Development and Evaluation of an Autonomous, Virtual Agent Based Social Skills Tutor for Children with Autism

by

Marissa Kate Milne

Thesis
Submitted to Flinders University
for the degree of

Doctor of Philosophy

College of Science and Engineering

16th March 2018

CONTENTS

Contents ... **ii**
Abstract ... **vi**
Declaration ... **vii**
Acknowledgements ... **viii**
Related Publications ... **ix**
List of Tables ... **x**
List of Figures ... **xii**
Chapter 1. Introduction ... **1**
Chapter 2. Background and Literature Review ... **7**
2.1 About Autism ... 7
2.1.1 Social Skills Deficit ... 9
2.1.2 Language and Communication Deficits ... 10
2.1.3 Repetitive Interests and Behaviours ... 10
2.1.4 Co-morbid Conditions ... 11
2.1.5 Generalisation and Maintenance ... 12
2.2 Existing Social Skills Interventions ... 15
2.2.1 Traditional Interventions ... 15
2.2.2 Technology Based Interventions ... 19
2.3 General Educational Considerations ... 25
2.3.1 What Human Tutors Do ... 26
2.3.2 Emotional Aspects of Learning ... 28
2.4 Advantages of Virtual Tutors ... 29
2.5 Existing Virtual Tutors ... 31
2.5.1 Language and Reading Tutors ... 31
2.5.2 Mathematics and Science Tutors ... 34
2.5.3 Social Skills Tutors ... 35
2.6 Using Virtual Tutors as Interventions ... 37
2.6.1 General Development Guidelines ... 37
2.6.2 Autism Specific Considerations ... 39
2.6.3 Socially Responsive Agents ... 41
2.6.4 Assistive Technologies ... 42
2.7 Existing Social Skills Curricula ... 44
2.7.1 Requirements for Software-Based Implementation ... 45
2.7.2 Considered Curricula ... 45
2.7.3 Selected Curricula ... 47
2.8 Assessment Tools and Techniques ... 50
2.8.1 Assessment within Software ... 50
2.8.2 Assessment for Evaluation of Software ... 57
2.9 Conclusion ... 58
Chapter 3. Research Aims ... **61**
Chapter 4. Research Objective 1 - Software Implementation ... **64**
4.1 Curriculum Implementation ... 64
4.2 Architecture Overview ... 68

4.3 Head X Virtual Agent 69
4.4 Lesson Interaction and Display 72
4.4.1 Lesson Authoring Process 73
4.4.2 Lesson Design Guidelines 75
4.5 Lesson Types 78
4.6 Other Whiteboard Functions 78
4.6.2 Data Logging and Use 82
4.6.3 Automated Assessment 82
4.6.4 Rewards and Reinforcement 86
Chapter 5. Evaluation Method 88
5.1 Overview and Justification 88
5.2 Methodology 89
5.2.1 Recruitment and Inclusion Criteria 89
5.2.2 Group Allocation 90
5.2.3 Selection and Design of Tools 91
5.2.4 Data Collection Schedule 93
5.3 Experimental Software 95
5.3.1 Typical Student Workflow 95
5.3.2 Data Collection Workflow 97
5.3.3 Experimental 'Social Content' Group 97
5.3.4 Control 'Maze Games' Group 98
5.4 Challenges Encountered During Evaluation 98
Chapter 6. Results and Analysis 100
6.1 Participant Demographics 100
6.2 Research Objective 1 - Software Implementation 101
6.3 Research Objective 2 - Changes in Knowledge 104
6.3.1 Demographics and Assumption Testing 105
6.3.2 Primary Analysis 109
6.3.3 Exploratory Analysis 110
6.3.4 Summary 120
6.4 Research Objective 3 - Changes in Behaviour 120
6.4.1 Demographics and Assumption Testing 121
6.4.2 Primary Analysis 121
6.4.3 Exploratory Analysis 122
6.4.4 Summary 128
6.5 Research Objective 4 - Maintenance of Skills 128
6.5.1 Task Completion 129
6.5.2 Primary Analysis 129
6.5.3 Exploratory Analysis 130
6.5.4 Summary 134
6.6 Research Objective 5 - Perceptions of Software 135
6.6.1 Pre-test Questionnaire 135
6.6.2 Post-test Questionnaire 136
6.6.3 Summary 145
Chapter 7. Discussion 146
7.1 Research Objective 1 - Software Implementation 146

7.2 Research Objective 2 - Changes in Knowledge 149
7.2.1 Content Quiz as an Assessment Tool 150
7.2.2 Response Subgroups 150
7.2.3 Correctness 151
7.2.4 Accuracy 155
7.2.5 Duration 156
7.2.6 Demographics 157
7.3 Research Objective 3 - Changes in Behaviour 158
7.3.1 Vineland-II as an Assessment Tool 158
7.3.2 Outcomes by Domain and Subdomain 159
7.3.3 Outcomes by Item 160
7.3.4 Adaptive Level and Demographics 161
7.4 Research Objective 4 - Maintenance of Skills 162
7.4.1 Content Quiz 162
7.4.2 Behavioural Assessment 164
7.5 Research Objective 5 - Perceptions of Software 165
7.5.1 Pre-test 166
7.5.2 Post-test 167
Chapter 8. Software Future Directions 173
8.1 Retained Features 173
8.2 General Modifications 174
8.3 Educational Content 175
8.4 Personalisation and Customisation 176
8.5 Assessment and Sequencing Algorithms 179
8.6 Authentic Interaction 180
8.7 Participatory Design 182
Chapter 9. Conclusion 183
9.1 Key Findings and Significance 183
9.2 Study Quality, Limitations and Recommendations 186
9.3 Closing Statement 188
Glossary 190
Appendices 191
Appendix A. Content Quiz Questions and Expected Answers 191
Appendix B. Pre-test Questionnaire 204
Appendix C. Post-test Questionnaire 205
Appendix D. Statistical Formulae and Conventions 206
Appendix E. Participant Completion of Data Collection Tasks 207
Appendix F. Additional Demographic Analyses 208
Appendix G. Primary Analysis Data 209
Appendix H. Detailed Content Quiz Correctness Tables 210
Appendix I. Detailed Content Quiz Duration Tables 211

