanese users prefer more anthropomorphic agents while US users not so much. It might worth exploring how users of different cultural backgrounds and whose first language is not English respond to an anthropomorphic interface such as an HECA.

Diverse population. As mentioned previously, participants who took part in this research were in their majority highly educated, with technological literacy and between 18 and 40 years old. It might worth exploring the illusion of humanness effect on older users or children and people of varying educational backgrounds as their response to the system might differ.

6.7 Implications for developers

The development of ECAs is a time-consuming process that developers might not be willing to invest in without evidence showing that it is worth the effort. In application development, assuring usability is a rather important part for the success of the interaction. In this thesis, HECAs were found to increase usability in an MSG.

The humanness and entertainment aspects of the agent persona instrument are the most useful in predicting usability scores, and these results are consistent for the two agents that were examined. In the paradigm reported in this thesis, increased usability is the result of the "illusion of humanness" effect which in turn results from high human likeness. High

human likeness is achieved by making the appropriate design choices from the ECADM.

Due to the methodological approach followed and the attention to the effect sizes, the number of evaluation participants was large enough to allow for a safe generalisation to the population. However, the generalisability of the evaluation findings to the general adult population should be treated with care. When developing usable spoken multimodal systems, the appropriateness of speech interaction must be decided for each application anew based on the purpose and environment of the application (Dybkjær et al. 2004). Weiss (2015) makes a similar claim that whether usability and quality are to be enhanced by using an ECA in a multimodal human-machine interface must be decided for each application anew. Since the platform for this evaluation was an MSG, no generalisation can be made about the “illusion of humanness” in other applications with different purposes or contexts. Nevertheless, the generalisation that can be made safely based on the evaluation findings is that contextually relevant spoken HECAAs of high human likeness with collaborative and instructional roles can induce illusion of humanness which results in increased usability in MSGs. A suggestion to developers for improving usability in similar contexts would be to incorporate spoken HECAAs with high human likeness by following the design decisions in Figure 36. Those decisions are not arbitrary as there is evidence from the literature on what results in high human likeness. As discussed in Chapter 5, Isbister and Doyle (2002) claim that an agent with physical appearance, sound and animation can cause a powerful visceral reaction on the user and evoke the “illusion of life”. By enhancing realism in movement, creating natural sounding speech and creating the right visual style that fits the application, user’s reaction to the agent can be amplified. Applying however

the same ECA design principles by following the ECADM under different circumstances (different media, different game genres, more diverse population etc.) would help determine the extent of the generalisability of the effect.

The ECADM and the spectrum of application interface design in relation to human likeness can be used to inform design decisions on the development of ECAs and the level of human likeness desired respectively. The ECAD model serves a dual function; apart from informing design decisions for designers it can act as a guide to categorise ECA research which will allow for better comparisons and analysis; in ECA research the characteristics of ECAs are not always reported or when they do they lack information that can be used for replication, analysis and comparison.

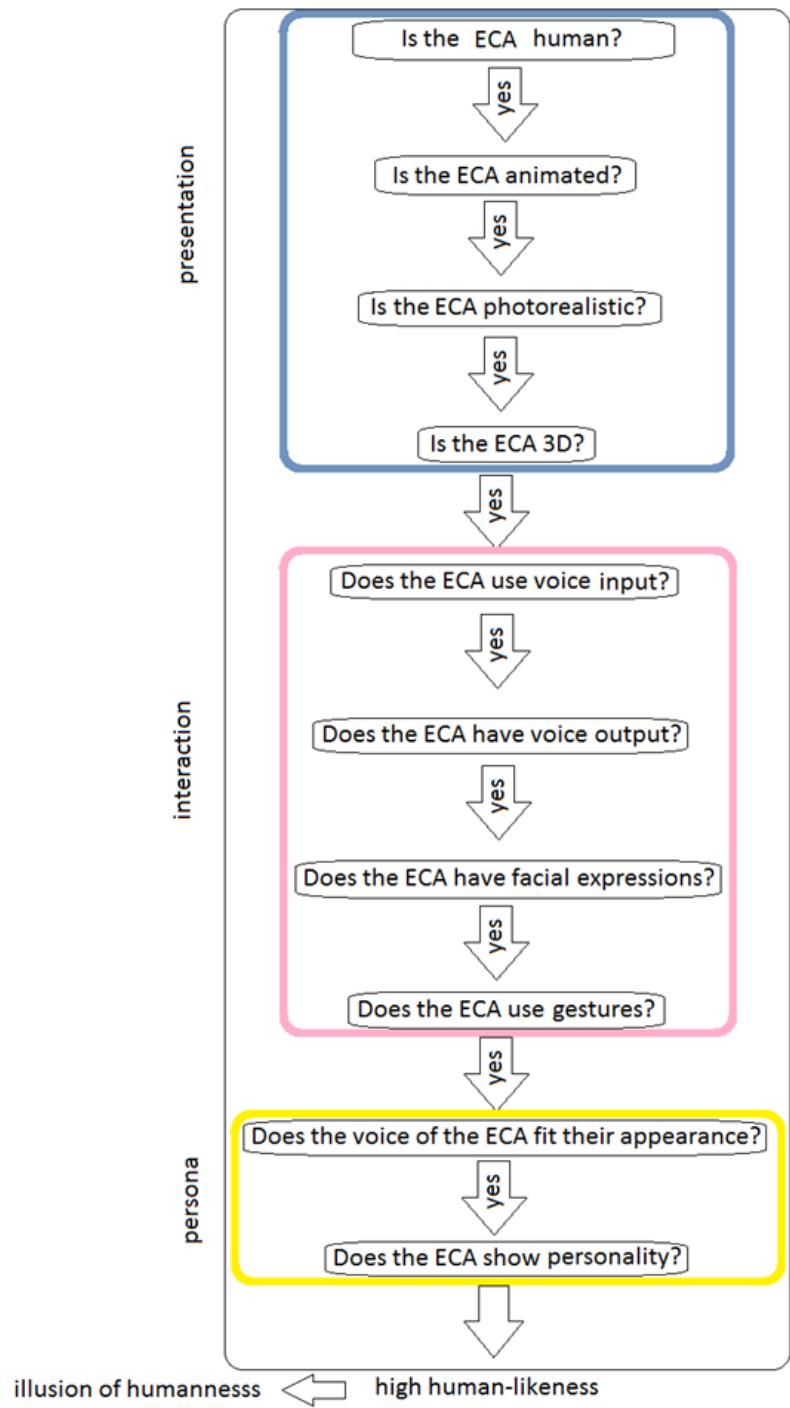


Figure 36-ECA design decisions that result in high human-likeness and in turn illusion of humanness

6.8 Conclusions

The primary aim of this research was to examine the extent to which spoken humanoid embodied conversational agents (HECAs) affect the usability of mobile serious game applications.

Mixed method analysis allowed for triangulation of findings. Following specific design decisions based on the ECADM model resulted in ECAs with high human likeness. High human likeness in turn resulted in the illusion of humanness effect. The findings revealed that ECAs with high human likeness evoked the illusion of humanness effect and improved the usability of an MSG.

The two ECA persona attributes that contributed more to usability were that the agents were regarded as human-like and entertaining. Embodied conversational agents with different roles had similar effects on usability. When the agent had the role of the instructor and the user did not interact with it, participants' expectations were lower, and a few preferred the text as it was deemed sufficient for instructions. Most of the users preferred to interact with the HECA collaborator agent due to the interactive nature of the task; the speech modality seemed to be a more fitting choice when having a conversation.

Results are consistent throughout analyses. The ECA version scored statistically significantly higher than the text version with a large effect size that shows that the results translate to a meaningful real-life difference. The regression analysis showed that the attributes "entertaining" and "human-like" contributed more to usability for both agents which supports the theory that the illusion of humanness has an impact on usability. All quantitative

results are supported and further explained by the qualitative data where users used pronouns when referring to the ECAs and justified saying that they were human-like and the interaction was more natural and fun because of them.

On one hand, HECA within an MSG do not add to the cognitive load since information is conveyed by verbal, non-verbal and extra-linguistic information. On the other hand, processing text increased the working memory load (Sweller, 1999) and participants had a harder time concentrating and remembering the information given.

Even though the experiment was experienced on a mobile phone, there were no comments about the size of the screen or the ECAs being small or an obstacle to the interaction. Also, results support the use of speech recognition as a mode of interaction with mobile applications and more specifically MSGs. Although some people whose native language was not English it was observed that they were among those who preferred the text version due to resemblance to subtitles while some suggested having both the ECA and text present.

In conclusion, ECAs on mobile devices have potential advantages over current interaction paradigms in improving usability because they provide a more "human-like" way of communicating with a complex system. However, further empirical investigation was required because the evidence on impact of ECAs on usability is lacking. The results from this thesis show that users prefer ECA versions of the Money World MSG over text, and rate it as more usable with a large effect size which shows a high practical significance (Cohen's $d=1.01$). The reason for this preference appears to be the agents' human-like attributes and the fact that they made the interaction more

entertaining. The implications of these findings are that developers should decide for each application anew if ECAs are fitting to the context and purpose of the application. However, developers should consider that in this context ECAs with high human likeness result in the illusion of humanness which in turn improves the overall usability.

References

- Abt, C.C., 1970. Serious game.
- Aldred, Jessica. (2011). From Synthespian to Avatar: Re-framing the Digital Human in Final Fantasy and The Polar Express. Mediascape.
- Altman, D., 1991. Practical Statistics for Medical Research.. London: Chapman and Hall (monograph).
- Alvarez, J., Michaud, L., 2008. Serious Games : Advergaming, edugaming, training and more, Idate. <https://doi.org/10.1145/1361083.1361093>
- Alvarez, J., Rampnoux, O., Jessel, J. P., Methel, G., 2007. Serious Game: Just a question of posture, in: Artificial & Ambient Intelligence, AISB 7: 420–423.
- Anderson, J., Davidson, N., Morton, H. & Jack, M., 2008. Language learning with inter-active virtual agent scenarios and speech recognition: lessons learned. Journal of Computer Animation and Virtual Worlds, Volume 19, pp. 605-619.
- Anderson, N.Davidson, H.Morton & M.A.Jack, 2008. Language learning with interactive virtual agent scenarios and speech recognition: lessons learned". Journal of Computer Animation and Virtual Worlds,, Volume 19, pp. 605-619.,
- Andre, E. et al., 2000. The automated design of believable dialogues for animated presentation teams. In: J. Cassell, ed. Embodied Conversational Agents. Cambridge, Massachusetts.: MIT Press, p. 220–255.
- Ardley, G., 1967. The Role of Play in the Philosophy of Plato. Philosophy 42, 226–244. <https://doi.org/10.1017/S0031819100001303>
- Argyle M. 1980. Bodily Communication. Methuen & Co Ltd, London.
- Arnab S., Berta R., Earp J., de Freitas S., Popescu M., Romero M., Stanescu I. and Usart M. "Framing the Adoption of Serious Games in Formal Education" Electronic Journal of e-Learning Volume 10 Issue 2, 2012, (pp159-171), available online at www.ejel.com
- Ball, G. et al., 1997. Lifelike computer characters: the persona project at Microsoft Research. In: J. Bradshaw, ed. Software Agents. Cambridge,MA: MIT Press.

Balme, L., Demeure, A., Barralon, N., Coutaz, J., Calvary, G., 2004. CAMELEON-RT: A Software Architecture Reference Model for Distributed, Migratable, and Plastic User Interfaces. *Ambient Intell.* 3295, 291–302. https://doi.org/10.1007/978-3-540-30473-9_28

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: Evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry*, 38(7), 813-822.

Bates, J., 1994. The role of emotion in believable agents. *Communications of the ACM*, 37(7), pp. 122-125.

Bartle, R. A. (2004). Designing virtual worlds. Berkeley, CA: New Riders

Baylor, A. & Chang, S., 2002. Pedagogical agents as scaffolds: The role of feedback timing, number of agents, and adaptive feedback. s.l.:s.n.

Baylor, A. & Ebbers, S., 2003. The Pedagogical Agent Split-Persona Effect: When Two Agents are Better than One. *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2003* (pp. 459-462). Chesapeake, VA, AACE. .

Baylor, A. & Kim, S., 2008. The effects of agent nonverbal communication on procedural and. In: Intelligent virtual agents. s.l.:Springer Berlin, pp. 208-214.

Baylor, A. & Kim, Y., 2005. Simulating Instructional Roles through Pedagogical Agents. *International Journal of Artificial Intelligence in Education*, Volume 15, pp. 95-115.

Baylor, A. & Ryu, J., 2003. The API (Agent Persona Instrument) for Assessing Pedagogical Agent Persona. s.l., ED-MEDIA 2003--World Conference on Educational Multimedia, Hypermedia & Telecommunications.

Baylor, A. L., 2000. Beyond butlers: Intelligent agents as mentors.. *Journal of Educational Computing Research*, 22(4), pp. 373-382.

Baylor, A., 2003. The impact of three pedagogical agent roles. Melbourne, Australia, AAMAS '03 Proceedings of the second international joint conference on Autonomous agents and multiagent systems.

Bellotti, F., Kapralos, B., Lee, K., Moreno-Ger, P. and Berta, R. (2013). Assessment in and of Serious Games: An Overview. *Advances in Human-Computer Interaction*, 2013, pp.1-11.

Bente, G. and Breuer, J. 2009 "Making the implicit explicit: embedded measurement in serious games," in Serious Games: Mechanisms and Effects, U. Ritterfield, M. J. Cody, and P. Vorderer, Eds., pp. 322–343, Routledge, New York, NY, USA

Bergeron, B., 2006a. Developing Serious Games (Game Development Series), Journal of Magnetic Resonance Imaging.

Bernsen, N. , 1994 Foundations of multimodal representations: a taxonomy of representational modalities, *Interacting with Computers*, Volume 6, Issue 4, December 1994, Pages 347–371, [https://doi.org/10.1016/0953-5438\(94\)90008-6](https://doi.org/10.1016/0953-5438(94)90008-6)

Beskow, J. & McGlashan, S., 1997. Olga - a conversational agent with gestures. Morgan KAufmann Publishers, San Francisco, IJCAI'97 workshop on Animated Interface Agents - Making them Intelligent, Nagoya, Japan.

Betz, N. E., & Luzzo, D. A. (1996). Career assessment and the Career Decision-Making Self-Efficacy Scale. *Journal of Career Assessment*, 4(4), 413-428.
<http://dx.doi.org/10.1177/106907279600400405>

Beun, R., de Vos, E. & Witteman, C., 2003. Embodied conversational agents: effects on memory performance and anthropomorphisation. In: T. R. e. al., ed. IVA 2003, LNAI 2792. Verlag Berlin Heidelberg 2003: Springer-, pp. 315-319.

Bickmore, T., Schulman, D. & Pfeifer, L., 2013. "Tinker – A Relational Agent Museum Guide". *Journal of Autonomous Agents and Multi-Agent Systems*, 27(2), pp. 254-276.

Bobrow, D., Kaplan, R., Kay, M., Norman, D., Thompson, H. and Winograd, T. (1977). GUS, a frame-driven dialog system. *Artificial Intelligence*, 8(2), pp.155-173.

Bogost, I., 2007. Persuasive Games: The Expressive Power of Videogames, *Literary Linguistic Computing*. <https://doi.org/10.1093/linc/fqn029>

Bohus, D., & Horvitz, E. (2010, November). Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction* (p. 5). ACM.

Bolt, R. A., 1980. "Put-that-there": Voice and gesture at the graphics interface. Seattle, Washington, USA, SIGGRAPH '80 Proceedings of the 7th annual conference on Computer graphics and interactive techniques.

Botturi, L., Loh, C.S., 2008. Once Upon a Game: Rediscovering the Roots of Games in Education. Games Purp. Potential Educ. pp.1-22.

Boyle, E., 2014. Psychological Aspects of Serious Games, in: Advances in Game-Based Learning : Psychology, Pedagogy, and Assessment in Serious Games. pp. 1–17. <https://doi.org/10.4018/978-1-4666-4773-2.ch001>

Boyle, E., Connolly, T.M., Hainey, T., 2011. The role of psychology in understanding the impact of computer games. Entertain. Comput. 2, 69–74. <https://doi.org/10.1016/j.entcom.2010.12.002>

Brogan, D., Metoyer, R. & Hodgins, J., 1998. Dynamically Simulated Characters in Virtual Environments. IEEE Computer Graphics & Applications, 18(5).

Cafaro, A., Hogni Vilhjalmsson, H. & Bickmore, T., 2006. First Impressions in Human-Agent Virtual Encounters. ACM Transactions on Computer-Human Interaction, 23(4).