Appendix J. Pre-Identified Vineland-II Items 213
Appendix K. Detailed Questionnaire Responses 216
Appendix L. Ethics Approvals 217
Appendix M. Recruitment Materials 219
Appendix N. Family Information Pack 220
Appendix O. Consent Form 225
Reference List 227

ABSTRACT

This is a multidisciplinary thesis comprising principally of a computer science component involving the development of educational software for teaching children with autism social skills using virtual humans, with a follow up evaluation of the software using social science methods. It addresses a mixed audience and provides technical detail as expected in the mathematical and computational sciences.

Individuals with autism experience difficulties with social skills and can find understanding the nonverbal cues and social behaviours of other people challenging. This makes building friendships and other appropriate relationships difficult, which can lead to isolation, social anxiety and depression, impacting their overall wellbeing. Further, many individuals with autism report an affinity for technology and exhibit high technology usage patterns. Using virtual humans to teach social skills to children with autism harnesses this preference for technology and provides a tool that can support the development of social skill knowledge and behaviour, ultimately aiming to improve individuals' everyday wellbeing.

Existing research with children with autism suggests that autonomous (self-directed) virtual humans can be used successfully to improve language skills (Bosseler and Massaro 2003) and authorable (researcher controlled) virtual humans can be used to improve social skills (Tartaro and Cassell 2006). The original contribution of this research is to combine these ideas and investigate the use of autonomous virtual humans for teaching basic social skills in the areas of greeting, conversation skills, and listening and turn taking.

The software in this research features three virtual humans who guide the learner through tasks and model social scenarios: a teacher, a peer with strong social skills, and a peer with developing social skills. Thirty one participants were assigned to either the control or experimental group using a matched pairs approach then asked to use the software for 10-15 minutes per day, 3-5 days per week for three weeks, with data collected before software use, at the end of the three week period, and again two and four months later. The data collected included a content quiz testing participant knowledge, the Vineland-II evaluating social behaviours as observed by caregivers, and questionnaires assessing participants' prior experience and expectations, and participant and caregiver perceptions following software use. The software itself also automatically recorded log data reflecting participant interaction with the system.

The Social Tutor was generally well-received by participants and caregivers, although more game-like elements and some adjustments to the virtual humans themselves and the lesson sequencing algorithm were requested for future development. Evaluation data likewise indicated positive trends, with a clear difference between performance of the experimental (social content) group and the control (non-social content) group in the content quiz. Vineland data was less clear with both groups performing similarly overall, although some encouraging trends were seen. Future work should focus on further generalisation support with the aim of converting the changes in knowledge demonstrated by participants to changes in real-world behaviour. A follow up evaluation with a larger sample size and longer software use period would also be beneficial to ascertain if any of the apparent promising trends observed in the data eventuate to significant outcomes.

DECLARATION

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Marissa Kate Bond (nee Milne)

Signed.....

Date.....

ACKNOWLEDGEMENTS

It has been a long journey and I am incredibly grateful to everyone who has supported me along the way. To my supervisors, Professor David Powers, Associate Professor Pammi Raghavendra and Dr Richard Leibbrandt, thank you for believing in me. This would not have been possible without your guidance and encouragement, and I truly appreciate all the time you have put into supporting me. I count myself very fortunate to have had the opportunity to learn from such skilled and passionate researchers.

To my colleagues and friends at Flinders University, thank you for your support throughout the years. You have helped me navigate the technical aspects of this journey, been there when I needed advice not only on completing the PhD itself but also on achieving some semblance of work-life balance. You have motivated me and helped me see the light at the end of the tunnel. More than anything, it has been a pleasure studying and working surrounded by such inspiring people whose company I enjoy so much.

To the authors of the social skills curricula implemented in the Social Tutor software, thank you for allowing me to draw on your hard work. Having access to the evidence-based, engaging content that you developed undoubtedly contributed to the positive outcomes evaluation participants achieved.

A huge thank you to the young people and families who volunteered to be involved in the evaluation of the Social Tutor software. Without you the evaluation and all the valuable insight it has provided would not have been possible. Meeting each and every one of you was an absolute highlight on this journey, and I thank you for inviting me into your homes. Your belief in the idea behind this research and your enthusiasm to see it continue into the future has been instrumental in driving this work to completion.

To all of my friends and family, thank you for being so understanding, especially at those crunch times when I fell off the face of the Earth for months on end. Thank you for your support, laughter, patience and love.

To my parents, thank you for encouraging me when things were tough, for being the voice of reason, and most of all for believing with me. Dad, thank you for being there when I needed someone to bounce coding problems around with. Mum, thank you for being so patient and letting me talk at you for hours on end when I had things to get off my chest. It's exactly what I needed. I love you both so much.

Finally, to my husband Allan and daughter Sage. I adore you both more than words can express. Allan, thank you for always taking care of me, especially making sure I was sufficiently fed during those intense periods of coding - I know I become totally unresponsive to the outside world - and for taking on all of the domestic duties over these last few months so I could focus on my thesis. I couldn't have done this without your support, thank you for always having my back and loving me even in my least lovable moments.

Please note: an editor has not been used in the construction of this thesis.