Calderon A. & Ruiz M., A systematic literature review on serious games evaluation: An application to software project management, Computers & Education (2015), doi: 10.1016/j.compedu.2015.07.011

Capod, 2017. ANALYSING LIKERT SCALE/TYPE DATA St. Andrews University. [Online] Available at: <https://www.st-andrews.ac.uk/media/capod/students/mathssupport/Likert.pdf>

Cassell, J. & Stone, M., 1999. Living hand to mouth: Psychological theories about speech and gesture in interactive dialogue systems. s.l., Proceedings of the AAAI Fall symposium '99, pp. 34-42..

Cassell, J. et al., 1999. Requirements for an architecture for embodied conversational characters.. Vienna, In Computer Animation and Simulation'99 (pp. 109-120). Springer.

Cassell, J. et al., 2001. More than just a pretty face:conversational protocols and the affordances of embodiment.. Knowledge-Based Systems,, 14(1-2), pp. 55-64.

Cassell, J. et al., 2002. MACK: Media lab Autonomous Conversational Kiosk. Monte Carlo, In Proceedings of IMAGINA'02,Jan 12-15.

Cassell, J., Bickmore, T., Vilhjálmsdóttir, H., & Yan, H. (2001). More than just a pretty face: conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, 55-64.

Cassell, J., Sullivan, J., Prevost & Churchill, E. F., 2000. *Embodied Conversational Agents*. s.l.:The MIT press.

Cavazza, M. et al., 2010. "How was your day?" An affective companion ECA. s.l.:s.n.

Chan, T. W. & Baskin, A. B., 1990. Learning companion systems. In: C. F. & G., ed. *Intelligent tutoring systems at the crossroads of artificial intelligence and education*. s.l.:NJ: Ablex Publishing Corporation, pp. 7-33.

Chan, T.-W., Roschelle, J., Hsi, S., Kinshuk, Sharples, M., Brown, T., Patton, C., Cher-Niavsky, J., Pea, R., Norris, C., Soloway, E., Balacheff, N., Scardamalia, M., Dillen-Bourg, P., Looi, C.-K., Milrad, M., Hoppe, U., 2006. One-To-One Technology-Enhanced Learning: An Opportunity For Global Research Collaboration. *Res. Pract. Technol. Enhanc. Learn.* 1, 3–29.
<https://doi.org/10.1142/S1793206806000032>

Charlene Jennett, 2012 Can 'serious games' be an effective tool for workplace learning? <http://www.ucl.ac.uk/news/news-articles/1208/22082012-TARGET-serious-game>

Charsky, D., 2010. From Edutainment to Serious Games: A Change in the Use of Game Characteristics. *Games Cult.* 5, 177–198.
<https://doi.org/10.1177/1555412009354727>

Chiang, Y.-T., J LIN Professor, S.S., Cheng, C.-Y., Zhi-Feng LIU Associate Professor, E., 2011. EX-PLORING ONLINE GAME PLAYERS' FLOW EXPERIENCES AND POSITIVE AFFECT. *Turkish Online J. Educ. Technol. Copyr.* □ *Turkish Online J. Educ. Technol.* 10.

Chin,J., Dukes,R., and Gamson, W. 2009 "Assessment in simulation and gaming: a review of the last 40 years," *Simulation & Gaming*, vol. 40, no. 4, pp. 553–568.

Clarebout, G. & Heidig (née Domagk), S., 2012. Pedagogical Agents. In: S. N. (eds), ed. *Encyclopedia of the Sciences of Learning*. Boston, MA: Springer, pp. 146-211.

Clark, D.B., Tanner-Smith, E.E., Killingsworth, S.S., 2016. Digital Games, Design, and Learning: A Systematic Review and Meta-Analysis. *Rev. Educ. Res.* 86, 79–122. <https://doi.org/10.3102/0034654315582065>

Clark, R. E. & Choi, S., 2005. Five design principles for experiments on the effects of animated pedagogical agents.. *Journal of Educational Computing Research*,, 32(3), p. 209–225.

Coakes, S., 2005. SPSS: analysis without anguish: version 12.0 for Windows. Singapore: John Wiley & Sons Australia Ltd..

Coe, R., 2002. It is the effect size stupid. What effect size is and why it is important. Exeter, Annual conference of the British Educational Research Association.

Cohen, J. E., 1988. Statistical Power Analysis for the Behavioral Sciences,. Hillsdale, NJ:: Lawrence Erlbaum Associates, Inc.

Cohen, J., 1992. Quantitative methods in psychology. *Psychological bulletin*, 112(1), pp. 155-159.

Cohen, P. & Oviatt, S., 1995. The role of voice input for human-machine communication. *Proc. Natl. Acad. Sci. USA*, Vol. 92, pp. 9921-9927, October 1995.

Collin, S., Jack, M. & Anderson, J., 2004. A comparison of the effectiveness of single and multiple 3D embodied synthetic agents in eBanking. In: In Proceedings of Third International Conference on Intelligent Agents, Web Technology and Internet Commerce (IAWTIC'2004)., s.l.:s.n., pp. 564-575.

Connolly, T. M., Stansfield, M. H., & Hainey, T., 2008. Development of a general framework for evaluating games-based learning. *Proc. 2nd Eur. Conf. games-based Learn.* 1–14.

Connolly, T.M., Boyle, E. a., MacArthur, E., Hainey, T., Boyle, J.M., 2012. A systematic literature review of empirical evidence on computer games and serious games. *Comput. Educ.* 59, 661–686.
<https://doi.org/10.1016/j.compedu.2012.03.004>

Coolican, H., 1994. Research Methods and Statistics in Psychology.. ISBN 0-340600-829. ed. London: Hodder & Stoughton: s.n.

Corti, K., 2006. Games-based Learning: a serious business application. *Inf. PixelLearning* 34(6), 1–20.

- Corti, K., Gillespie, A., 2015. A truly human interface: interacting face-to-face with someone whose words are determined by a computer program. *Front. Psychol.* 6, 1–18. <https://doi.org/10.3389/fpsyg.2015.00634>
- Courgeon, M., 2008. Multimodal Affective and Reactive Character. [Online] Available at: <http://matthieu.courgeon.free.fr/MARC/main.html> [Accessed 25 March 2018].
- Craig, S., Gholson, B., Ventura, M. & Graesser, A., 2000. Overhearing dialogues and monologues in virtual tutoring systems. *Journal of Artificial Intelligence in Education*, Volume 11, pp. 242-253.
- Creswell, J. W. & Plano Clark, V. L., 2006. Designing and Conducting Mixed Methods Research.. Thousand Oaks, CA: Sage.
- Creswell, J. W., 2011. Research Design Qualitative, Quantitative, and Mixed Methods Approaches. s.l.: SAGE Publications, Inc.
- Crowe, A. R. (2007). Learning to teach with mobile technology: A teacher educator's journey. In M. van't Hooft & K. Swan (Eds.), *Ubiquitous computing in education* (pp. 127-144). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Csete, J. W. Y. & V. D., 2004. Mobile devices in and out of the classroom. Lugano, In L. Cantoni & C. McLoughlin (Eds.), *Proceedings of ED-MEDIA 2004--World Conference on Educational Multimedia, Hypermedia & Telecommunications* (pp. 4729-4736).
- Csikszentmihalyi, M., 1975. Beyond Boredom and Anxiety: Experiencing Flow in Work and Play. *Jossey-Bass Behav. Sci. Ser.* 231. <https://doi.org/10.2307/2065805>
- D'Angour, A., 2013. Plato and Play: Taking Education Seriously in Ancient Greece. *Am. J. Play* 5, 293–307.
- Dahl, Y., Alsos, O. and Svanæs, D. (2010). Fidelity Considerations for Simulation-Based Usability Assessments of Mobile ICT for Hospitals. *International Journal of Human-Computer Interaction*, 26(5), pp.445-476.
- Dautenhahn, K., 2002. Socially intelligent agents creating relationships with computers and robots.. Issue Boston, Mass.: Kluwer Academic Publishers..
- Davidson, N. McInnes, F. R. & Jack, M. A., 2004. Usability of dialogue design strategies for automated surname capture.. *Speech Communication*, Volume 43, pp. 55-70.

- De Angelis, Antonella & Johnson, Graham & Coventry, Lynne. (2001). The Unfriendly User: Exploring Social Reactions to Chatterbots.
- De Freitas, S., 2006. Learning in Immersive worlds A review of game-based learning Prepared for the JISC e-Learning Programme. JISC eLearning Innov. 3.3, 73. <https://doi.org/10.1111/j.1467-8535.2009.01024>.
- De Lope, R.P., Medina-Medina, N., 2017. A Comprehensive Taxonomy for Serious Games. *J. Educ. Comput. Res.* 55, 629–672.
<https://doi.org/10.1177/0735633116681301>
- De Vos, E., 2002. Look at that Doggy in my Windows. PhD Thesis: Utrecht University.
- Deb, S. (2016). فاعلية التعلم عن بعد في البلدان النامية باستخدام تكنولوجيا الاتصالات الخلوية و الوسائط. Effective Distance Learning in Developing Countries Using Mobile and Multimedia Technology. 5(10), pp.205-219.
- Deci, E. & Ryan, R., 1985. *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.
- Deci, E. L., Koestelr, R., & Ryan, R. M. (1999). A metanalytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125, 627-668.
- Dehn, D. M. & Van Mulken, S., 2000. The impact of animated interface agents: a review of empirical research.. *International Journal of Human-Computer Studies*, Volume 52, pp. 1-22.
- Dempsey, J. V., Lucassen, B., Gilley, W., & Rasmussen, K. (1993). Since Malone's theory of intrinsically motivating instruction: What's the score in the gaming literature? *Journal of Educational Technology Systems*, 22(2), 173–183.
- Deniozou, 2016. Investigating the potential of mobile games as learning environments for independent adult skill development. University of Edinburgh.
- Deterding, S., Khaled, R., Nacke, L., Dixon, D., 2011. Gamification: toward a definition. Chi 2011 12–15. <https://doi.org/978-1-4503-0268-5/11/0>
- Diner, S. & Doanay, A., 2017. The effects of multiple-pedagogical agents on learners academic success, motivation, and cognitive load. *Computers & Education*, 111(C), pp. 74-100.

- Djaouti, D., Alvarez, J., Jessel, J.-P., 2011. Classifying serious games: The G/P/S model. *Handb. Res. Improv. Learn. Motiv. through Educ. games Multidiscip. approaches* 118–136. <https://doi.org/10.4018/978-1-60960-495-0.ch006>
- Dohen, M. (2009). Speech through the ear, the eye, the mouth and the hand. In *Multimodal signals: Cognitive and algorithmic issues* (pp. 24-39). Springer, Berlin, Heidelberg.
- Donovan, L. 2012. *The Use of Serious Games in the corporate sector*, Dublin: Learnovate Centre.
- Doolin, S., 2014. *Usability Engineering for Embodied Conversational Agents with Older Users*, Edinburgh: University of Edinburgh.
- Dörner, R., Göbel, S., Effelsberg, W., Wiemeyer, J., 2016b. Introduction, in: *Serious Games: Foundations, Concepts and Practice*. pp. 1–34. https://doi.org/10.1007/978-3-319-40612-1_1
- Doumanis, I., 2013. Evaluating humanoid embodied conversational agents in mobile guide applications.. Middlesex University: PhD Thesis.
- Doumanis, I., Serengul, S., 2015, Framework for Research in Gamified Mobile Guide Applications using Embodied Conversational Agents (ECAs)
- Doumanis, I., Smith, S., 2015. A Framework for Research in Gamified Mobile Guide Applications using Embodied Conversational Agents (ECAs). *Int. J. Serious Games* 2, 21–40.
- Dryer, Christopher. (1999). Getting Personal with Computers: How to Design Personalities for Agents.. *Applied Artificial Intelligence*. 13. 273-295. 10.1080/088395199117423.
- Dutton, R. T., Foster, J. C., Jack, M. A. & Stentiford, F. W. M., 1993. Identifying usability attributes of automated telephone services. s.l., s.n.
- Eisenberg, N., Guthrie, I. K., Murphy, B. C., Shepard, S. A., Cumberland, A., & Carlo, G. (1999). Consistency and development of prosocial dispositions: A longitudinal study. *Child development*, 70(6), 1360-1372.
- Elborji, Y., Khaldi, M., 2014. An IEEE LOM Application Profile to Describe Serious Games «SG-LOM». *Int. J. Comput. Appl.* 86, 1–8. <https://doi.org/10.5120/15042-3404>

Embodied conversational agents (ECAs) and relevant literature References

Encarnacao, L.M., 2009. On the Future of Serious Games in Science and Industry. Proc. Cgames 2009 Usa - 14th Int. Conf. Comput. Games Ai, Animat. Mobile, Interact. Multimedia, Educ. Seri-ous Games IEEE Comp Soc, TCSIM; IEEE Comp Soc, Louisville Ch.

Enfield, J., Myers,R.D., Lara, M., and Frick, T.W. 2012 "Innovation diffusion: assessment of strategies within the diffusion simulation game," Simulation & Gaming, vol. 43, no. 2, pp. 188–214, 2012.

Erickson, T., 1997. Designing agents as if people mattered. In: J. M. Bradshaw, ed. Software agents. Menlo Park: AAAI/MIT, p. 79–96.

Erman, L., Hayes-Roth, F., Lesser, V. & Reddy, D., 1980. The Hearsay-II speech-understanding system: integrating knowledge to resolve uncertainty.. ACM Comput Surv , 12(2), p. 213–253.

Felicia, P. (2010)."Handbook of research on improving learning and motivation through educational games; multidisciplinary approaches; vol. 2 ." Reference & Research Book News,Aug.2011. AcademicOneFile,
http://link.galegroup.com/apps/doc/A263163121/AONE?u=ed_itw&sid=AONE&xid=413883e2. Accessed 4 Feb. 2018.

Field, A., 2013. Discovering Statistics Using IBM SPSS Statistics. s.l.:SAGE Publications Ltd,;

Findlater, Leah & McGrenere, Joanna. (2008). Impact of screen size on performance, awareness, and user satisfaction with adaptive graphical user interfaces. Conference on Human Factors in Computing Systems - Proceedings. 1247-1256. 10.1145/1357054.1357249.

Findlater, Leah & McGrenere, Joanna. (2008). Impact of screen size on performance, awareness, and user satisfaction with adaptive graphical user interfaces. Conference on Human Factors in Computing Systems - Proceedings. 1247-1256. 10.1145/1357054.1357249.

Flanagan, (1995), Research in speech communication, Proceedings of the National Academy of Sciences Oct 1995, 92 (22) 9938-9945; DOI: 10.1073/pnas.92.22.9938

- Fogg, B.J., 2002. Persuasive technology. *Ubiquity* 2002, 2.
- <https://doi.org/10.1145/764008.763957>
- Fogg, B.J., 2003. Persuasive Technology: Using Computers to Change What We Think and Do, *Persuasive Technology: Using Computers to Change What We Think and Do*. <https://doi.org/10.1016/B978-1-55860-643-2.X5000-8>
- Foo, N., Douglas, G. & Jack, M., 2008. 'Incentive schemes in the financial services sector: moderating effects of relationship norms on customer-brand relationship. *International Journal of Bank Marketing*, 26(2), pp. 99-118.
- Foster, J. C. et al., 1998. An experimental evaluation of preferences for data entry method in automated telephone services.. *Behaviour and Information Technology*, Volume 17, pp. 82-92.
- Foster, M. E. (2007, July). Enhancing human-computer interaction with embodied conversational agents. In *International Conference on Universal Access in Human-Computer Interaction* (pp. 828-837). Springer, Berlin, Heidelberg.
- Frasson C., Blanchard E.G. (2012) Simulation-Based Learning. In: Seel N.M. (eds) *Encyclopedia of the Sciences of Learning*. Springer, Boston, MA
- Freeman, D. (2003). Creating emotions in games. The craft and art of emotioneering.
- Gaitatzes A., Christopoulos D., Papaioannou G.: *The Ancient Olympic Games: Being Part of the Experience*. In *VAST 2004: The 5th International Symposium on Virtual Reality, Archaeology and Cultural Heritage* (2004), pp. 19–28. 4, 5
- Gamelearn: Gamebased learning courses for soft skills training, 'The future of serious games through the lens of mobile devices', 2015. [Online]. Available: <https://gamelearn.com/thefutureofseriousgamesthroughthelensofmobiledevices/> .[Accessed: 04Dec2015].
- Gardner, J.E.: Can the Mario Bros. help? Nintendo games as an adjunct in psychotherapy with children. *Psychother. Theory Res. Pract. Train.* 28, 667 (1991)
- Garris, R., Ahlers, R., Driskell, J.E., 2002. Games, Motivation, and Learning: A Research and Practice Model. *Simul. Gaming* 33, 441–467.
<https://doi.org/10.1177/1046878102238607>
- Garvey, C. (1990). *The developing child series. Play (Enlarged ed.)*. Cambridge, MA, US: Harvard University Press.

Gee, J. P. (2005). Learning by design: Good video games as learning machines. *E-learning and Digital Media*, 2(1), 5-16.