RELATED PUBLICATIONS

Conference Papers

Milne, M., Leibbrandt, R., Raghavendra, P., Luerssen, M., Lewis, T. and Powers, D. (2013). Lesson Authoring System For Creating Interactive Activities Involving Virtual Humans: Thinking Head Whiteboard. In H Hagras & V Loia, ed. 2013 IEEE Symposium on Intelligent Agent (IA) Piscataway, USA: IEEE. IEEE Symposium Series on Computational Intelligence. Singapore. Apr 2013, pp. 13-20. [10.1109/IA.2013.6595184] Google Scholar citations: 2

Milne, M.K., Luerssen, M.H., Lewis, T.W., Leibbrandt, R.E. and Powers, D.M. (2010). Development of a Virtual Agent Based Social Tutor for Children with Autism Spectrum Disorders. In Proceedings of the International Joint Conference on Neural Networks 2010. USA: IEEE. WCCI 2010 IEEE World Congress on Computational Intelligence. Barcelona, Spain. Jul 2010, pp. 1555-1563. [10.1109/IJCNN.2010.5596584] Google Scholar citations: 23

Milne, M.K., Powers, D.M. and Leibbrandt, R.E. (2009). Development of a software-based social tutor for children with autism spectrum disorders. In Jesper Kjeldskov, Jeni Paay and Stephen Viller, ed. Conference proceedings : Australian Conference on Computer Human Interaction (OZCHI 2009) Melbourne 2009: CHISIG. Australian Conference on Computer Human Interaction. Melbourne. Nov 2009, pp. 265-268. [10.1145/1738826.1738870] Google Scholar citations: 6

Book Chapters

Milne, M.K., Luerssen M.H., Lewis, T.W., Leibbrandt, R.E. and Powers, D.M. (2011). Designing and Evaluating Interactive Agents as Social Skills Tutors for Children with Autism Spectrum Disorder. In Diana Perez-Marin and Ismael Pascual-Nieto, ed. Conversational Agents and Natural Language Interaction: Techniques and Effective Practices. Hershey, USA: IGI Global, pp. 23-48. [10.4018/978-1-60960-617-6.ch002] Google Scholar citations: 8

Milne, M.K., Luerssen, M.H., Leibbrandt, R.E., Lewis, T.W. and Powers, D.M. (2011). Embodied Conversational Agents for Education in Autism. In Mohammad-Reza Mohammadi, ed. A Comprehensive Book on Autism Spectrum Disorders. 1st ed. Rijeka, Croatia: InTech, pp. 387-412. [10.5772/18688] Google Scholar citations: 3

LIST OF TABLES

Table 1: Severity levels for autism spectrum disorder in DSM-5 8
Table 2: Categories of generalisation 13
Table 3: Examples of existing interventions for children with autism 16
Table 4: Examples of existing tutoring software 32
Table 5: Overview of activity structure indicating number of lessons in each category 67
Table 6: Current, original and sample FaceGen models 72
Table 7: Placeholders and their purposes 76
Table 8: Lesson activity types available in the Thinking Head Whiteboard software 79
Table 9: Data collection schedule 94
Table 10: Extracted software log file data summary 102
Table 11: Calculated software log file data summary 102
Table 12: Summary of lesson interaction data 103
Table 13: Sequence of objectives as taken by experimental group participants 104
Table 14: Experimental participants grouped into response levels 107
Table 15: Summary of correctness, accuracy and duration data from pre-test and post-test content quiz 109
Table 16: Mean change in content quiz correctness scores from pre-test to post-test 112
Table 17: Summary of mean change in duration in seconds by response subgroup 114
Table 18: Comparison of change in content quiz question answering duration by topic 115
Table 19: Change in content quiz duration in seconds from pre- to post-test by questions and group 116
Table 20: Change in content quiz duration from pre- to post-test by question and response subgroups 116
Table 22: Summary of mean duration, correctness and accuracy data by age bucket 118
Table 21: Summary of type and complexity of content quiz questions 118
Table 23: Comparison of correctness, accuracy and duration data by age bucket 119
Table 24: Summary of mean change in v-sum and standard scores for Vineland-II domains and overall 122
Table 25: Summary of mean change in v-scale scores for Vineland-II subdomains 122
Table 26: Summary of Vineland-II domain change from pre-test to post-test by response subgroups 123
Table 27: Summary of Vineland-II subdomain change from pre-test to post-test by response subgroup 123
Table 28: Vineland-II items of marginal and statistical significance for change from pre-test to post-test 124
Table 29: Comparison of pre-identified and non-selected Vineland-II items 125
Table 30: Scoring for adaptive levels of domains in Vineland-II Caregiver Survey 125
Table 31: Summary of participant data where a change in adaptive level occurred from pre- to post-test 126
Table 32: Summary of demographic data for Vineland-II responses 127
Table 33: Summary of longitudinal content quiz completion by group 129
Table 34: Summary of longitudinal Vineland-II completion by group 129
Table 35: Summary of longitudinal data for content quiz by group 131
Table 36: Summary of longitudinal data for Vineland-II by group 132

Table 37: Post-test questionnaire completion rates 137
Table 38: Mean ratings of 'time spent on lessons' and 'lesson difficulty' 140
Table 39: Summary of strength-focussed comments from open-ended post-test survey 141
Table 40: Summary of challenge-focussed comments from open-ended post-test survey 142
Table 41: Summary of suggestions from post-test survey responses 144
Table 42: Significance values for experimental group whole-quiz correctness data 209
Table 43: Significance values for control group whole-quiz correctness data 209
Table 44: Comparison of pre-test and post-test content quiz correctness scores 210
Table 45: Comparison of change in content quiz question answering duration by topic 211
Table 46: Change in content quiz duration in seconds from pre- to post-test by questions and group 211
Table 47: Change in content quiz duration from pre-test to post-test by question and response subgroups . 212
Table 48: Summary of raw count and percentage pre-test questionnaire data 216
Table 49: Mean responses to post-test questionnaire Likert-style items 216

LIST OF FIGURES

Figure 1: Social Tutor login screen 62
Figure 2: PALS characters (Cooper et al. 2003a) 65
Figure 3: Screenshot of the Whiteboard topic selection screen for a new account 66
Figure 4: Social Tutor architecture overview 69
Figure 5: An extract from a user's summary XML file 70
Figure 6: Head X can model many facial expressions 71
Figure 7: Directory structure of the Social Tutoring software 73
Figure 8: A basic XML lesson definition file 74
Figure 9: The output of the definition file shown in Figure 8 74
Figure 10: Condensed screenshot of the View Progress screen 81
Figure 11: A sample curriculum definition XML file 83
Figure 12: Screenshot of a homework activity 87
Figure 13: Student workflow 96
Figure 14: Control group topic options 98
Figure 15: Example of a control group maze activity 99
Figure 16: Box plots showing distribution of participant ages by group 106
Figure 17: Quantile-comparison plots showing pre-test to post-test change in content quiz correctness. 106
Figure 18: Change in content quiz correctness score from pre-test to post-test by group 107
Figure 19: Distribution of correctness scores for the experimental and control groups 108
Figure 20: Time spent on content quiz questions by data collection point and group 114
Figure 21: Box plots displaying variation in Vineland-II scores by gender 128
Figure 22: Mean correctness score at each data collection period by group and subgroup 131
Figure 23: Mean v-sum scores at each data collection point by group 133
Figure 24: Mean v-sum scores at each data collection point by response subgroup 134
Figure 25: Visual summary of numerical pre-test questionnaire items 136
Figure 26: Participant purposes for computer use 136
Figure 27: Mean responses Likert-style post-test questionnaire items by group 137
Figure 28: Frequency of experimental group responses to Likert-style post-test questionnaire items 138
Figure 29: Frequency of control group responses to Likert-style post-test questionnaire items 138
Figure 30: Ratings of 'time spent on lessons' by group 140
Figure 31: Ratings of 'difficulty of lessons' by group 140
Figure 32: Example of a self-model created using FaceGen, with a photo of the target on the right. 178

CHAPTER 1. INTRODUCTION

Autism spectrum disorder (ASD) is characterised by deficits in communication, social skills and a tendency towards repetitive behaviours, making social interactions extremely challenging for those on the spectrum (Kroncke et al. 2016). Individuals with autism have difficulty not only recognising and interpreting facial expressions and body language, but also understanding the motivations and feelings of others, skills that often appear to come naturally to their neurotypical peers (Kroncke et al. 2016). Putnam and Chong (2008) conducted a survey asking 12 adults with autism and 114 family members of children with autism what they wanted from software and technologies designed for individuals on the spectrum, and interestingly found that children with autism often reported feeling more comfortable interacting with computers and technology than their peers, a finding that has also been reported elsewhere (Baron-Cohen et al. 2009). A more recent survey of 172 parents with neurotypical children and 139 parents of children with autism found that those with autism had higher daily usage patterns of certain electronics, Internet and video games than their neurotypical peers (MacMullin et al. 2016). Further, Parsons et al. (2016) developed a mobile application to ask individuals with autism and those who support them "if there was one new technology to help people with autism, what would it be?" and found that, second only to academic skills, there were many requests for technology to support individuals with their social communication and interaction skills.

Work with authorable virtual agents for social skills education and autonomous virtual agents for language learning have both demonstrated positive outcomes for children with autism (Bosseler and Massaro 2003, Tartaro and Cassell 2006). Authorable agents are defined as those requiring external input to interact with the user, such as a researcher observing the interaction and controlling the virtual human like a puppet, while autonomous virtual humans are fully self-contained and require no such external input. Inspiration for the current work was gained from combining these existing studies, with the aim being to develop an autonomous virtual agent for social skills learning. The original contribution of the research presented in this thesis is to design, implement and evaluate the effectiveness of a computer program that can be deployed on existing desktop computers in both home and school environments and be used by children with autism to support their social skill development, in particular focussing on skills that are necessary and appropriate for individuals in mainstream schooling.

Given the multidisciplinary nature of the current research, it should be noted that this work is based in the domain of computer science and draws on social science methods for evaluating the software developed. This thesis aims to address a mixed audience, providing background information appropriate for those approaching it from both a computer science background and a social sciences background, and provides technical detail to the level expected in the domain of mathematical and computational science.

It is hoped that a computer program providing a virtual-agent based social skills tutor will capitalise on the affinity for technology that individuals with autism have reported, while working towards improved social proficiency. This approach is aimed to provide a motivating, judgement-free environment for developing

social skills in, while offering a tool that can be used independently by the individual, consequently relieving pressure from carers, educators and therapists working with the learner. The goal is that the software developed here will target more general social skills and concepts, allowing educators and therapists to focus their efforts on specific and nuanced areas of need that the learner has. The software is not intended to reduce or replace real social interaction in any sense, but instead, given the nature of the virtual humans the software is centred around, to be used as a stepping stone between theoretical knowledge and real-world practice.

The first aim of the research presented in this thesis was to design and implement a Social Tutor software program to teach the basic skills of greeting others, listening and turn taking, and beginning, ending and maintaining conversations. To ensure that the software adhered to current best practice and drew on existing knowledge and experiences in the fields of social skills education, technology for use with autism, and implementation of virtual agents for education more generally, a literature review was conducted into these areas and is presented in this thesis. From this review a framework for the Social Tutor was designed that included mechanisms for automatically assessing student knowledge, a dynamic lesson sequencing system based on this automated assessment, and methods for providing appropriate feedback and progress tracking. To ensure that the software could respond to student needs rather than taking a 'one size fits all' approach, it was determined that an effective and achievable method would be to create a large number of very short, self-contained lesson activities that could be mixed and matched according to the learner needs at any time.