Gee, J., 2005. Why are videogames good for learning? Spectrum 32, 25–32.

Gee, J.P., 2003. What video games have to teach us about learning and literacy. *Comput. Entertain.* 1, 20. <https://doi.org/10.1145/950566.950595>

George, S., Serna, A., 2011. Introducing mobility in serious games: Enhancing situated and collaborative learning, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 12–20. https://doi.org/10.1007/978-3-642-21619-0_2

Glass, G. (1972). The wisdom of scientific inquiry on education. *Journal of Research in Science Teaching*, 9(1), pp.1-18.

Goodhue, D. & Loiacono, E., 2002. Randomizing Survey Question Order Vs. Grouping Questions by Construct:An Empirical Test of the Impact On Apparent Reliabilities and Links to Related Constructs,. s.l., Proceedings of the 35th Hawaii International Conference.

Gould, A. (1995). PLANNING AND REVISING THE SAMPLE SIZE FOR A TRIAL. *Statistics in Medicine*, 14(9), pp.1039-1051.

Grace-Martin & M, 2017. Can Likert Scale Data ever be Continuous?. [Online] Available at: <http://www.theanalysisfactor.com/can-likert-scale-data-ever-be-continuous/> [Accessed 6 June 2017].

Gratch, J., Hartholt, A., Dehghani, M., & Marsella, S. (2013). Virtual humans: a new toolkit for cognitive science research. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 35, No. 35).

Gravetter, F. & Wallnau, L., 2014. Essentials of statistics for the behavioral sciences. 8th ed. Belmont, CA: Wadsworth.

Gredler, M. E. (1996). Educational Games and Simulations: A Technology in Search of a (Research) Paradigm. *Technology*, 39, 521-540.

GREEN, B. and WOLF, A. (1961). BASEBALL:AN AUTOMATIC QUESTION-ANSWERER. Ft. Belvoir: Defense Technical Information Center.

Griffiths, M.: The therapeutic use of videogames in childhood and adolescence. *Clin. Child Psychol. Psychiatry* 8, 547–554 (2003)

Griffiths, M.: Video games and clinical practice: issues, uses and treatments. Br. J. Clin. Psychol. 36, 639–642 (1997)

Gris Sepulveda, I., 2015. Physical engagement as a way to increase emotional rapport in interactions with embodied conversational agents, Texas: The University of Texas at El Paso, ProQuest Dissertations Publishing.

Gulz, (2004). Benefits of virtual characters in computer-based learning environments: Claims and evidence. International Journal of Artificial Intelligence in Education, 14(3–4), 313–334

Gulz, A. & Haake, M., 2006. Design of animated pedagogical agents—A look at their look. International Journal of Human-Computer Studies, 64(4), pp. 322-339.

Gulz, A. et al., 2011. Building a social conversational pedagogical agent: design challenges and methodological approaches . In: D. Perez-Marin & I. & Pascual-Nieto, eds. Conversational agents and natural language interaction. Hersey: IGI Global, pp. 128-155.

Gunson, N., Marshall, D., Morton, H. & Jack, M., 2011. User perceptions of security and usability of single-factor and two-factor authentication in automated telephone banking. Computers & Security, 30(4), pp. 208-220.

Haake, M. & Gultz, A., 2009. A look at the roles of look & roles in embodied pedagogical agents. International Journal of Artificial Intelligence in Education, Volume 19, pp. 39-71.

Hainey, T., Connolly, T., Stansfield, M., Boyle, L., 2011. The Use of Computer Games in Education : A Review of the Literature. Handb. Res. Improv. Learn. Motiv. through Educ. Games Multidiscip. Approaches (2 Vol. 29–50. <https://doi.org/10.4018/978-1-60960-495-0.ch002>

Hair, J. et al., 1998. Multivariate data analysis. s.l.:Pearson.

Hamilton, A., 2017. Listen up: the future of voice technology, The Drum Network. [Online] Available at: <http://www.thedrum.com/opinion/2017/11/15/listen-up-the-future-voice-technology> [Accessed 3 March 2018].

Hanson, D., Olney, A., Prilliman, S., Mathews, E., Zielke, M., Hammons, D., ... & Stephanou, H. (2005, July). Upending the uncanny valley. In AAAI (Vol. 5, pp. 1728-1729).

- Hayes, B., Bonner, A. & Douglas, C., 2013. An introduction to mixed methods research for nephrology nurses.. Renal Society of Australasia Journal, Volume 9, pp. 8-14..
- Hayes-Roth, B., 1998. Jennifer James, celebrity auto spokesperson. Orlando, Florida, USA — July 19 - 24, 1998 , SIGGRAPH '98 ACM SIGGRAPH 98 Conference abstracts and applications.
- Hayes-Roth, B., Amano, K., S. R. & S. T., 2004. Training brief intervention with a virtual coach and virtual patients.. In: W. BK & R. G., eds. Annual Review of CyberTherapy and Telemedicine. San Diego, CA: Interactive Media Institute, p. 85–96.
- Hays, R. T. 2005 "The effectiveness of instructional games: a literature review and discussion," Tech. Rep. 2005-004, Naval Air Warfare Center, Training Systems Division,.
- Hays, R.T., 2005. The effectiveness of instructional games: a literature review and discussion. Nav. Air Warf. Cent. Train. Syst. Div. 1–63. <https://doi.org/citeulike-article-id:3089090>
- Heidig, S. & Clarebout, G., 2011. Do pedagogical agents make a difference to student motivation and learning?. Educational Research Review, 6(1), pp. 27-54.
- Helpern, D. et al., 2012. Operation ARA: A computerised learning game that teaches critical thinking and scientific reasoning. Thinking skills and Creativity, Volume 7, pp. 93-100.
- Helpern, D. et al., 2012. Operation ARA: A computerised learning game that teaches critical thinking and scientific reasoning. Thinking skills and Creativity, Volume 7, pp. 93-100.
- Hershfield, H., Goldstein, D., Sharpe, W., Fox, J., Yeykelis, L., Carstensen, L. and Bailenson, J. (2011). Increasing Saving Behavior Through Age-Progressed Renderings of the Future Self. *Journal of Marketing Research*, 48(SPL), pp.S23-S37.
- Herz, J.C., 1997. Joystick nation : how videogames ate our quarters, won our hearts, and re-wired our minds. Little, Brown Company, Bost. 240.
<https://doi.org/10.1007/s13398-014-0173-7.2>

- Hill, R. et al., 2003. Virtual humans in the mission rehearsal exercise system.. Künstliche Intelligenz (KI Journal), Special issue on Embodied Conversational Agents. , 17(4), p. 5–10.
- Huang, H., 2018. Embodied Conversational Agents. In: K. L. N. J. Kirakowski, ed. The Wiley Handbook of Human Computer Interaction, 1. s.l.:John Wiley & Sons Ltd., pp. 601-614.
- Huang, W.D., Johnson, T.E., Han, S.H.C., 2013. Impact of online instructional game features on college students' perceived motivational support and cognitive investment: A structural equation modeling study. Internet High. Educ. 17, 58–68. <https://doi.org/10.1016/j.iheduc.2012.11.004>
- Hubal, R. et al., 2008. How Do Varied Populations Interact with Embodied Conversational Agents? Findings from Inner-city Adolescents and Prisoners. Comput Human Behav., 24(3), p. 1104–1138.
- Hyers, K. & Mawston, N., 2017. Strategy Analytics-Google Leads in AI Powered Smartphones. [Online] Available at: [https://www.strategyanalytics.com/strategy-analytics/news/strategy-analytics-press-releases/strategy-analytics-press-release/2017/10/25/google-leads-in-ai-poweredsmartphones?utm_source=Triggermail&utm_medium=email&utm_campaign=Post%20Blast%20%28bii-apps-\[Accessed 5 March 2018\].](https://www.strategyanalytics.com/strategy-analytics/news/strategy-analytics-press-releases/strategy-analytics-press-release/2017/10/25/google-leads-in-ai-poweredsmartphones?utm_source=Triggermail&utm_medium=email&utm_campaign=Post%20Blast%20%28bii-apps-[Accessed 5 March 2018].)
- Hylén, J., 2017. 2 arguments for using mobile phones and social media in adult learning. [Online] Available at: <https://ec.europa.eu/epale/en/blog/2-arguments-using-mobile-phones-and-social-media-adult-learning> [Accessed 26 January 2018].
- IAI, n/a. *The information architecture institute.* [Online] Available at: <https://www.iainstitute.org/what-is-ia> [Accessed 10 August 2019].
- Ilagan, J., 2014. TARDIS (Training young Adult's Regulation of emotions and Development of social Interaction Skills). [Online] Available at: <http://www.ucl.ac.uk/ioe/research/featured-research/tardis> [Accessed 25 March 2018].
- Isbister, K. & Doyle, P., 2002. Design and evaluation of embodied conversational agents: A proposed taxonomy. Bologna, Italy, AAMAS '02, July 15-19, 2002.

Iso.org. (1998). Ergonomics of human-system interaction. [online] Available at: <https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en> [Accessed 10 Sep. 2019].

Jack, M. A., Foster, J. C. & Stentiford, F. W. M., 1993. Usability analysis of intelligent dialogues for automated telephone services.. s.l., s.n.

Jack, M. et al., 2005. Research in usability engineering workshop notes, s.l.: s.n.

Jacob Cohen (1988). Statistical Power Analysis for the Behavioral Sciences (second ed.). Lawrence Erlbaum Associates.

Jaldemark, J., Sofia, , Eriksson, , 1, B., Peter, , Mozelius, , 2017. Applying mobile devices and game-based learning in formal educational settings : Playing Pokémon Go as a tool for learning in a Swedish 11–14.

Jamieson, S., 2004. Likert scales: How to (ab)use them.. Medical Education, Volume 38, p. 1217–1218..

Jekel, J., DL, K. & JG, E., 2001. Epidemiology, Biostatistics and Preventive Medicine. Philadelphia: W.B. Saunders Company.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and the-oretical perspectives. Handbook of personality: Theory and research, 2(1999), 102-138.

Johnson, G. & Valente, A., 2008. Tactical alnguage and culture training systems: using artificial inteligence to teach foreign languages and cultures.. Melno, Park, CA, Proceedings of the 20th national conference on innovative applications of artificial intelligence..

Johnson, S. C., Slaughter, V., & Carey, S. (1998). Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds. Developmental Science, 1, 233–238.

Judd, C. M., McClelland, G. H. & Ryan, C. S., 2009. Data analysis: A model comparison approach., New York: Routledge..

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2009). Data analysis: A model comparison approach (2nd ed.). New York, NY, US: Routledge/Taylor & Francis Group.

Junlan Feng, D. Hakkani-Tur, G. Di Fabrizio, M. Gilbert and M. Beutnagel, (2006)"Webtalk: Towards Automatically Building Spoken Dialog Systems

Through Miningwebsites," 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, 2006, pp. I-I.

Jurafsky, D. & Martin, J., 2017. Dialog systems and chatbots. In: Speech and Language Processing. s.l.:s.n., p. Chapter 29.

Juul, J. (2012). *A casual revolution*. Cambridge, Mass.: MIT Press.

Kafai, Yasmin. (1994). Minds In Play: Computer Game Design as a Context for Children's Learning.

Kalyuga, Chandler & Sweller, 1999. Managing split-attention and redundancy in multimedia instruction. *Applied cognitive psychology* , 13(4), pp. 351-371.

Kapp, K.M., 2007. Tools and techniques for transferring know-how from boomers to gamers. *Glob. Bus. Organ. Excell.* 26, 22–37. <https://doi.org/10.1002/joe.20162>

Kasap, Z. & Magnenat-Thalmann, N., 2008. Intelligent Virtual Humans with Autonomy and Personality : State-of-the-Art. *Intelligent decision Technologies*, 1(2), p. 43–84.

Ke, F. (2009). A Qualitative Meta-Analysis of Computer Games as Learning Tools, in Ferdig, R.E. (Ed), *Handbook of research on effective electronic gaming in education*, Information Science Reference; Hershey, PA, 1-32.

Ke, F., 2008. Computer games application within alternative classroom goal structures: Cognitive, metacognitive, and affective evaluation. *Educ. Technol. Res. Dev.* 56, 539–556. <https://doi.org/10.1007/s11423-008-9086-5>

Kelley, K., Clark, B., Brown, V. & Sitzia, J., 2003. Good practice in the conduct and reporting of survey research.. *International Journal for Quality in health care*, 15(3), pp. 261-266.

Kickmeier-Rust, M., 2009. Talking digital educational games. ... *Adapt. Digit. Educ. Games*

Kickmeier-Rust, M.D., Peirce, N., Conlan, O., Schwarz, D., Verpoorten, D., Albert, D., n.d. Immersive Digital Games: The Interfaces for Next-Generation E-Learning? Univers. Access Human-Computer Interact. Appl. Serv. 647–656. https://doi.org/10.1007/978-3-540-73283-9_71

Kiili, K., 2005. Educational Game Design : Experiential gaming model revised. Building.

- Kim, C. & Baylor, A., 2008. A Virtual Change Agent: Motivating Pre-service Teachers to Integrate Technology in Their Future Classrooms. *Educational Technology & Society*, 11(2), pp. 309-321.
- Kipp, M., Kipp, K., Ndiaye, A. & Gebhard, P., 2006. Evaluating the tangible interface and virtual characters in the interactive COHIBIT exhibit.. *Intelligent Virtual Agents*, Volume Springer Berlin Heidelberg, pp. 434-444.
- Klopfer, E., Squire, K., 2008. Environmental detectives-the development of an augmented reality platform for environmental simulations. *Educ. Technol. Res. Dev.* 56, 203–228. <https://doi.org/10.1007/s11423-007-9037-6>
- Knutson, Brian. (1996). Facial expressions of emotion influence interpersonal trait inferences. *J. Nonverbal Behav.* 20. 165-182. 10.1007/BF02281954.
- Koda, T., & Maes, P. (1996, August). Agents with faces: The effects of personification of agents. In *Proceedings of HCI* (Vol. 96, pp. 98-103).
- Kopp, S., Gesellensetter, L., Krämer, N. & Wachsmuth, I., 2005. A conversational agent as museum guide—design and evaluation of a real-world application. In: *International Workshop on Intelligent Virtual Agents*. Berlin Heidelberg: Springer, p. 329–343.
- Kopp, S., Gesellensetter, L., Krämer, N. & Wachsmuth, I., 2005. A conversational agent as museum guide—design and evaluation of a real-world application. In: *International Workshop on Intelligent Virtual Agents*. Berlin Heidelberg: Springer, p. 329–343.
- Kopp, S., Jung, B., Lebmann, N. & Wachsmuth, I., 2003. Max – A Multimodal Assistant in Virtual Reality Construction. s.l.: KI 4/03: 11-17. .
- Korre, D., 2012. The impact of realistic virtual humans within immersive environments. University of Edinburgh.
- Kramer, N., Rosenthal-von der Putter, A. & Hoffmann, L., 2015. Social effects of virtual and robot companions. In: S. Sundar, ed. *The handbook of the psychology of communication technology*. s.l.:John Wiley & Sons, Inc..
- Kroenke, D. & Auer, D., 2009. *Database Concepts*. New Jersey: Prentice Hall.
- Kukulska-Hulme, A., Pettit, J., 2006. Practitioners as innovators: Emergent practice in personal mobile teaching, learning, work and leisure. *Mob. Learn. Transform. Deliv. Educ. Train.* 135–155.

Laamarti, F., Eid, M., El Saddik, A., 2014. An overview of serious games. Int. J. Comput. Games Technol. <https://doi.org/10.1155/2014/358152>

Landauer, T., 1988. Research methods in human-computer interaction. . In: Handbook of human-computer interaction. s.l.:s.n., pp. 905-28.

Landay, J., 2016. Stanford experiment shows speech recognition writes texts more quickly than thumbs, s.l.: Stanford .

Lane, H. et al., 2011. Intelligent tutoring goes to the museum in the big city: A pedagogical agent for informal science education. International journal of artificial intelligence in education, pp. 155-162.

Larsen, L. B., 1999. Combining objective and subjective data in evaluation of spoken dialogues.. Irsee, Germany, ESCA Workshop on Interactive Dialogue in Multi-Modal Systems..

Larsen, L. B., 2003. Assessment of spoken dialogue system usability - what are we really measuring?. s.l., EUROSPEECH '93..

Lenhart, A., Kahne, J., Middaugh, E., Rankin Macgill, A., Evans, C., Vitak, J., 2008. Teens, Video Games, and Civics: Teens' gaming experiences are diverse and include significant social interaction and civic engagement. Pew Internet Am. Life Proj. 1–64. <https://doi.org/10.1016/j.chembiol.2006.01.005>

Lepper, M(1988) Motivational Considerations in the Study of Instruction, Cognition and Instruction, 5:4, 289-309, DOI: 10.1207/s1532690xci0504_3

Lester, J. et al., 1997. The persona effect: Affective impact of animated pedagogical agents. New York, NY: ACM Press. pp. 359–366, Proceedings of the Human Factors in Computing Systems Conference.