From the literature review, the importance of inclusive design was also emphasised. This is particularly relevant given the sensory issues that many individuals with autism also experience, which can range from abnormally low sensitivity which leads to sensory-seeking behaviours, to abnormally high sensitivity which leads to sensory-avoidance behaviours, and can apply to any of the senses (Robertson and Simmons 2013). To ensure that the software was suitable for the widest range of individuals possible, the interface was deliberately kept clean with a basic colour palate, no background music, and a simple layout. Images were used extensively but were also used mindfully, typically to indicate the purpose of buttons and interactive elements. Combining sensory difficulties with the communication difficulties that can also come hand-in-hand with autism, it is not uncommon to have a diverse range of preferred interaction modes within a population of individuals on the spectrum. For example, some may be strong readers and prefer text-based instruction, while others may not and may prefer spoken cues, while others again may be sensitive to sound and prefer image-based cues. Thus, by providing simple image and text-based cues, along with the ability to have the virtual characters read instructions aloud, it is hoped that most learners will be supported enough to use the software independently.

In the same vein, many individuals with autism experience poor fine-motor skills which can make typing and using a mouse challenging (Posserud et al. 2016). While tablets and other mobile computing devices were not in widespread use when the Social Tutor implementation began and thus the software was designed for use on a standard desktop or laptop computer, the Social Tutor was nevertheless designed to be touch-screen compatible and thus has the potential to be easily ported to these mediums. By designing it this way the

interaction modes were kept simple, focussing on drag and drop and simple button clicking rather than typing or complex inputs, and minimising the impact that fine motor skill difficulties would have on software interaction. A number of other potentially assistive technologies were investigated for inclusion in the Social Tutor, particularly speech recognition and gesture recognition, however at the time that implementation began the benefits of these technologies in terms of accessibility were outweighed by their lack of robustness to error and their reliance on specialised equipment. The technology in these areas has since improved and the equipment required is now often integrated into mobile and standard computing devices, so they may be worth consideration in future iterations of this and related software.

In addition to the features already discussed, the Social Tutor software was also designed with customisation in mind, and as such lesson activities are written in a basic XML-style language and provide sensible default behaviours. The goal of this was to make it relatively easy for non-programmers, for example caregivers and educators, to both create their own lesson activities and modify existing ones. By modifying existing lessons and creating their own, educators and caregivers can ensure that the content of activities presented in the Social Tutor software aligns well with the educational and intervention activities that learners are partaking in through other channels, and that content is as engaging and relevant as it can be. For example if a student is learning a particular greeting in an intervention at school, then the Social Tutor lessons on that topic could be modified to ensure that their target greeting sentence is included prominently. Another example could be that the images in lessons could be replaced with custom ones of the learners' own friends or favourite cartoon characters, increasing engagement and relevance. While this feature has a lot of potential, it was determined that for the initial evaluation of the Social Tutor consistency across individuals was paramount, and as such lesson customisation was not used in the evaluation presented in this thesis, instead all participants received the lesson content developed by the researcher with no individual personalisation.

Developing and validating a social skills curriculum for implementation in the Social Tutor software was determined to be too large an undertaking to fit within the scope of the current research, essentially consisting of a separate research project in and of itself, and thus the decision was made to identify existing social skills curricula that were evidence-based, contained content that aligned with the educational aims of the research, and that were presented in a way that could be delivered successfully using a software environment. Ultimately three evidence-based curricula were chosen for inclusion, namely the 'Playing and Learning to Socialise' (PALS) curriculum (Cooper et al. 2003a), the 'Skillstreaming' curriculum (McGinnis and Goldstein 2012) and the 'Social Decision Making / Social Problem Solving' (SDM/SPS) curriculum (Elias and Butler 2005), with each complimenting and building on the content of the others. An investigation into assessment tools and automated techniques that could be applied to determine student understanding and adapt dynamically to student needs was also conducted to inform the development of the dynamic lesson sequencing system included in the Social Tutor software.

Human peer tutoring is a well-established evidence-based practice that has demonstrated effectiveness for learners both with and without disabilities and across a wide range of ages, settings and topics (Bowman-

Perrott et al. 2013) and a technique that inspiration can be drawn from in development of the Social Tutor software. There are two main hypotheses regarding what makes human tutoring effective: the tutor action hypothesis that suggests it is primarily the tutor's actions that result in positive learning gains, and the student action hypothesis that suggests it is instead the student's ability to build connections between concepts and construct knowledge that leads to these gains. Seemingly in conflict with the tutor action hypothesis, research has been shown that even when tutors are only allowed to prompt students without providing explanations or feedback, the students learned effectively (Chi et al. 2001). However, in contrast to the student action hypothesis, it has been observed that learners rarely make effective use of their tutor beyond confirming that they are taking the right steps or to confirm a piece of information (VanLehn 2011). A tutor action hypothesis that aligns with these observations and appears more promising is that it is the tutor's ability to identify when the student is about to make a mistake and provide immediate feedback and prompting to get them back on the right pathway. Research suggests that this approach helps minimise frustration, stalling, confusion and the need for backtracking and re-doing work (Chi et al. 2001, Bowman-Perrott et al. 2013).

Closely related to this, the concept of scaffolding is also core to many effective educational strategies. Scaffolding can take many forms and broadly refers to the concept of providing a learner with the guidance and structure they need to achieve a learning outcome (van de Pol et al. 2015). In a practical sense this often manifests as taking a large, complex task and breaking it down into basic subtasks. Learners can then master each of these basic subtasks in turn, ultimately building up to being able to understand and perform the complex target task (Jackson et al. 2010a, van de Pol et al. 2015). By ensuring that the Social Tutor performs ongoing assessment of student knowledge so that difficulties are detected and remedied quickly, and structuring activities in a scaffolded manner, some of the benefits of human tutoring techniques can be realised in the Social Tutor software. Interestingly, most human tutors lack formal training, and yet the technique remains effective. This is encouraging and suggests that even an imperfect virtual tutor should be capable of benefitting learners.

While an imperfect virtual tutor can still be beneficial to learners, in any autonomous system where judgements are made about the correctness of an answer or appropriateness of a behaviour, care must be taken to avoid inadvertently reinforcing undesirable responses. In the case of the Social Tutor, this means that in cases where uncertainty exists, the preferred approach is for the virtual humans to explain the desirable response, typically without commenting on the students' original input. In this way the system ensures that learners are presented with the information they need without accidentally reinforcing less preferred behaviours or suggesting that a correct choice of the students' is incorrect.

Two widely acknowledged issues for any intervention intended for individuals on the autism spectrum are those of generalisation and maintenance. Generalisation refers to the ability to apply knowledge and skills learned in one context to another, with 'near transfer' referring to applying skills to tasks that are similar to the original learning task, and 'far transfer' referring to applying skills to more distinct contexts such as real-

world situations and novel environments (Whyte et al. 2015). It has often been found that individuals with autism become quite good at 'doing the intervention' but then fail to use their new skills in situations outside of the intervention environment. Maintenance is another area that can be challenging, with learners forgetting what they have learned in the intervention over time unless explicit reinforcement is provided.

Despite being known issues, recent reviews have shown that both generalisation and maintenance are still often overlooked, not measured, or treated as an afterthought in existing research (Neely et al. 2016). During development of the Social Tutor software a number of mechanisms were put in place to help support generalisation, such as ensuring a diverse set of visual supports were used, presenting the same concepts in a variety of different activity types, and ensuring activities were closely related to real-world situations that learners are likely to encounter.

Following completion of the Social Tutor software an evaluation was conducted into its effectiveness. During the software evaluation period, participants were asked to use the software for 10-15 minutes a day, 3-5 days a week, for three weeks. The aims of the evaluation were to determine if changes in knowledge and behaviour occurred due to interaction with the Social Tutor, and whether these changes were maintained after use of the Social Tutor ended. An investigation into participant and caregiver perceptions of the software was also carried out to inform future development of this and related software. To ensure any changes to knowledge or behaviour that occurred following software use could be attributed to use of the Social Tutor, both an experimental group and a control group were included. The experimental group received content designed to explicitly teach social skills, while the control group received simple non-social game-like maze activities instead.

Again, a review of existing literature was performed prior to this process to determine the most appropriate evaluation methodology to employ and the tools necessary to measure data relating to knowledge and behaviour. The outcomes are presented in this thesis. Ultimately a bespoke content quiz was developed to determine student knowledge, and this was supported by use of the Vineland Adaptive Behaviour Scales (Vineland-II) to measure changes in behaviour as observed by participants' caregivers. These measures were administered prior to participants' use of the software, immediately at the end of the software use period, and then again both two and four months after the software use period to specifically address the issue of maintenance. Along with these measures, participants and caregivers were asked to complete questionnaires to provide insight into their expectations of the software prior to use and their experiences with the Social Tutor after the software use period ended. The Social Tutor software additionally recorded automatic log data to provide insight into participant usage patterns across the intervention period.

Results of the software evaluation indicate that the Social Tutor was generally well received by participants and caregivers, with caregivers being particularly supportive of the aims of the study. Improvements to virtual human voices, the lesson sequencing system and increased gamification of educational content are all recommended to enhance future iterations of the Social Tutor. Data from the content quiz indicated that the experimental group made performance improvements from pre-test to post-test that were statistically

significant while the control group performed similarly at both time points, supporting the notion that social skills content explicitly taught by the Social Tutor was responsible for this change in participant knowledge. Vineland-II data indicated that in general the experimental and control groups performed similarly over the course of the study and thus no notable improvements in behaviour can be attributed to use of the Social Tutor alone. However, given the short three-week intervention period, this may simply indicate that more time with the software is required before the gains in knowledge of social skills translates into changes in social behaviour, or may indicate that the Vineland-II is not suitably sensitive to this change for this purpose.

As previously stated, the overall aim of this research is to design and implement evidence-based software for improving conversation-focussed social skills in children on the autism spectrum that utilises autonomous virtual humans, then evaluate this software for its effectiveness. The purpose of this thesis is to provide a description of the Social Tutor software and its features, including insight into the decisions that led to the current software design, then to present the procedure and findings of the evaluation of this software. In Chapter 2 the literature review covering both of these activities is presented, with the research aims of this study and how they are addressed provided in Chapter 3. The first aim of this research was to design and implement social tutoring software based on current research and best-practice guidelines. This software is unique in that it utilises virtual characters to assist children with autism to improve their social skills, a novel application of autonomous virtual humans. Following this, the primary aims of this study were to determine if broad scale changes in knowledge and behaviour occurred due to interaction with the software, and if these changes were maintained over a 4 month period. The study also investigated participant and caregiver perceptions of the Social Tutor software. Following initial data analysis a number of secondary questions were raised and further exploratory data analysis was conducted, in particular around the possible categorisation of users into high, average, and low response subgroups. Further exploratory analysis was conducted in an attempt to identify possible characteristics of users that would assist in recognising who would benefit most from the software, and to provide insight into which aspects of the software should be retained and which could be improved to lead to better educational outcomes and more positive user experiences. Technical details of the Social Tutor software are then offered in Chapter 4, with an explanation of the software evaluation that was conducted provided in Chapter 5. Following this, the data gathered from the software evaluation is presented in Chapter 6, with a detailed discussion of the implications of these findings in Chapter 7 and a thorough discussion of recommendations for future directions in Chapter 8. Finally, a reflection on the limitations and outcomes of the research as a whole is presented in Chapter 9.

CHAPTER 2. BACKGROUND AND LITERATURE REVIEW

To develop successful social skills tutoring software a wide range of elements must be drawn together, including an understanding of autism itself, educational theories behind learning, existing social skills interventions, and knowledge of current computer science techniques and technologies relating to the development of virtual humans and computer-aided learning. The purpose of this chapter is to discuss these critical elements and provide an explanation for the choices made while developing and evaluating the Social Tutor software in the current research, in particular the selection of educational materials, assessment tools, and technical approaches incorporated into the software.