Lester, J., 1996. The persona effect: Affective impact of animated pedagogical agents.. Proceedings of the Conference of Human Factors in Computer Systems (CHI-97) , pages 359–366, Atlanta, GA, 1996., Volume Atlanta, GA, p. 359–366.

Likert, R., 1932. A technique for the measurement of attitudes.. In: Archives of Psychology. s.l.:s.n.

Linser, R., Ree-Linsdtad, N., Vold, T., 2008. The Magic Circle - Game Design Principles and Online Role-play Simulations. Proc. World Conf. Educ. Multimedia, Hypermedia Telecommun. 5290–5297.

Louwense, M., Graesser, A., McNamara, D. & Lu, S., 2008. Embodied Conversational Agents as Conversational Partners. *APPLIED COGNITIVE PSYCHOLOGY*, Issue Published online in Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/acp.1527.

Love, S. et al., 1992. Towards a usability measure for automated telephone services.. s.l., Proceedings of Institute of Acoustics Speech and Hearing Workshop..

Lubke, Gitta & Muthén, Bengt. (2004). Applying Multigroup Confirmatory Factor Models for Continuous Outcomes to Likert Scale Data Complicates Meaningful Group Comparisons. *Structural Equation Modeling-a Multidisciplinary Journal - STRUCT EQU MODELING*. 11. 514-534. 10.1207/s15328007sem1104_2.

Ma, M., Oikonomou, A., Jain, L., 2011. Innovations in Serious Games for Future Learning, in: Se-rious Games and Edutainment pp. 3–7.
<https://doi.org/10.1007/978-1-4471-2161-9>

MacDorman, K. and Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, 7(3), pp.297-337.

Maes. 1997. Pattie Maes on Software Agents: Humanizing The Global Computer. *IEEE Internet Computing* 1, 4 (July 1997), 10-19.
DOI=<http://dx.doi.org/10.1109/MIC.1997.612209>

Magnenat-Thalmann, N., Kasap, Z., 2009. Virtual humans in serious games. 2009 Int. Conf. CyberWorlds, CW '09 71–79. <https://doi.org/10.1109/CW.2009.17>

Malone, T. W. & Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. *Aptitude, learning, and instruction*, 3(1987), 223-253.

Malone, T.W., 1981. Toward a theory of intrinsically motivating instruction. *Cogn. Sci.* 5, 333–369. [https://doi.org/10.1016/S0364-0213\(81\)80017-1](https://doi.org/10.1016/S0364-0213(81)80017-1)

Manderlink, G., & Harackiewicz, J. M. (1984). Proximal versus distal goal setting and intrinsic motivation. *Journal of personality and social psychology*, 47(4), 918.

Manganello, F., Falsetti, C., Spalazzi, L., Leo, T., 2013. PKS: An ontology-based learning construct for lifelong learners. *Educ. Technol. Soc.* 16, 104–117.
<https://doi.org/10.2307/jeductechsoci.16.1.104>

Marczewski, A. (2013). Gamification: A Simple Introduction and a Bit More. E-Book.

Matthews, A., Anderson, N., Anderson, J. & Jack, M., 2008. The effects of personality and individualised product portrayals in the usability of 3D embodied conversational agents in an eBanking scenario'. In: International Conference on Intelligent Virtual Agents, (IVA-08). Tokyo, Japan. : s.n., pp. 516- 517.

Matthews, A., Anderson, N., Anderson, J. & Jack, M., 2008. The effects of personality and individualised product portrayals in the usability of 3D embodied conversational agents in an eBanking scenario'. In: International Conference on Intelligent Virtual Agents, (IVA-08). Tokyo, Japan. : s.n., pp. 516- 517.

McBreen, H., 2002. Embodied conversational agents : extending the persona metaphor to virtual retail applications. PhD thesis: University of Edinburgh .

McCollum, C. et al., 2004. Developing an immersive, cultural training system. Arlington, VA: National Training Systems Association. 2004.Dec 1–2, Proceedings of the Interservice/Industry Training, Simulation, and Education.

McLaughlin, T., Smith, D., Brown, I. a., 2010. A framework for evidence based visual style development for serious games. Proc. Fifth Int. Conf. Found. Digit. Games - FDG '10 132–138. <https://doi.org/10.1145/1822348.1822366>

Michael, D.R., Chen, S.L., 2005. Serious Games: Games That Educate, Train, and Inform, Education. <https://doi.org/10.1021/la104669k>

Min Kim, C., 2012. Virtual Change Agents. In: N. M. Seel, ed. Encyclopedia of the Sciences of Learning. s.l.:Springer US.

Moreno, R. & Flowerday, T., 2006. Students' choice of animated pedagogical agents in science learning: a test of the similarity-attraction hypothesis on gender and ethnicity. Contemporary Educational Psychology, 31(2), p. 186–207.

Moreno, R., 2005. Multimedia learning with animated pedagogical agents.. In: R. Mayer, ed. The Cambridge handbook of multimedia learning . Cambridge, UK: Cambridge University Press., p. 507–524.

Moreno, R., M. R., S. H. & L. J., 2001. he case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents?. Cognition and Instruction., Volume 19, p. 177–213.

Moreno-Ger, P., Torrente, J., Hsieh, Y. G. & T., W. L., 2012. Usability Testing for Serious Games: Making Informed Design Decisions with User Data.. *Advances in Human-Computer Interaction*, 1(13).

Morewedge, C., Preston, J. and Wegner, D. (2007). Timescale bias in the attribution of mind. *Journal of Personality and Social Psychology*, 93(1), pp.1-11.

Mori, (1970) "The uncanny valley," *Energy*, vol. 7, no. 4, pp. 33–35, (in Japanese).

Mori, J., Prendinger, H. & Ishizuka, M., 2003. Evaluation of an Embodied Conversational Agent with, s.l.: Paper presented at AAMAS-03, Second International Joint Conference on Autonomous Agents and Multi-agent systems.

Morton, H. & Jack, M., 2005. 'Scenario-based spoken interaction with virtual agents'. *Journal of Computer Assisted Language Learning*, 18(3), pp. 171-191.

Morton, H., Mcbreen, H. M. & Jack, M. A., 2004. Experimental evaluation of the use of embodied conversational agents in eCommerce applications.. In: Z. &. P. C. Ruttkay, ed. *From Brows till Trust: Evaluating Embodied Conversational Agents.* s.l.:s.n.

Morton,H & M.A.Jack, 2005. Scenario-based spoken interaction with virtual agents. *Journal of Computer Assisted Language Learning*, 18(3), pp. 171-191.

Moundridou, M. & Virvou, M., 2002. Evaluating the persona effect of an interface agent in an intelligent tutoring system.. *Journal of Computer Assisted Learning*, 18(3), p. 253–261.

Murano, ., 2006. Why anthropomorphic user interface feedback can be effective and . University of Salford, s.n.

Nakamura, J., Csikszentmihalyi, M., 2009. Flow theory and research. Oxford Handb. Posit. Psy-chol.

<https://doi.org/10.1093/oxfordhb/9780195187243.013.0018>

Nass, C. & Moon, Y., 2000. Machines and Mindlessness: Social Responses to Computers. *The Society for the Psychological Study of Social Issues*, 56(1), pp. 81-103.

Nass, C. et al., 1997. Computers are social actors:A review of current research. In: B. Friedman, ed. *Moral and ethical issues in human computer interaction* . Stanford, CA: : CSLI Press, p. 137–162.

Nass, C., Steuer, J. & Tauber, E., 1994. Computers are social actors. s.l., CHI '94 Human Factors in Computing Systems.

Nielsen Norman Group. (1994). Guerrilla HCI: Article by Jakob Nielsen. [online] Available at: <https://www.nngroup.com/articles/guerrilla-hci/> [Accessed 21 Aug. 2019].

Nielsen Norman Group. (2007). High-Cost Usability Sometimes Makes Sense. [online] Available at: <https://www.nngroup.com/articles/when-high-cost-usability-makes-sense/> [Accessed 21 Aug. 2019].

Nishida, T., Nakazawa, A., Ohmoto, Y. & Mohammad, Y., 2014. History of Conversational System Development. In: T. Nishida, ed. Conversational Informatics: A Data-Intensive Approach with Emphasis on Nonverbal Communication. s.l.:Springer Japan , pp. 43-62.

NIST, S., 2017. Histogram Interpretation: Skewed (Non-Normal) Right. (2017). Engineering statistics handbook. [Online] Available at: <http://www.itl.nist.gov/div898/handbook/eda/section3/histogr6.htm> [Accessed 6 June 2017].

Noma, T., Zhao, L. & Badler, N. I., 2002. Design of a virtual human presenter.. s.l.:s.n.

Norman, D., 1997. How might people interact with agents.. In: J. M. Bradshaw, ed. Software agents. Menlo Park, CA: MIT Press, pp. 49-56.

Olsen, J. (2014). Health Coaching: A Concept Analysis. *Nursing Forum*, 49(1), pp.18-29.

Pagani, M., 2009. Encyclopedia of Multimedia Technology and Networking,. Second Edition ed. Bocconi University, Italy: IGI Global.

Pallant, J., 2013. SPSS survival manual. s.l.: Open University Press McGraw-Hill Education.

Park, Y., 2011. A pedagogical framework for mobile learning: Categorizing educational applications of mobile technologies into four types. Int. Rev. Res. Open Distance Learn. 12, 78–102. [https://doi.org/10.3394/0380-1330\(2006\)32](https://doi.org/10.3394/0380-1330(2006)32)

Pea, R.D., Maldonado, H., 2006. WILD for learning : Interacting through new computing devices anytime , anywhere. Cambridge Handb. Learn. Sci. 852–885.

Pelachaud, C., De Carolis, B., de Rosis, F. & Poggi, I., 2002. Embodied contextual agent in information delivering application.. Proceedings of the First International Joint Conference on Autonomous Agents and Multi-agent Systems, Volume ACM Press, p. 758–765.

Perez-Martin, D. & Pascual-Nieto, I., 2011. Conversational agents and natural language interaction. 1 ed. Hersley: IGI Global.

PhD, R. (2019). Why Do We Anthropomorphize?. [online] Psychcentral.com. Available at: <https://psychcentral.com/news/2018/03/01/why-do-we-anthropomorphize/11766.html> [Accessed 10 Sep. 2019].

PIXELearning Limited. Why serious games work, 2011.
<http://www.pixelearning.com/Resources/White%20Papers/Why%20Games%20work%20for%20learning%20and%20development.pdf>.

Poggi, I. et al., 2005. Greta a believable embodied conversational agent. In: O. Stock & M. Zancanaro, eds. Multimodal Intelligent Information Presentation. Dordrecht: Springer.

Popescu, M., Bellotti, F., 2012. Approaches on Metrics and Taxonomy in Serious Games, in: Pro-ceedings of The 8th International Scientific Conference - eLearning and Software for Education. pp. 351–358.
<https://doi.org/10.5682/2066-026X-12-171>

Preece, J., Rogers, Y. & Sharp, H. .., 2002. Interaction design: beyond human-computer interaction.. NY: : Wiley..

Prendinger, H., Mayer, S., Mori, J. & Ishizuka, M., 2003. Persona Effect Revisited Using Bio-Signals to Measure and Reflect the Impact of. s.l., T. Rist et al. (Eds.): IVA 2003, LNAI 2792, pp. 283–291, 2003..

Prendinger, H., Mori, J. & (2005)., M. I., 2005. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game.. International Journal of Human-Computer Studies, Volume 62, pp. 231-245.

Prensky, M., 2003. Digital game-based learning. Comput. Entertain. 1, 21.
<https://doi.org/10.1145/950566.950596>

Provoost, S., Ming Lau, H., Ruwaard, J. & Riper.H, 2017. Embodied Conversational Agents in Clinical Psychology: A Scoping Review. Journal of Medical Internet Research, 19(5).

Rai, S., Wong, K., Cole, P., 2006. Game construction as a learning tool. ... 2006 Int. Conf. Game ... 231–236.

Raybourn, E., H. J., D. E. & K, M., 2005. Adaptive thinking & leadership simulation game training for Special Forces officers. Orlando, FL, Proceedings of the Interservice/Industry Training, Simulation and Education Conference..

Reeves, B. and Read, J. (2009). *Total Engagement*. Boston: Harvard Business Review Press.

Reeves, Byron & Nass, Clifford. (1996). The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Bibliovault OAI Repository, the University of Chicago Press.

Rehm, Matthias & Wissner, Michael. (2005). Gamble — A Multiuser Game with an Embodied Conversational Agent. 180-191. 10.1007/11558651_18.

Rickel, J. & Johnson, W., 1997. Steve: an animated pedagogical agent for procedural training in virtual environments. ACM SIGART Bulletin, Volume 8 (Issue 1-4).

Rickel, J. & Johnson, W., 2000. Task-oriented collaboration with embodied agents in virtual worlds. In: J. Cassell, ed. *Embodied conversational agents*. s.l.:MIT Press, pp. 96-122.

Rieber, L. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Educational Technology Research and Development*, 44(2), pp.43-58.

Ritterfeld, U., Cody, M., Vorderer, P., 2009. Serious games: Mechanisms and effects, *Serious Games: Mechanisms and Effects*.
<https://doi.org/10.4324/9780203891650>

Robertson, J. & Kaptein, M., 2016. Modern Statistical Methods for HCI. s.l.:Springer.

Robertson, J. & Wiemer-Hastings, R., 2002. Feedback on Children's Stories via Multiple Interface Agents. Springer-Verlag London, UK ©2002 , ITS '02 Proceedings of the 6th International Conference on Intelligent Tutoring Systems.

Rosas, Ricardo & Nussbaum, Miguel & Cumsille, Patricio & Marianov, Vladimir & Correa, Mónica & Flores, Patricia & Grau, Valeska & Lagos, Francisca & Lopez, Ximena & López, Verónica & Rodríguez, Patricio & Salinas, Marcela. (2003).

Beyond Nintendo: Design and assessment of educational video games for first and second grade students. Computers & Education. 40. 71-94. 10.1016/S0360-1315(02)00099-4.

Rossi, P. H., Wright, J. D. & Anderson, A. B. ., 1983. Handbook of survey research. New York: Academic Press.

Rouillard, J., Serna, A., David, B., Chalon, R., 2014. Rapid Prototyping for Mobile Serious Games, in: Learning and Collaboration Technologies. Technology-Rich Environments for Learning and Col-laboration SE - 20. pp. 194–205.
https://doi.org/10.1007/978-3-319-07485-6_20

Rowe, J., Shores, L., Mott, B. & Lester, J., 2010. Intergrating learning, problem solving and engagement in narrative-centered learning environments.. International journal of artificial intelligence in education, Volume 20, pp. 166-177.

Rowe, J., Shores, L., Mott, B. & Lester, J., 2010. Intergrating learning, problem solving and engagement in narrative-centered learning environments.. International journal of artificial intelligence in education, Volume 20, pp. 166-177.

Ruan, S. et al., 2017. Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. Journal Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies archive, Volume 1(Issue 4).

Ruttkay, Z., Dormann, C. & Noot, H., 2004. Embodied conversational agent on a common ground: A framework for design and evaluation. In: Z. Ruttkay & C. Pelachaud, eds. From brows to trust. Netherlands: Kluwer Academic Publishers, pp. 27-66.

Ryan, R. & Deci, E., 2000. *Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions*. s.l., Contemporary Educational Psychology 25, 54–67.

Sakpal, R. & Wilson, D. M., 2011. Virtual game show host: Dr. Chestr.. Virtual Reality Conference (VR) 2011 IEEE, Singapore, March 2011, pp. 237 -238.

Salinas, A. & S. J., 2006. PDAs and ubiquitous computing in the school.. Pori, Finland. , Proceed-ings of the human centered technology workshop 2006.

Salvucci, S. et al., NCES 97-464. Measurement Error Studies at the National Center For Education Statistics, Washington, DC: United States Department of Education, Office of Education.

Sánchez, J., Olivares, R., 2011. Problem solving and collaboration using mobile serious games. *Comput. Educ.* 57, 1943–1952.
<https://doi.org/10.1016/j.compedu.2011.04.012>

Sanford, K., Starr, L.J., Merkel, L., Kurki, S.B., 2015. Serious games: Video games for good? *E-Learning Digit. Media* 12, 90–106.
<https://doi.org/10.1177/2042753014558380>

SantosPerez, GonzalezParada, E. & J. Canogarcia, 2013. 'Mobile embodied conversational agent for task specific applications'. *IEEE Transactions on Consumer Electronics* , , 59(3).

Sawyer B., Keynote to the Serious Games Summit, GDC, San Francisco, 2007.

Sawyer B., Keynote: Identifying The Serious Games Opportunity: Positioningfor Success, Singa-pore Serious Games Conference, Singapore
2010.<http://www.asiaevents.com.sg/seriousgames2010/>, 2010 (retrieved 1.09.10).

Sawyer, B and Smith,P., 2008. Serious Games Taxonomy, in: Pre-Conference Workshops : May 7 Virtual Worlds & Health Games Accessibility Day. pp. 1–54.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2011). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social cognitive and affective neuroscience*, 7(4), 413-422.

Schroeders, N., Adesope, O. & Gilbert, R., 2013. HOW EFFECTIVE ARE PEDAGOGICAL AGENTS FOR. *Journal of Educational Computing Research*, 49(1), pp. 1-39.

Scott, M., Pereira, L. & Oakley, I., 2015. Show me or tell me: designing avatars for feedback.. *Interacting with computers.*, 27(4), p. 458–469.

Selya, A. S. et al., 2012. A Practical Guide to Calculating Cohen's f², a Measure of Local Effect Size, from PROC MIXED.. *Frontiers in Psychology*, , 111(3), p.
<http://doi.org/10.3389/fpsyg.2012.00111>.

Seng, W.Y., Yatim, M.H.M., 2014. Computer Game as Learning and Teaching Tool for Object Oriented Programming in Higher Education Institution. *Procedia - Soc. Behav. Sci.* 123, 215–224. <https://doi.org/10.1016/j.sbspro.2014.01.1417>

Sepulveda, I., 2015. Physical engagement as a way to increase emotional rapport in interactions with embodied conversational agents. PhD Thesis: The University of Texas at El Paso.

Serious Games Adoption in Corporate Training. Available from:
https://www.researchgate.net/publication/262249501_Serious_Games_Adoption_in_Corporate_Training [accessed Feb 05 2018].

Shabanah, S., 2014. Computer Games for Algorithm Learning. *Handb. Res. Improv. Learn. Motiv. through Educ. Games Multidiscip. Approaches I.*
<https://doi.org/10.4018/978-1-60960-495-0>

Shaffer, D.W., Squire, K.R., Halverson, R., Gee, J.P., 2005. Video Games and The Future of Learning.

Shimoda, T. A., White, B. Y. & Frederiksen, J. R., 1999. Acquiring and transferring intellectual skills with modifiable software agents in a virtual inquiry support environment.. Proceedings of the 32nd International Conference on System Sciences, Los Alamos, CA, IEEE Computer Society..

Shneiderman, B. & Plaisant, C., 2004. Designing the User Interface – Strategies for Effective Human Computer Interaction. 4th ed. s.l.:Pearson Addison Wesley.

Shneidermann, B. & Maes, P., 1997. Direct manipulations vs. interface agents: excerpts from debates at IUI'97 and CHI'97. *Interactions*, Volume 4, pp. 42-61.

Sims, E., 2007. Reusable, lifelike virtual humans for mentoring and role-playing.. *Computers & Education*, 49(1), pp. 75-92.

Sitzmann, T., 2011. A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Pers. Psychol.* 64, 489–528.
<https://doi.org/10.1111/j.1744-6570.2011.01190.x>

Soanes, C., & Stevenson, A. (Eds.). (2005). Oxford dictionary of English (2nd ed.). New York: Oxford University Press. Sonnenschein, S., & Whitehurst, G. J. (1984). Developing referential communication: A hierarchy of skills. *Child Development*, 55, 1936– 1945.

Spence, J.: The use of computer arcade games in behaviour management. *Mal. Ther. Educ.* 6, 64–68 (1988)

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., & Waters, K. (1996). When the interface is a face. *Human-Computer Interaction*, 11(2), 97-124.

Squire, K., Jenkins, H., 2003. Harnessing the power of games in education. *Insight* 3, 5–33. <https://doi.org/10.1080/09500690110067011>

Stair, R. & Reynolds, G., 2001. *Principles of Information Systems*. Boston: Course Technology.

Steckenfinger, S. A., & Ghazanfar, A. A. (2009). Monkey visual behavior falls into the uncanny valley. *Proceedings of the National Academy of Sciences*, 106(43), 18362-18366.

Steinkuehler, C. & Squire, K., 2013. Video Games and Learning MOOC. s.l.:University of Wisconsin-Madison.

Stockley, D., n.a.. E-learning Definition and Explanation (Elearning, Online Training, Online Learn-ing). [Online]

Sturm, J. & Boves, L., 2005. Effective error recovery strategies for multimodal form-filling applications.. *Speech Communication*, Volume 45, pp. 289-303.

Sung, Y., Chang, K. and Liu, T. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education*, 94, pp.252-275.

Susi, T., Johannesson, M., Backlund, P., 2007. Serious games—An overview. Skövde Univ. Skövde Tech. Rep. HSIKTR07001 28.

Suzuki, S. & Yamada, S., 2004. Persuasion through overheard communication by life-like agents. Procs of IEEE/WIC/ACM, s.n.

Swartout, W., 2010. Lessons learned from virtual humans. *AI. Magazine.*, , 31(1), p. 9–20.

Sweetser, P., Wyeth, P., 2005. GameFlow: A Model for Evaluating Player Enjoyment in Games. *Comput. Entertain.* 3, 3–3.
<https://doi.org/10.1145/1077246.1077253>

- Sweller, J., J van Merriënboer, J. & Paas, F. G. W. C., 1998. Cognitive Architecture and Instructional Design. *Educational Psychology Review*, 10(3), p. 251–296.
- T. Bickmore, L. P. D. S., 2011. Relational agents improve engagement and learning in science museum visitors. Berlin Heidelberg, International Workshop on Intelligent Virtual Agents, Springer, p. 55–67.
- Tabachnick, B. G. & Fidell, L. S., 2001. Using Multivariate Statistics. Boston: Allyn and Bacon..
- Takeuchi, A. & Naito, T., 1995. Situated facial displays: Towards social interaction. s.l., Conf. on Human Factors in Computing Systems.
- Thorisson, K., 1996. Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills.. PhD Thesis: MIT Media Laboratory.
- Tibshirani, R. & Hastie, T., 2016. Linear Model Selection and Regularization.. [Online] Available at https://lagunita.stanford.edu/c4x/HumanitiesScience/StatLearning/asset/model_selection.pdf [Accessed 5 June 2017].
- Tibshirani, R., n.d. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Volume Series B (Methodological), pp. 267-288.
- Tramonti, Michela & Lavalle, Arturo. (2014). siLang project – Situated learning and "serious games" towards an effective multicultural communication.
- Trochim, W. M. & Donnelly, J. P., 2006. The research methods knowledge base. 3rd ed. Cincinnati: OH:Atomic Dog.
- Trochim, W. M. & Donnelly, J. P., 2006. The research methods knowledge base. 3rd ed. Cincinnati: OH:Atomic Dog.
- Tseklevs, E., Cosmas, J., & Aggoun, A. (2014). Benefits, barriers and guideline recommendations for the implementation of serious games in education for stakeholders and policymakers. *British Journal of Educational Technology*, 47(1), 164–183. doi:10.1111/bjet.12223
- Tuah, N. M., 2018. The Framework of Anthropomorphic Interface in Gamification Application for, Southampton: University Of Southampton, Faculty Of Physical Sciences And Engineering.
- Ulicsak Mary , Games in Education: Serious Games, Senior researcher, Futurelab

- Uskov, A., Sekar, B., 2014. Serious games, gamification and game engines to support framework activities in engineering: Case studies, analysis, classifications and outcomes, in: IEEE Interna-tional Conference on Electro Information Technology. pp. 618–623. <https://doi.org/10.1109/EIT.2014.6871836>
- Valetsianos, G. & Miller, C., 2008. Conversing with pedagogical agents: A phenomenological exploration of interacting with digital entities.. British Journal of Educational Technology, 39(6), pp. 969-986.
- Van Eck, R. & Adcock, A., 2003. Reliability and factor structure of the Attitude Toward Agent Scale (ATAS).. Chicago, IL, AERA.
- Van Mulken, S., André, E. & & Muller, J., 1998. The Persona Effect: How substantial is it?. In: Proceedings Human Computer Interaction (HCI-98). Sheffield, UK: s.n., pp. 53-66.
- Van Mulken, S., André, E., & Müller, J. (1998). The persona effect: How substantial is it?. In People and computers XIII (pp. 53-66). Springer, London.
- Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods*, 49(2), pp.653-673.
- Veletsianos, G. & Miller, C., 2008. Conversing with pedagogical agents: A phenomenological exploration of interacting with digital entities. British Journal of Educational Technology, Volume 39,(Issue 6), pp. 969-986.
- Veletsianos, G., 2006. Contextual pedagogical agents: stereotypes and first impressions and their impact on student learning and perceptions of agent persona. s.l.:Master thesis.
- Veletsianos, G., 2010. Veletsianos, G. Contextually Relevant Pedagogical Agents: Visual Appearance, Stereotypes, And First Impressions And Their Impact On Learning.. Computers & Education, pp. 576-585.
- Virvou, M., Katsionis, G., Manos, K., 2005. Combining software games with education: Evalua-tion of its educational effectiveness. Educ. Technol. Soc. <https://doi.org/10.1016/j.corsci.2007.02.007>
- Vlachopoulos, D., Makri, A., 2017. The effect of games and simulations on higher education: a systematic literature review. Int. J. Educ. Technol. High. Educ. <https://doi.org/10.1186/s41239-017-0062-1>

Vogel, J.J., Vogel, D.S., Cannon-Bowers, J., Bowers, C.A., Muse, K., Wright, M., 2006. Computer Gaming and Interactive Simulations for Learning: A Meta-Analysis. *J. Educ. Comput. Res.* 34, 229–243. <https://doi.org/10.2190/FLHV-K4WA-WPVQ-H0YM>

Waleed Kadous, M. & Sammut, C., 2002. Mobile Conversational Characters, s.l.: s.n.

Walker. 1994. Discourse and deliberation: testing a collaborative strategy. In Proceedings of the 15th conference on Computational linguistics - Volume 2 (COLING '94), Vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA, 1205-1211. DOI: <https://doi.org/10.3115/991250.991347>

Walldén, S., Soronen, A., 2004. Edutainment: From television and computers to digital television. Univ. Tampere Hypermedia Lab.

Wang, H. & Sun, C. T., 2011. Game reward systems: Gaming experiences and social meanings.. s.l., DiGRA Conference..

Wanner, D., 2014. Serious economic games: Designing a simulation game for an economic experiment, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 782–793. https://doi.org/10.1007/978-3-319-07626-3_74

Waytz, A., Cacioppo, J. & Epley, N., 2014 . Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism. *Perspect Psychol Sci.*, 5(3), p. 219–232..

Weir, C. S., Douglas, G., Carruthers & M. & Jack, M. A., 2009. User perceptions of security, convenience and usability for eBanking authentication tokens.. *Journal of Computers and Security*, Volume 28, pp. 47-62.

Weir, C. S., Douglas, G., Carruthers & M. & Jack, M. A., 2009. User perceptions of security, convenience and usability for eBanking authentication tokens.. *Journal of Computers and Security*, Volume 28, pp. 47-62.

Weiss, B., Wechsung, I., Kühnel, C. & Möller, S., 2015. Evaluating Embodied Conversational Agents in Multimodal Interfaces. *Computational Cognitive Science*, 1(6), pp. 1-21.

- Weizenbaum, J., 1966. ELIZA-A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), pp. 36-45.
- Weizenbaum, Joseph (1976). Computer Power and Human Reason: From Judgment to Calculation. New York: W.H. Freeman and Company. pp. 2, 3, 6, 182, 189. [ISBN 0-7167-0464-1](#).
- Weng, M. et al., 2011. Conceptual Design of Multi-Agent based Personalized Quiz Game. 11th IEEE International Conference on Advanced Learning Technologies,, pp. 19-21.
- White, ., B. Y. & Frederiksen, J. R., 1998. Inquiry, modeling, and metacognition: Making science accessible to all students.. *Cognition & Instruction*, 16(1), pp. 3-118.
- White, B. et al., 2002. Inquiry Island: Affordances of a Multi-Agent Environment for Scientific Inquiry and Reflective Learning. Proceedings of the Fifth International Conference of the Learning Sciences (ICLS), Mahwah, NJ, Erlbaum.
- White, B. Y., 1993. Thinkertools: Causal models, conceptual change, and science education.. *Cognition & Instruction*, 10(1), p. 1–100.
- Whiteside, J., Bennet, J. & Holtzblatt, K., 1988. Usability engineering: Our experience and evolution.. In: M. Helander, ed. *Handbook of Human–Computer Interaction*. Amsterdam: North Holland: s.n.
- Wiemeyer, J., Nacke, L., Moser, C., "Floyd" Mueller, F., 2016. Player Experience, in: Serious Games. pp. 243–271. https://doi.org/10.1007/978-3-319-40612-1_9
- Wik, P., 2011. The Virtual Language Teacher Models and applications for language learning using embodied conversational agents. Doctoral Thesis ed. Stockholm, Sweden: KTH School of Computer Science and Communication.
- Wilkinson, P., 2016. Entertainment Computing and Serious Games 9970, 17–41. <https://doi.org/10.1007/978-3-319-46152-6>
- Willis, J. & Todorov, A., 2006. First impressions: making up your mind after 100 ms exposure to a face.. *Psychological Science*, 17(7), p. 592–598.
- Winograd, T. (1971) Procedures as a Representation for Data in a Computer Program for Understanding Natural Language

Wisdom, J. P., Cavaleri, M. A., Onwuegbuzie, A. J. & Green, C. A., 2012. Methodological reporting in qualitative, quantitative, and mixed methods health services research articles. *Health Services Research*, Volume 47, pp. 721-745.

Wolfe, J., 1997. The effectiveness of business games in strategic management course work. *Simul. Gaming* 28, 360–376.

<https://doi.org/10.1177/1046878197284003>

Woods, W. A. (1973). An experimental parsing system for transition network grammars. In "Natural Language Processing" (R. Rustin, ed.), pp. 111- 154. Algorithmics Press, New York

Wooldridge.(1999) Intelligent Agents In G. Weiss, editor: Multiagent Systems, The MIT Press

Wouters, P., van Nimwegen, C., van Oostendorp, H., van der Spek, E.D., 2013. A meta-analysis of the cognitive and motivational effects of serious games. *J. Educ. Psychol.* 105, 249–265. <https://doi.org/10.1037/a0031311>

Yahoo Finance UK, 'Mobile education platforms key to continued success in the serious game market', 2015. [Online].

Available:<https://uk.finance.yahoo.com/news/mobileeducationplatformskeycontinued000000181.html>. [Accessed: 04Dec2015].

Yan Ru Guo, D. H. Goh and B. Luyt, (2014) "Using affective embodied agents in information literacy education," IEEE/ACM Joint Conference on Digital Libraries, London, 2014, pp. 389-398.doi: 10.1109/JCDL.2014.6970195

Yang, J.C., Chen, C.H., Chang Jeng, M., 2010. Integrating video-capture virtual reality technology into a physically interactive learning environment for English learning. *Comput. Educ.* 55, 1346–1356.

<https://doi.org/10.1016/j.compedu.2010.06.005>

Yankelovich, N., Karat, C. & Lai, J., 2007. CONVERSATIONAL SPEECH INTERFACES AND TECHNOLOGIES. In: A. Sears & J. Jacko, eds. *The Human-Computer Interaction Handbook*. Boca Raton: CRC Press, pp. 383-391.

Yankelovich, N., Karat, C. & Lai, J., 2007. CONVERSATIONAL SPEECH INTERFACES AND TECHNOLOGIES. In: A. Sears & J. Jacko, eds. *The Human-Computer Interaction Handbook*. Boca Raton: CRC Press, pp. 383-391.

Yee, N., Bailenson, J. N., & Rickertsen, K. (2007, April). A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 1-10). ACM.

Yu, A., 2017. "Multicollinearity: SAS Tips. [Online] Available at: Creative-wisdom.com[Accessed 5 June 2017].

Zara, A., Maffiolo, V., MArtins & Devilles, 2007. Collection and Annotation of a Corpus of Human-Human Multimodal Interactions. In: A. Paiva, PradaR & P., eds. ACII 2007, LNCS 4738, . Verlag Berlin Heidelberg: Springer-, p. 464–475.

Zhang, D. & Adipat, B. (2005) Challenges, Methodologies, and Issues in the Usability Testing of Mobile Applications, International Journal of Human-Computer Interaction, 18:3, 293-308, DOI: [10.1207/s15327590ijhc1803_3](https://doi.org/10.1207/s15327590ijhc1803_3)

Zyda, M., 2005. From visual simulation to virtual reality to games. USC Inf. Sci. Inst. 38, 25–32. <https://doi.org/10.1109/MC.2005.297>

Appendices

Appendix A

MoneyWorld mobile Researcher Procedure

Give consent form to participants (MUST BE SIGNED)

Researcher: position chair so it is close to the desk, face on to the monitor.