First, the nature of autism and the typical learning needs of individuals on the spectrum must be understood in order to inform the development of a tutor that behaves in a way that is tailored to their specific needs and is intuitive to use. For individuals with autism, ensuring that unnecessary sensory stimulation is avoided and available interaction modes within the software are supportive of the difficulties associated with autism is also important. Next, an investigation of existing social skills interventions, both traditional and technology-based, provides an understanding of the gaps and opportunities in the field and enables lessons to be learned from successful approaches and applied to the Social Tutor developed in the current research.

Given that the interaction style of the Social Tutor software is intended to echo that of personalised one-on-one tutoring, obtaining an understanding of human tutoring behaviour and the educational theories underpinning its widely acknowledged success is vital (Graesser et al. 1999, Chi et al. 2001, VanLehn 2011). Following this, an investigation into the ability of virtual tutors to demonstrate successful outcomes on par with human tutors and the unique benefits that virtual tutors offer is provided, followed by discussion of a select sample of noteworthy virtual tutors from a range of application areas. Next, the unique considerations involved in using virtual tutors as social skills interventions for children with autism are addressed.

A comparison of existing social skills curricula is then presented with an eye towards selecting appropriate content for the Social Tutor software, and with strategies for encouraging both deep learning and generalisation to real-world contexts being a priority. To this end, the aim was to identify content that is evidence-based, meets the learning goals of the program being developed, and can be incorporated into a software-based environment successfully. Following this, assessment strategies and tools are investigated with two goals in mind. First, procedures and techniques that can be incorporated into the software as part of the ongoing automated assessment and dynamic lesson sequencing system are investigated. Finally, the behavioural assessment tool most appropriate for use in the evaluation of the software itself is determined.

2.1 About Autism

Autism spectrum disorder is a pervasive developmental disorder characterised by impairment in social communication skills and the presence of restricted, repetitive patterns of interest and behaviour. Autism affects individuals from very early in their lives, with diagnosis often happening at around two to three years

old. Being a spectrum disorder, individuals with autism can range from having limited or no functional speech and often having a low IQ, to having an IQ in the normal range or above and displaying functional speech (American Psychiatric Association 2013).

In 2013 the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) was released, bringing with it significant changes to the way autism spectrum disorders are classified (American Psychiatric Association 2013). Under the previous edition (DSM-IV-TR), autism spectrum disorder was considered to cover several conditions including high-functioning and low-functioning classic autism, Asperger Syndrome (AS) and Pervasive Developmental Order Not Otherwise Specified (PDD-NOS) (American Psychiatric Association 2000). Under the newer DSM-5, the labels of Asperger Syndrome and PDD-NOS are no longer present, however it is expected that most individuals who would previously have fallen under these diagnoses will now be covered by the broader autism definition (Joshi 2013). Under DSM-5 autism now has a severity ranking, details of which can be seen in Table 1.

Table 1: Severity levels for autism spectrum disorder in DSM-5

Severity level	Social communication	Restricted, repetitive behaviours
Level 3 "Requiring very substantial support"	Severe deficits in verbal and nonverbal social communication skills cause severe impairments in functioning, very limited initiation of social interactions, and minimal response to social overtures from others. For example, a person with few words of intelligible speech who rarely initiates interaction and, when he or she does, makes unusual approaches to meet needs only and responds to only very direct social approaches	Inflexibility of behaviour, extreme difficulty coping with change, or other restricted/repetitive behaviors markedly interfere with functioning in all spheres. Great distress/difficulty changing focus or action.
Level 2 "Requiring substantial support"	Marked deficits in verbal and nonverbal social communication skills; social impairments apparent even with supports in place; limited initiation of social interactions; and reduced or abnormal responses to social overtures from others. For example, a person who speaks simple sentences, whose interaction is limited to narrow special interests, and who has markedly odd nonverbal communication.	Inflexibility of behavior, difficulty coping with change, or other restricted/repetitive behaviors appear frequently enough to be obvious to the casual observer and interfere with functioning in a variety of contexts. Distress and/or difficulty changing focus or action.
Level 1 "Requiring support"	Without supports in place, deficits in social communication cause noticeable impairments. Difficulty initiating social interactions, and clear examples of atypical or unsuccessful response to social overtures of others. May appear to have decreased interest in social interactions. For example, a person who is able to speak in full sentences and engages in communication but whose to-and-fro conversation with others fails, and whose attempts to make friends are odd and typically unsuccessful.	Inflexibility of behavior causes significant interference with functioning in one or more contexts. Difficulty switching between activities. Problems of organization and planning hamper independence.

From Autism Speaks (2013)

The software developed in this research is designed to be most appropriate for individuals with a diagnosis of level 1 autism and in some cases may be beneficial to those with a diagnosis at level 2, but is not anticipated to be appropriate for those with a diagnosis of level 3 severity as the level of prerequisite language skill required is likely to be too advanced.

For this research, individuals with an existing diagnosis of autism or Asperger Syndrome were included in the evaluation portion of the study. While Asperger Syndrome is no longer formally recognised, this criteria was included to ensure that individuals who have not yet received a recent diagnosis under the new criteria could still participate, particularly as the software was in development prior to the release of DSM-5 and originally targeted those with Asperger Syndrome and those with high functioning autism spectrum disorder.

In the remainder of this section the core impairments of autism and the various challenges associated with it are discussed to provide an understanding of what children on the spectrum experience and how the software developed for this research can best support their social skill development.

2.1.1 Social Skills Deficit

Social skills can be defined as "specific, identifiable skills that result in socially competent behaviour" (Wilkins 2010). Social interactions are a major challenge for individuals with autism, as they not only struggle to use nonverbal communication appropriately themselves but also find it challenging to understand the nonverbal cues of others (Rapin and Tuchman 2008, American Psychiatric Association 2013). For individuals who do not have autism, commonly referred to as neurotypical, understanding of facial expressions, body language and eye gaze is developed through their everyday experiences, however for children with autism these skills need to be explicitly taught, and one possible explanation for this is that children with autism fail to demonstrate typical joint attention with interaction partners and this limits their opportunities to learn related social skills (Davis et al. 2014, Rice et al. 2016). The failure to develop a typical 'theory of mind' has also been suggested as a contributing factor to the difficulties that individuals with autism face. 'Theory of mind' refers to the ability to understand that other people have thoughts and feelings that are different from your own, and without this understanding, predicting other people's motivations and how they will respond to situations is a major difficulty (Leslie 1987).