Select relevant ID (next one on the list) for the participant.

Participant Induction

Thank you for coming to help us today with our research.

My name is, and I'm going to be taking you through the experiment today.

Today we are looking at a mobile based application called MoneyWorld. First there is a short tutorial on the topic. And then there are two versions of the application I would like you to try. Afterwards I would like you to tell us what you think about each of them.

We can stop the session at any time if do not want to continue. Just let me know.

OK, so we'll just start with the tutorial.

[Initial Tutorial]

If you are ready to begin, I'll start it.

Researcher: Start TUTORIAL

Researcher: take note on your sheet of any problems.

Ok, thanks. Now I would just like to ask you a couple of questions about that.

Researcher: ASK questions about Tutorial

[Version1]

OK, thank you. Now I would like you to try the first version of application. If you are ready to begin, I'll start it.

Researcher: Start correct Version according to schedule

Researcher: take note on your sheet of any problems.

Researcher: Enter ID in laptop launcher.

Computer should direct participant to laptop to complete USAB_1 and then AGENT1.

[Version2]

Thanks. Now I'll ask you to try the next version.

If you are ready to begin, I'll start it.

Researcher: Start correct Version according to schedule

Researcher: take note on your log sheet of any problems.

Researcher: Enter ID in laptop launcher.

Computer should direct participant to laptop to complete USAB_2 and then AGENT2

Thank you. So now I'd like to ask you a few questions about your experience today.

Ask questions from Exit Interview

Thank you for helping us today with our research.

Give £10 and receipt slip (MUST BE SIGNED).

Additional notes:

- 1) A scenario version is considered completed (for the purposes of continuing the sessions and administering the questionnaire) when at least one shopping item has been bought (not including when shop-keeper takes over).
- 2) Make sure that the applications needed to run the experiment are connected to the machine and work properly before starting the experiment.
- 3) Check that the mic is working properly.
- 4) Be very careful that you load the correct version of Money World as specified in the 'participant schedule document'.
- 5) In between Money World experience and Usab questionnaire completion, you may have to assist in moving the questionnaire laptop for the participant.

- 6) Make a note of any problems the participant experiences as we may have to take this into consideration when looking at their data.
- 7) Complete data entry as soon as your session is completed.

Version Key

V1	ECA with list A
V2	Text agent with list A
V3	ECA with list B
V4	Text agent with list b

Appendix B

Game playing and devices survey

"Welcome to this technographic survey.

Thank you for agreeing to take part in this survey collecting technographic data for academic research. This survey aims to collect data on the use of technology and games. This survey will take approximately 5-7 minutes to complete. All your answers will be kept in strict confidentiality. The research experiments will be conducted in accordance with the Data Protection Act. Any comments you supply will remain anonymous on a secure university computer. Your information will remain confidential and will be accessible only by me and my supervisors. Any comments you supply will remain anonymous and your information will not go to any third parties. At no time during or after the project will any attempt be made to sell you any products or services as a result of your participation. For any inquires please contact *****@ed.ac.uk. If you wish to continue please click 'Next'."

* 1. Please select your age group:

- 19-25
- 26-30
- 31-35
- 36-40
- 41+
- Prefer not to say

* 2. What gender do you identify with?:

- Male
- Female
- Prefer not to say

* 3. Have you played any computer/console/mobile games (including tablets) in the last 6 months?

- Yes
- No

* 4. In what device do you usually play games ?(You can choose more than one)

- Tablet
- Smart phone
- Smart TV
- PC/Mac desktop
- Game console
- Laptop
- Smart watch
- N/A

* 5. Please indicate which games have you played (You can choose more than one):

- First/third person shooters (e.g. Halo, Call of duty)
- Action/Sport games (e.g. Mario Kart, God of war,
- FIFA)
Real time strategy games (e.g Star craft, Age of Empires,
- Civilization) Casual, Puzzle (e.g. Flappy bird, Puzzle Quest,
- Bejeweled, Solitaire) Simulation/Social games (e.g. Sims,
- Farmville)
Role play games/ Fantasy (e.g. World of warcraft, Final
- Fantasy) Music games (e.g. Guitar Hero, Rock Band)

Other (please specify)

* 6. How many hours do you play per week?(Please select one)

- <2 (up to 2 hours a
week)
- 2-10
(between 2 and 10)
- >10 (over 10)

* 7. Which ones of the following do you own? (You can choose more than one)

- Tablet
- Smart
- phone
- Smart TV
- PC/Mac
- desktop
- Game
console
- Laptop
- Smart watch
- None

Other (please specify)

* Please state how many hours you use every device you own, daily (on average). Please select one group for each device. (work and leisure)

	0-2	2-4	4-6	6+
Tablet	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Smart phone	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Smart TV	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
PC/Mac desktop	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Game console	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Laptop	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Smart watch	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Other (please specify)

* 9. Please state the reason you use every device daily (in case you own one):

	Tablet	Smartphone
Media (e.g. Music, Video)	<input type="checkbox"/>	<input type="checkbox"/>
Communication (Calls/SMS)	<input type="checkbox"/>	<input type="checkbox"/>
Social Media (e.g. Facebook, LinkedIn)	<input type="checkbox"/>	<input type="checkbox"/>
Reading books or documents (e.g. PDF, WORD)	<input type="checkbox"/>	<input type="checkbox"/>
Email	<input type="checkbox"/>	<input type="checkbox"/>
Map/navigation applications	<input type="checkbox"/>	<input type="checkbox"/>
Organiser/calendar	<input type="checkbox"/>	<input type="checkbox"/>
Games	<input type="checkbox"/>	<input type="checkbox"/>
Photography/Camera	<input type="checkbox"/>	<input type="checkbox"/>

Other (please specify)

* 10. Screen size of your devices :

	3"- 4.9"	5"- 6.9"	7"- 10" or more
Tablet	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Smartphone	<input type="radio"/>	<input type="radio"/>	

Other (please specify)

Appendix C

Main experiment: Usability questionnaire Assumption testing

Assumption of normality: One-Sample Kolmogorov-Smirnov Test

One-Sample Kolmogorov-Smirnov Test

		ECA_MEAN	TEXT_MEAN
N		90	90
Normal Parameters ^{a,b}	Mean	5.32	4.40
	Std. Deviation	.76	1.02
Most Extreme Differences	Absolute	.072	.071
	Positive	.070	.065
	Negative	-.072	-.071
Test Statistic		.072	.071
Asymp. Sig. (2-tailed)		.200 ^{c,d}	.200 ^{c,d}

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

d. This is a lower bound of the true significance.

One-way ANOVA³⁹

Research Question:

Is there a statistically significant difference on ECA mean and Text mean by order?

H_0 : There is not a statistically significant difference on ECA mean and Text mean by order.

H_a : There is a statistically significant difference on ECA mean and Text mean by order.

Data Analysis

To examine the research question, an Analysis of Variance (one-way ANOVA) was conducted to determine if there is a significant difference on the ECA mean and Text mean by order. One-way ANOVA is an appropriate statistical analysis when the purpose of research is to assess if mean differences exist on one continuous dependent variable by an independent variable with two or more discrete groups. The dependent variables in this analysis are ECA mean and Text mean, and the independent variable is the order of experience (ECA first, Text first). The assumptions of normality and homogeneity of variance were assessed (Statistics Solutions. (2013)). Normality was assessed using the one-sample Kolmogorov-Smirnov test on both mean scores as reported previously. As seen in Table 40, there was no significant difference

³⁹ Even though ANOVAs are usually used when comparing more than two groups, when having two groups and all assumptions (normal distribution etc.) are checked t-tests and F-tests provide the same results. During the analysis, ANOVAs were preferred in some cases as they control better Type I errors and the results can be more reliable (Laerd statistics, 2013). In this research, both t-tests and F-tests were run in order to verify that the results were the same either way and all the appropriate tests were run in order to secure the validity of these tests.

either on the ECA mean score by order ($df=1$; $F=1.91$; $p.=0.170$) or the Text mean score by order ($df=1$; $F=3.30$; $p.=0.073$). Therefore, the null hypothesis cannot be refuted.

ANOVA

		df	F	Sig.
	Between Groups	1	1.91	.170
ECA_MEAN	Within Groups	88		
	Total	89		
	Between Groups	1	3.30	.073
TEXT_MEAN	Within Groups	88		
	Total	89		

Table 40-ANOVA

Appendix D

Main experiment: API questionnaire

Descriptive statistics and assumption testing

- Is there a statistically significant difference on ECA mean and Text mean by order?
- Is there a difference on perceived agent persona by version (ECA vs. Text)?

Descriptive statistics

Collaborator agent

Tables 41,42 detail the descriptive statistics for the mean scores of the two versions.

Descriptive Statistics				
	order	Mean	Std. Deviation	N
		3.4497	.49484	24
	ECA-TEXT	3.4663	.69866	21
Collaborator agent ECA		3.4574	.59156	45
MEAN		3.8949	.56139	23
	TEXT-ECA	3.8750	.42258	22
		3.8852	.49286	45

		3.6676	.56902	47
	Total	3.6754	.60371	43
		3.6713	.58255	90
		2.3385	.54815	24
	ECA-TEXT	2.4802	.53141	21
		2.4046	.53901	45
		3.3279	.66859	23
Collaborator agent TEXT MEAN	TEXT-ECA	3.1193	.43782	22
		3.2259	.57106	45
		2.8227	.78354	47
	Total	2.8072	.57864	43
		2.8153	.68948	90

Table 41-Descriptive statistics for the collaborator agent.

Instructor agent

Descriptive Statistics				
	order	Mean	Std. Deviation	N
		3.4097	.40763	24
	ECA-TEXT	3.4444	.70580	21
		3.4259	.56000	45
Instructor agent ECA MEAN		3.5344	.60793	23
	TEXT-ECA	3.7822	.54020	22
		3.6556	.58289	45
	Total	3.4707	.51362	47

		3.6172	.64211	43
		3.5407	.57995	90
		2.4878	.54431	24
	ECA-TEXT	2.4980	.49327	21
		2.4926	.51526	45
		3.3315	.64445	23
Instructor agent TEXT MEAN	TEXT-ECA	3.3295	.48777	22
		3.3306	.56676	45
		2.9007	.72698	47
	Total	2.9234	.64164	43
		2.9116	.68379	90

Table 42-Descriptive statistics for the instructor agent.

Assumption testing: One-Sample Kolmogorov-Smirnov Test

Collaborator

One-Sample Kolmogorov-Smirnov Test

	ECA_MEAN	TEXT_MEAN
N	90	90
Normal Parameters ^{a,b}	Mean	3.671
	Std. Deviation	0.582
Asymp. Sig. (2-tailed)		0.463
		0.280

a. Test distribution is Normal.

b. Calculated from data.

Table 43-One-Sample Kolmogorov-Smirnov Test for collaborator agent version means.

Instructor

One-Sample Kolmogorov-Smirnov Test

	ECA_MEAN	TEXT_MEAN
N	90	90
Normal Parameters ^{a,b}		
Mean	3.541	2.911
Std. Deviation	0.580	0.683
Asymp. Sig. (2-tailed)	0.538	0.739

a. Test distribution is Normal.

b. Calculated from data.

Table 45-One-Sample Kolmogorov-Smirnov Test for instructor agent version means.

One-way ANOVA

Is there a statistically significant difference on ECA mean and Text mean by order?

H_0 : There is not a statistically significant difference on ECA mean and Text mean by order.

H_a : There is a statistically significant difference on ECA mean and Text mean by order.

Data Analysis

In order to examine the research question, a one-way ANOVA was conducted to determine if there is a significant difference on the ECA mean and Text mean by order. The dependent variables in this analysis were the ECA mean

and Text mean, and the independent variable was the order of experience (ECA first, Text first). The assumptions of normality and homogeneity of variance were assessed. Again, normality was assessed using the one-sample Kolmogorov-Smirnov test on both mean scores as reported previously.

Collaborator agent

As seen in Table 36, there was a significant difference both on the ECA mean score by order ($df=1$; $F=49.224$; $p.=0.000$) and the Text mean score by order ($df=1$; $F=13.890$; $p.=0.000$). Therefore, it was assumed that the alternative hypothesis was true, so there was a statistically significant difference on the ECA and Text means by order.

ANOVA

		df	F	Sig.
	Between Groups	1	49.224	0.000
ECA_MEAN	Within Groups	88		
	Total	89		
	Between Groups	1	13.890	0.000
TEXT_MEAN	Within Groups	88		
	Total	89		

Table 36-One-way ANOVA on collaborator agent persona.

Instructor Agent

As seen in Table 37, there was not a significant difference on the ECA mean score by order ($df=1$; $F=3.632$; $p.=0.060$), while there was a significant

difference on the Text mean score by order ($df=1$; $F=53.857$; $p.=0.000$).

Therefore, the null hypothesis was accepted for the ECA and rejected for the Text.

ANOVA

		df	F	Sig.
	Between Groups	1	3.632	.060
ECA_MEAN	Within Groups	88		
	Total	89		
	Between Groups	1	53.857	.000
TEXT_MEAN	Within Groups	88		
	Total	89		

Table 37-One-way ANOVA on instructor agent persona.

Individual statements Type I and Type II error

Type I error

In order to avoid a Type I error for multiple t-tests (for all 24 statements), a Bonferroni Correction was run.

Collaborator agent

A post-hoc Bonferroni Correction test showed that all ECA statements' scores were found to be statistically significant compared to the Text statements' scores apart from 5 items (encouraged me to reflect, focus, improve my knowledge, helpful, useful) and one where the Text version was statistically significant over the ECA version (The agent was instructor like).

Instructor agent

Post-hoc Bonferroni Correction test showed that all ECA statements' scores were found to be statistically significant compared to the Text statements' scores.

Calculate Bonferroni Correction

Alpha: 0.05

R: 24

With no correction the chance of finding one or more significant differences in 24 tests = 0.708 (70.8%).

Bonferroni's adjustment:

Lower the 0.05 to **0.0020833**

Significant Effects in API Attributes

Collaborator agent

The main effect of version was significant for 19 out of the 24 attributes. The analysis showed significant main effect for the order factor for all the attributes apart from 1 (instructor-like). The main effect of shopping list was not significant for any of the 24 attributes. The interaction between the

version and order of experience was significant for 8 out of 24 attributes (made interesting, encouraged to reflect, improved knowledge, natural emotion, personality, human-like, motivating, friendly). There was also significant interaction between the version and the shopping list for 1 of the 24 attributes (needs a lot of improvement). Also significant was the interaction between the order and the shopping list for 1 of the 24 attributes (intelligent).

The ECA agent was rated significantly better than the Text agent in all the cases except for one (instructor-like).

Overall, many significant results were found for the two versions. These are summarised in Table 42.

Attribute	Significant Differences
The agent kept my attention. -	Order (df=1; F= 18.676 ; p = 0.000)
The agent made the instruction interesting. -	Order (df=1; F= 26.755; p = 0.000), Version*order (df=1; F=7.050; p.=0.009)
The agent presented the material effectively. -	Order (df=1; F= 4.624; p = 0.034)
The agent helped me to concentrate on the presentation. -	Order (df=1; F=26.387; p = 0.000)
The agent was knowledgeable. -	Order (df=1; F= 9.157; p = 0.000)
The agent encouraged me to reflect what I was learning. -	Version*order (df=1; F=16.347; p.=0.000), Order (df=1; F= 8.344; p = 0.005)
The agent was enthusiastic. -	Order (df=1; F= 36.665 ; p = 0.000)
The agent led me to think more deeply about the presentation. -	Order (df=1; F= 5.426; p = 0.022)
The agent focused me on the relevant information. -	Order (df=1; F= 6.475; p = 0.013)

The agent improved my knowledge of the content. -	Version*order (df=1; F= 14.165; p.=0.000), Order (df=1; F= 13.313; p = 0.000)
The agent was interesting. -	Order (df=1; F= 35.098 ; p = 0.000)
The agent was enjoyable. -	Order (df=1; F= 30.314; p = 0.000)
The agent was instructor-like. -	-
The agent was helpful. -	Order (df=1; F= 6.985; p = 0.010)
The agent was useful. -	Order (df=1; F= 10.125; p = 0.002)
The agent showed emotion. -	Order (df=1; F= 37.780 ; p = 0.000)
The agent has a personality. -	Version*order (df=1; F= 4.766; p.=0.032), Order (df=1; F= 27.028 ; p = 0.000)
The agent's emotion was natural. -	Version*order (df=1; F=6.898; p.=0.010), Order (df=1; F= 17.566; p = 0.000)
The agent was human-like. -	Version*order (df=1; F= 34.563; p = 0.000)
The agent was expressive. -	Order (df=1; F= 24.799 ; p = 0.000)
The agent was entertaining. -	Order (df=1; F= 25.621; p = 0.000)
The agent was intelligent. -	Order*list order (df=1; F=8.178; p.=0.005), Order (df=1; F= 11.464; p = 0.000)
The agent was motivating. -	Version*order (df=1; F=4.453; p.=0.038), Order (df=1; F= 24.896; p = 0.000)
The agent was friendly. -	Version*order (df=1; F=19.959; p.=0.000), Order (df=1; F= 45.631 ; p = 0.000)

Table 42-Summary of Significant Effects per Attribute.

Instructor agent

The main effect of version was significant for all the attributes. The analysis showed a significant main effect for the order factor for all the attributes apart from 3 (instructor-like, helpful, useful). The main effect of shopping list was not significant for any of the 24 attributes. The interaction between version and order of experience was significant for all the attributes apart from 5 (instructor-like, helpful, useful, kept my attention, personality, emotion).

The ECA agent was rated significantly better than the Text agent in all the cases.

Overall, many significant results were found for the two versions. These are summarised in Table 43.

Attribute	Significant Differences
The agent kept my attention. -	Order (df=1; F= 18.983 ; p = 0.000)
The agent made the instruction interesting. -	Order (df=1; F= 20.972; p = 0.000), Version*order (df=1; F=13.858; p.=0.000)
The agent presented the material effectively. -	Version*order (df=1; F=5.497; p.=0.021) Order (df=1; F= 4.281; p = 0.042)
The agent helped me to concentrate on the presentation. -	Version*order (df=1; F=14.460; p.=0.000), Order (df=1; F=18.153; p = 0.000)
The agent was knowledgeable. -	Version*order (df=1; F=7.601; p.=0.007) Order (df=1; F= 4.671; p = 0.033)
The agent encouraged me to reflect what I was learning. -	Version*order (df=1; F=18.247; p.=0.000), Order (df=1; F= 4.119; p = 0.045)
The agent was enthusiastic. -	Version*order (df=1; F=4.191; p.=0.044), Order (df=1; F= 26.106 ; p = 0.000)
The agent led me to think more deeply about the presentation. -	Version*order (df=1; F=6.902; p.=0.010), Order (df=1; F= 8.192; p = 0.005)
The agent focused me on the relevant information. -	Version*order (df=1; F=7.754; p.=0.007), Order (df=1; F= 4.616; p = 0.034)
The agent improved my knowledge of the content. -	Version*order (df=1; F= 14.191; p.=0.000), Order (df=1; F= 7.764; p = 0.007)
The agent was interesting. -	Version*order (df=1; F=7.379; p.=0.008), Order (df=1; F= 34.758 ; p = 0.000)
The agent was enjoyable. -	Version*order (df=1; F=6.308; p.=0.014), Order (df=1; F= 30.079; p = 0.000)
The agent was instructor-like. -	-

The agent was helpful. -	-
The agent was useful. -	-
The agent showed emotion. -	Order (df=1; F= 18.988 ; p = 0.000)
The agent has a personality. -	Order (df=1; F= 27.615 ; p = 0.000)
	Version*order (df=1; F=4.913; p.=0.029), Order
The agent's emotion was natural. -	(df=1; F= 17.676; p = 0.000)
	Version*order (df=1; F=4.923; p.=0.029), Order
The agent was human-like. -	(df=1; F= 34.270; p = 0.000)
	Version*order (df=1; F=5.717; p.=0.019), Order
The agent was expressive. -	(df=1; F= 29.334; p = 0.000)
	Version*order (df=1; F=5.100; p.=0.000), Order
The agent was entertaining. -	(df=1; F= 5.100; p = 0.026)
	Version*order (df=1; F=7.049; p.=0.009), Order
The agent was intelligent. -	(df=1; F= 8.085; p = 0.006)
	Version*order (df=1; F=11.100; p.=0.001), Order
The agent was motivating. -	(df=1; F= 15.254; p = 0.000)
	Version*order (df=1; F=19.315; p.=0.000), Order
The agent was friendly. -	(df=1; F= 20.952; p = 0.000)

Table 43-Summary of Significant Effects per Attribute.

Repeated measures ANOVA

Research Question:

RQ: Is there a difference on perceived agent persona by version (ECA vs. Text)?

H₀: There is no difference on perceived agent persona by version (ECA vs. Text).

H_a: There is a difference on perceived agent persona by version (ECA vs. Text).

Data Analysis

To examine this research question, a repeated-measures analysis of variance (ANOVA) was conducted to assess if mean differences existed on the agent persona by version (ECA vs. Text). Normality was checked with skewness and kurtosis values, and sphericity was assessed through a Mauchly's Test of Sphericity. Since there were only two conditions, sphericity was 1.

Collaborator agent

The results of the ANOVA for the within-subjects variable show that the significance for version was $p.=0.000$ ($df=1$; $F=186.212$; $p.=0.000$) and a statistically significant effect of relationship between version and order ($df=1$; $F=9.551$; $p.=0.003$). The statistically significant difference of the perceived agent persona between the two versions for the collaborator agent led to the rejection of the null hypothesis and the acceptance that there was a difference on the agent persona by version. Also, there was a statistically significant between-subjects effect of order of experience ($df=1$; $F=39.820$; $p.=0.000$); therefore, the difference between the two orders was statistically significant. Since there are only two conditions, sphericity is not an issue in this experiment and, therefore, the sphericity-assumed data are explored.

Tables 45 and 46 give the Tests of Within-Subjects Effects and Tests of Between-Subjects Effects, respectively.

Tests of Within-Subjects Effects

Source		df	F	Sig.
version	Sphericity Assumed	1	186.212	.000
version * order	Sphericity Assumed	1	9.551	.003

Table 46-Tests of Within-Subjects Effects for collaborator agent.

Tests of Between-Subjects Effects

Source	df	F	Sig.
Intercept	1	4351.635	.000
order	1	39.820	.000

Table 47-Tests of Between-Subjects Effects for collaborator agent.

Instructor agent

Again, the results of the ANOVA for the within-subjects variable show that the significance for version was $p.=0.000$ ($df=1$; $F=86.332$; $p.=0.000$). The statistically significant difference of the perceived agent persona between the two versions led to the rejection of the null hypothesis and the acceptance that there was a difference on the instructor-agent persona by version. Also, a repeated measures ANOVA on the overall mean scores found that the relationship between the version and order of experience was significant with

$p.=.000$ ($df=1$; $F=19.934$; $p.=0.000$). The between-subjects effect of order of experience was also statistically significant with $p.=000$ ($df=1$; $F=30.579$; $p.=0.000$) that indicates a difference between orders. Since there are only two conditions, sphericity is not an issue in this experiment and, therefore, the sphericity-assumed data are studied.

Tables 47 and 48 give the Tests of Within-Subjects Effects and Tests of Between-Subjects Effects, respectively.

Tests of Within-Subjects Effects

Source		df	F	Sig.
version	Sphericity Assumed	1	86.332	.000
version * order	Sphericity Assumed	1	19.934	.000

Table 48-Tests of Within-Subjects Effects for instructor agent.

Tests of Between-Subjects Effects

Source	df	F	Sig.
Intercept	1	4460.497	.000
order	1	30.579	.000

Table 50-Tests of Between-Subjects Effects for instructor agent.

Appendix E

Main experiment: Regression analysis

Regression model assessment for collaborator and instructor agents

Shopkeeper-collaborator agent

Multiple Linear Regression

Assessing the regression model I: diagnostics

Fig.1 gives the scatterplot of the dependent variable and the regression standardised predicted value from the full model of the shopkeeper-interaction partner agent (9 independent variables); no outliers and residuals were identified. Also, upon further examination for influential cases, none were detected.

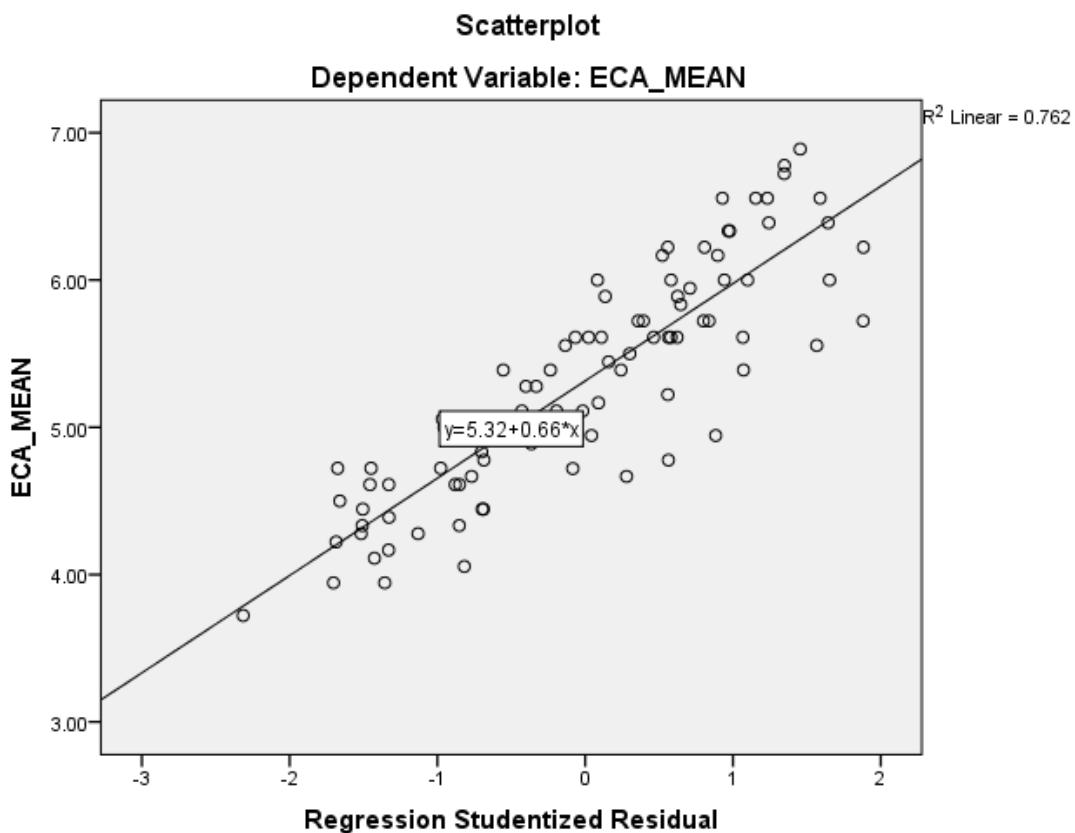


Figure 37-Scatterplot of the dependent variable and the regression standardised predicted value from the full model (9 independent variables).

Assessing the regression model II: generalisation

How much of the Usability can be explained by the 9 API attributes?

The relevant assumptions of this analysis were tested prior to the multiple regression analysis.

In this research, all predictor variables are quantitative, and the dependent variable is an aggregated score of the Likert scale which means it is quantitative and continuous but bounded since the data collected vary between 1 and 7. The assumption of non-zero variance was met as the predictors vary in value.

Outliers and influential cases were identified in an initial data screening and were modified. An examination of the Mahalanobis distances indicated no multivariate outliers. However, one case had a value of more than 28 and with a sample of less than 100 and 9 predictors, values greater than 27.88 are considered problematic (Field, 2013). Upon further investigation though, the Mahalanobis values were compared to chi square distributions and none was lower than 0.001; thus, it was deemed that no multivariate outliers existed.

None of the external variables correlated too highly (>0.8) with the ones selected in the model. Yet, the nature of this questionnaire was such that the items were correlated at some level. The assumption of independence was also met – all the values of the outcome variable were independent. Residual and scatter plots indicate that the assumptions of linearity, homoscedasticity and normality were met as seen in the following figures (fig38,39) (Hair, et al., 1998; Pallant, 2013).

The assumption for independent errors was deemed to be inconclusive. This is because the closer the Durbin-Watson value to 2 is, the better, and for these data the value was 1.636. Upon further investigation for models with intercept, Savin and White (1977) suggest a lower limit (dL) of 1.312 and an upper limit (dU) of 1.741. Over dU, the null hypothesis that the residuals from an ordinary least-squares regression are not autocorrelated is not rejected. Since the test statistic value from this model was 1.636, that is between dL and dU, the test is inconclusive.

For models with an intercept and observed test statistic value lower than 2 (Savin and White (1977))	K=9	
	dL	dU
N=90	1.312	1.741

Table 21-Positive serial correlation Durbin-Watson Five Per Cent Minimal Bound (William N. Evans, Econometrics University of Notre Dame) N : number of samples, K: number of predictors.

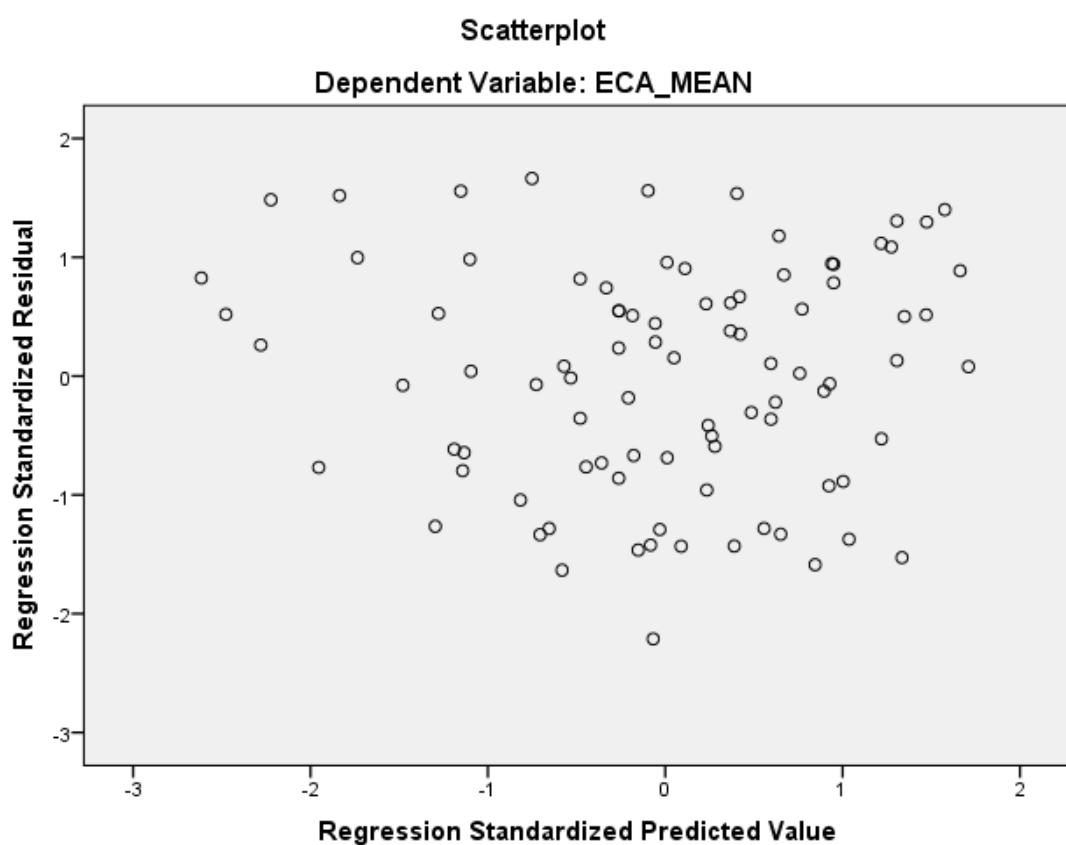


Figure 38-Scatterplot showing that homoscedasticity has been met.

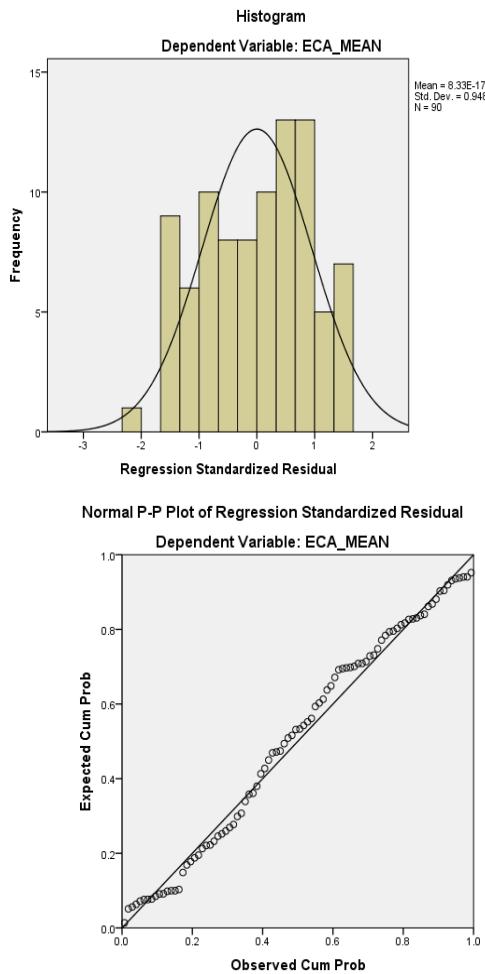


Figure 39-Histogram and normal P-P plots showing the normal distribution of the residuals.

An examination of the correlations between the independent variables revealed that none were highly correlated (>0.8). All correlations were positive and small to moderate, ranging from .36 ("The agent emotion was natural" and "The agent was entertaining") to .72 ("The agent showed emotion" and "The agent was expressive"). However, since they were correlated to a degree, the collinearity statistics (i.e., Tolerance and VIF) were examined and all were found within accepted limits. Thus, the assumption of multicollinearity has been met (Hair, et al., 1998; Coakes, 2005).

The correlations between the dependent variable (mean usability) and the 9 independent variables, were all positive and small to moderate ranging from .16 (The agent was friendly) to .38 (The agent was human-like). This is an indication that the data were suitably correlated with the dependent variable in order to be examined with multiple linear regression.

Alex- instructor agent

Multiple Linear Regression

Assessing the regression model I: diagnostics

As seen in fig. 1 from the scatterplot of the dependent variable and the regression standardised predicted value from the full model for the Alex-instructor agent (9 independent variables), no outliers and residuals were identified. Also, upon further examination for influential cases none were detected.

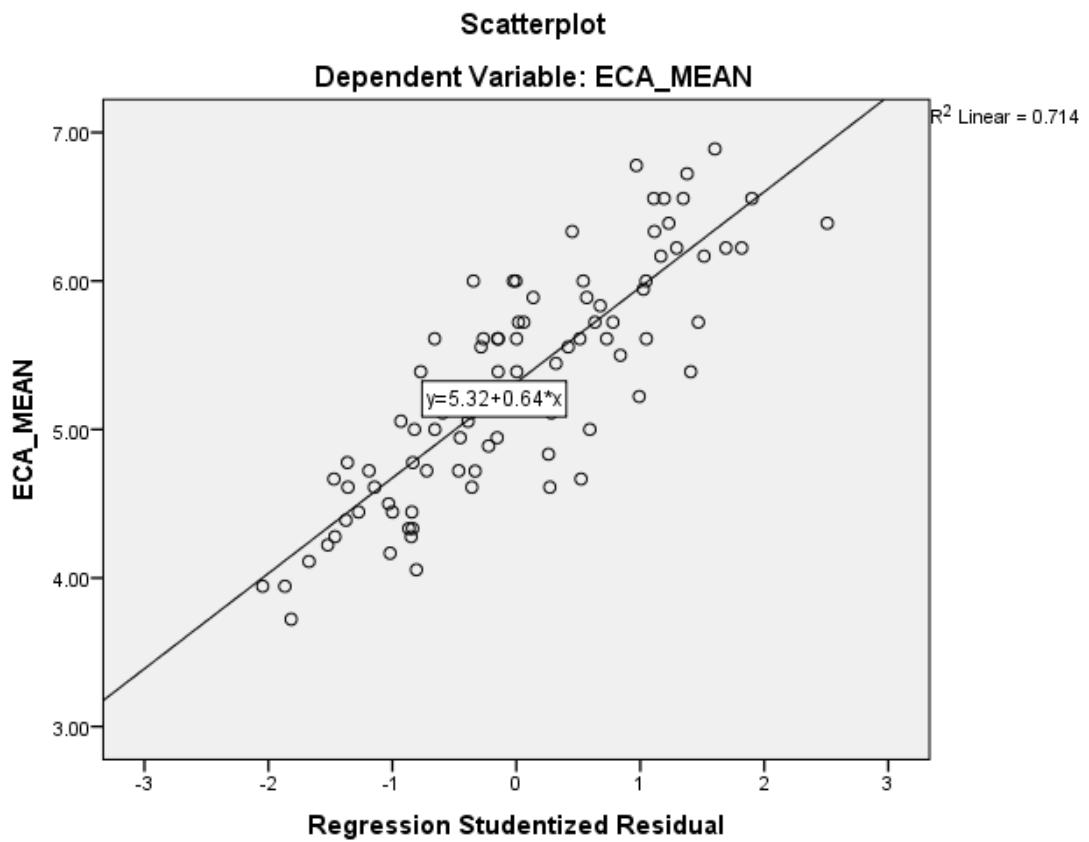


Figure 40-Scatterplot of the dependent variable and the regression standardized predicted value from the full model for Alex- instructor agent (9 independent variables).

Assessing the regression model II: generalization

How much of the Usability can be explained by the 9 API attributes?

The relevant assumptions of this analysis were tested prior to the multiple regression analysis. The assumption of non-zero variance was met as the predictors have variation in value.

Outliers and influential cases identified in initial data screening and modified as mentioned previously. An examination of the Mahalanobis distances indicated no multivariate outliers.

None of the external variables were highly correlated (> 0.8) with the ones selected in the first model. The assumption of independence was also met as all the values of the outcome variable are independent. Residual and scatter plots indicated the assumptions of linearity, homoscedasticity and normality were met as seen in the following figures (fig.41) (Hair, et al., 1998; Pallant, 2013).

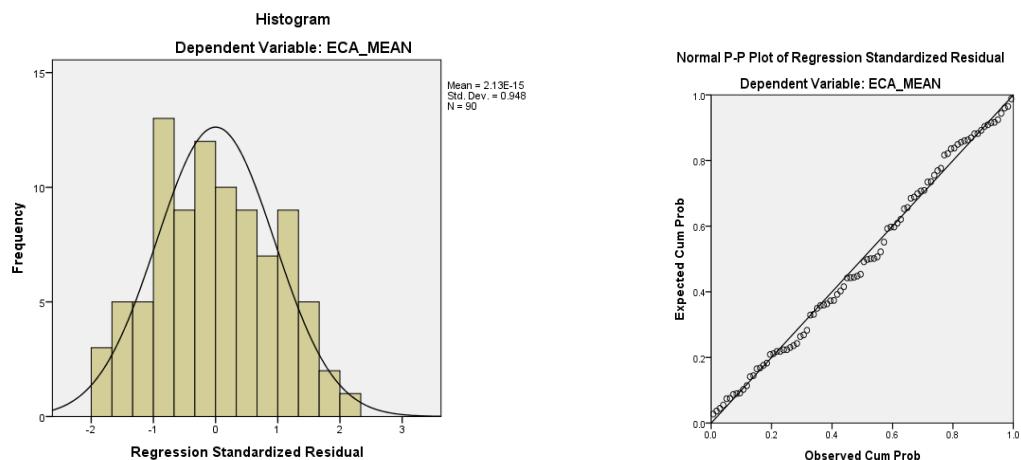


Figure 41-Histogram and normal P-P plots showing the normal distribution of the residuals.

The assumption for independent errors was deemed to be met with a Durbin-Watson value of 1.764. The value is also over the upper limit suggested by Savin and White (1977); thus, the null hypothesis of the residuals from an ordinary least-squares regression being not autocorrelated was not rejected.

For models with an intercept and observed test statistic value lower than 2 (Savin and White (1977))	K=9	
	dL	dU
N=90	1.312	1.741

Table 52-Positive serial correlation Durbin-Watson Five Per Cent Minimal Bound (William N. Evans, Econometrics University of Notre Dame) N : number of samples, K: number of predictors.

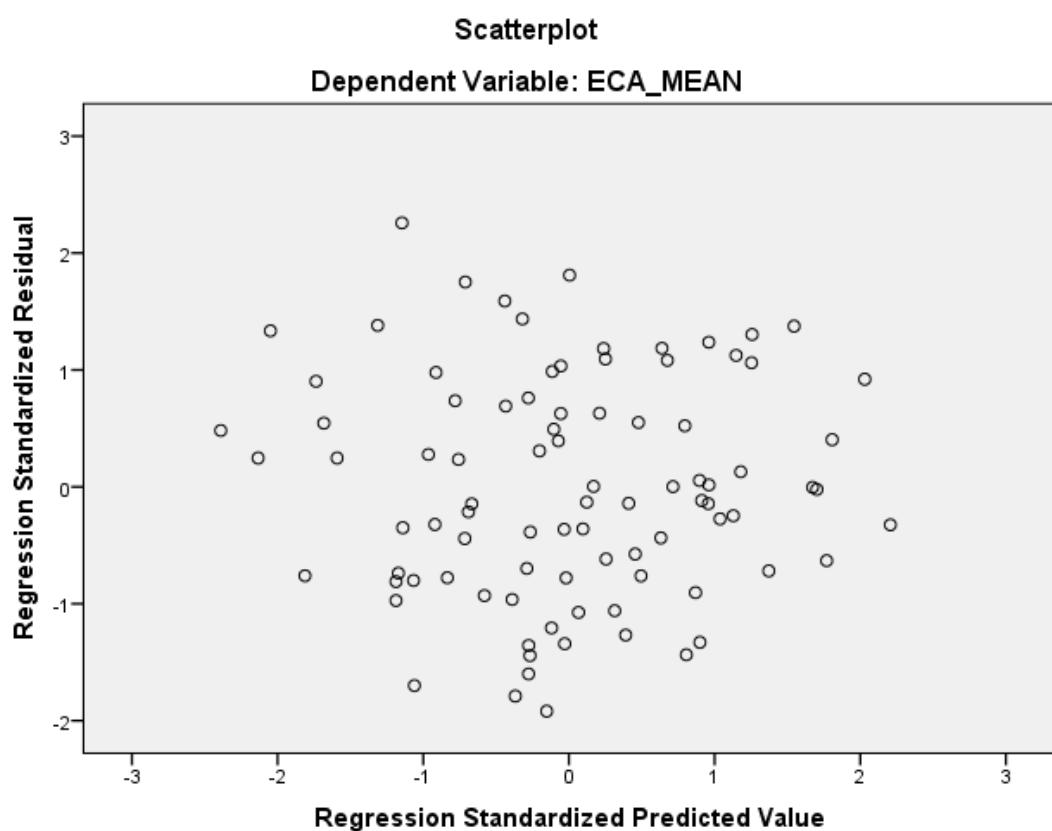


Figure 42-Scatterplot showing that homoscedasticity has been met.

An examination of the correlations between the independent variables revealed that none were highly correlated (> 0.8). All correlations were positive and small to moderate, ranging between .33 (The agent emotion was friendly and The agent's emotion was natural) and .72 (The agent was human-like and The agent showed emotion). However, since they were correlated to some degree, the collinearity statistics (i.e., Tolerance and VIF) were examined and all were found to be within accepted limits. Thus, the assumption of multicollinearity has been met (Hair, et al., 1998; Coakes, 2005).

The correlations between the dependent variable (ECA version mean usability) and the 9 independent variables were all positive and small to moderate, ranging from .19 (The agent was friendly) to .45 (The agent was entertaining). This is an indication that the data are suitably correlated with the dependent variable in order to be examined with multiple linear regression.

Appendix F

Preliminary work: Pilot study 1

Examination of individual attributes

Individual attributes

Although the main test compared the overall means of each version (Game-Learn), it indicated which attributes were significant or not as it was an omnibus statistical test. To examine any differences for each of the individual attributes on the questionnaire between the versions, a t-test was run on the mean scores of each question; version was the within-participants factor and order of experience was the between-participants' factors. The results of these tests were reported in Table 10.

Research Question:

Is there a statistically significant difference for each of the individual attributes on the questionnaire between the two versions?

H_0 : There is not a statistically significant difference for each of the individual attributes on the questionnaire between the two versions.

H_a : There is a statistically significant difference for each of the individual attributes on the questionnaire between the two versions.

Data Analysis

Based on the data from the paired samples t-test summarised in Table 11, the null hypothesis was rejected for the three attributes "I enjoyed using Moneyworld", "I thought Moneyworld was fun" and "I found the use of

Moneyworld stimulating" meaning that the difference between versions was statistically significant for the three attributes. The Game version was rated significantly better than the Learn version in all the cases.

Overall, many significant results were found for the two versions. These are summarised in Table 5.

Attribute	Significant Differences
Confusion	-
Concentration	Version * order (df=1; F=14.114; p < 0.001)
Flustered	-
Stressed	-
Relaxed	-
Nervous	Version * order (df=1; F=7.439; p = 0.008)
Frustrating	-
Embarrassed	Version * order (df=1; F=14.368; p < 0.001)
Knew what to do	-
Felt in control	Version * order (df=1; F=15.703; p < 0.001)
Happy to use again	Order (df=1; F=7.230; p = 0.009)
Needs improvement	-
Enjoyment	Version (df=1; F=4.053; p = 0.049) Order (df=1; F=9.696; p = 0.003)
Fun	Version (df=1; F=10.055; p = 0.002)
Felt part of	-
Stimulating	Version (df=1; F=4.152; p = 0.046)

Order (df=1; F=4.958; p = 0.030)

Easy to use -

Complicated -

Table 53: Summary of Significant Effects per Attribute.

Appendix G

Exit questionnaire sample

MoneyWorld3 ID Procedure

Participant ID: **102**

For the first part of the experiment, **Launch Tutorial**

After the tutorial is completed, ask the questions about the tutorial.

Then launch the first version of the game: **V2**

After this version, get participant to complete **USAB_1b**
Agent1_b

Then launch the second version of the game: **V3**

After this version, get participant to complete **USAB_2b**
Agent2_b

After both versions are completed, ask the exit interview.

NOTES

Please take a note of any errors, problems or observations made by the participant.

Exp1:

- 1) Not sure if she should talk
- 2) Missed the cue on speech recognition

Exp2:

- 3) Spoke the right time

Exit interview order 2

MoneyWorld Mobile: Interview Questionnaire (please make sure you record all the comments)

Tutorial questions

Q1. What did you think of the tutorial? (elaborate)

It was interesting. Did not know about the old currency.

Q2. Did you find the tutorial helpful?

- Yes
- No
- No answer

Comments:

If no, why is that?

Q3. Did you feel you understood the old money?

- Yes
- No
- No answer

Comments:

If no, why is that?

Q1. General preference

Today you experienced two versions of Moneyworld. Which version did you prefer?

First Second no preference

Please give me reasons for your answer:

The shopkeeper made me feel relaxed. It was more interactive and enjoyable.

Q2.Text

In the first version, the interface that you interacted with in order to buy the items on the list looked like this (show text SK).

What did you think about it?

Good but poor compared to the ECA version which was an improvement.

Q3. In the first version, the interface that assist you with the list looked like this (show text Alex).

What did you think about it?

Good but poor compared to the ECA that was better.

Q4. ECA

In the second version, the interface that you interacted with in order to buy the items on the list looked like this (show photorealistic SK).

What did you think about it?

Really liked him. He was polite and funny.

Q5. In the second version, the interface that assist you with the list looked like this (show photorealistic Alex). What did you think about it?

It was better, more interactive.

(show screenshots of all four agents in pairs)

Q6.a. Which system did you prefer to be assisted from? **(do not read them out)**

- System with characters
- System with text only
- No preference

Q6.b. Can you please give me reasons for your answer?

Felt more interactive. I felt connected. Versions with characters was more enjoyable and amusing.

Q7.a. Which system did you prefer to interact with on the shop? **(do not read them out)**

- System with characters
- System with text only
- No preference

Q7.b. Can you please give me reasons for your answer?

Felt more interactive. I felt connected. Versions with characters was more enjoyable and amusing.

Q8.a. Do you use agents/assistants such as Siri/Cortana/Speaktoit on your smartphone in your everyday life?

- Yes
- No
- Sometimes

Q8.b. Which ones do you use?

Siri

Q8.c. For what tasks?

Calls, fun

Q9.a. What do you like

It was humorous, helpful and a more natural interaction.

or dislike about this kind of interface?

I had to repeat often

Q9.b. Why?

I use natural language and talking instead of writing.

Q10. Would you use a game with speech recognition like Moneyworld on your phone?

Yes

(Do not ask, used for verification purposes)

Gender

male female

Age 29

Appendix H

Participation Acceptance Form

Research Experiments at The University of Edinburgh

Participation Acceptance Form

Yes, I am interested in taking part in research experiments for the Doctoral studies of Miss Danai Korre at The University of Edinburgh.

My details are:

Full name:	
Telephone number(s):	
E-mail address:	
Date of birth:	

I understand that the experiments are being carried out by,

Miss Danai Korre as part of her for her Doctoral studies.

I understand that I will be offered a small compensation for any individual research experiments in which I take part.

I accept that the research experiments will be conducted in accordance with the Data Protection Act and the Code of Conduct of the Market Research Society.

I understand that any comments I supply will remain anonymous.

I understand that my information will not go to any third parties.

I understand that at no attempt will be made to sell me any products or services as a result of my participation.

I understand that I will be able to withdraw from my involvement in the experiments at any time.

Signature:

A large, empty rectangular box with a thin black border, intended for a handwritten signature.