These difficulties understanding other people's motivations and interacting with them means that individuals with autism often fail to develop friendships and other social relationships appropriate to their developmental level. This can lead to many serious impacts on their overall wellbeing, for example loneliness, isolation, social anxiety and depression. Bauminger and Kasari (2000) note that individuals with autism are typically self-conscious of their differences in social functioning and this feeds into their feelings of loneliness and their difficulties developing friendships. A recent study by Jackson and Dritschel (2016) further indicates that difficulties with social problem solving contribute significantly to vulnerability for depression. Jackson and Dritschel (2016) also suggest that the increased social demands during adolescence and early adulthood compound the risk, making depression a serious issue for individuals on the spectrum. In addition to anxiety

and depression, because of their social difficulties adolescents and adults with autism are at increased risk for a variety of secondary negative outcomes including isolation, rejection, bullying, low-self esteem, school dropout and unemployment (Mitchel et al. 2010, Schohl et al. 2014). Clearly difficulties with social interaction can lead to significant and life-long impacts on individuals' mental health and wellbeing, and a tool that can help mitigate this has the potential to positively impact quality of life in many areas.

Related learning areas identified as being of particular need for individuals on the autism spectrum include developing conversation skills such as initiating and shifting conversation topics, social skills such as reading and responding to nonverbal cues, and emotional skills such as regulating and expressing emotions and developing coping strategies for stressful situations (Rubin 2007).

2.1.2 Language and Communication Deficits

Typically language skill development is delayed in individuals with autism, with some lower functioning individuals failing to develop spoken language skills at all, and with several studies suggesting that higher IQ and language skills at a young age can be predictive of better adult outcomes (Magiati et al. 2014, Pickles et al. 2014). Prior to DSM-5, there was some suggestion that individuals with Asperger Syndrome have a higher verbal IQ than nonverbal IQ, while individuals with autism display the reverse, having a higher nonverbal IQ than verbal IQ (Rubin 2007). Under the new criteria, it is recognised that verbal and nonverbal IQ both develop on a unique basis for each individual.

Individuals on the autism spectrum often display repetitive use of language including echolalia, and interpret language very literally, which can result in serious misunderstandings and difficulties when metaphors and sarcasm are used (American Psychiatric Association 2013). The use of visual supports and provision of clear, to-the-point instructions is recommended to assist those with lower verbal IQ to understand what they are being told, while providing clear spoken instructions can assist those with higher verbal IQ, such as that previously associated with a diagnosis of Asperger Syndrome (Grandin 1995, Myles et al. 2005). Simple written language appears to be a beneficial medium for those with both higher and lower verbal IQ. Communication is not only a challenge in itself, but also compounds the difficulty individuals on the spectrum experience with social skills (Rapin and Tuchman 2008).

2.1.3 Repetitive Interests and Behaviours

Individuals with autism often display repetitive behaviours, often sensory related and sometimes stereotyped such as rocking or hand flapping, and it is common for them to fixate on a single interest or small range of restricted interests rather than the broader range of interests commonly seen in neurotypical individuals (Kohls et al. 2014, Kirby et al. 2017). Due to their restricted and repetitive patterns of behaviour, individuals with autism exhibit a very strong preference for sameness, which can result in high levels of anxiety and stress when confronted with a change to their routine or environment. Caregivers of individuals with autism often report that these restricted and repetitive behaviours and interests are one of the most challenging

aspects of their everyday life and can negatively impact not only the individual with autism themselves but also the well-being of their family (Kohls et al. 2014).

One technique that can greatly ease transitions between activities is to prepare the individual beforehand by explaining what is going to happen and what is expected of them, and providing visual supports to assist the explanation (Myles et al. 2005, Pierce et al. 2013, Knight et al. 2015). It has been said that individuals with autism lack imagination; however, in reality they can exhibit great imagination although it is often non-social in nature or within the context of their special interest (Shaughnessy and Trimingham 2016, Ten Eycke and Müller 2016). In the past, numerous interventions have focussed on attempting to stop these intense fixations, with little success. More recently the focus has instead turned more towards harnessing this special interest to the learner's advantage and connecting learning activities to it in a variety of ways, for example an individual interested in trains might count trains in mathematics lessons, read books about trains in English lessons and paint or draw trains in art (Grandin 1995, Myles et al. 2005, Gunn and Delafield-Butt 2016). This has the added benefit of increasing enjoyment for the individual, and a positive mood has been shown to broaden attention and creative thinking, both of which are beneficial to learning (Seligman et al. 2009).

Connected to the notion of rigid behaviours and thoughts, individuals often need new concepts explicitly taught to them step by step and can find it very challenging to adapt their skills to novel situations (Rapin and Tuchman 2008). When developing new concepts, it is important to take this into consideration and teach in a way that best suits the individual learner while providing opportunities to practice the skill in a variety of contexts including a range of environments, using different materials, and with different interaction partners.

2.1.4 Co-morbid Conditions

Individuals on the autism spectrum often experience additional challenges due to conditions that are co-morbid with autism. Such conditions include sensory difficulties, poor fine and gross motor skills, auditory processing impairments, anxiety disorders, seizures, and intellectual disability, amongst other difficulties (Robertson and Simmons 2013, Posserud et al. 2016). When developing interventions for high functioning individuals, the difficulties with motor skills and sensory tolerance are particularly important to bear in mind (Grandin 1995, Robertson and Simmons 2013). Sensory tolerance can range from very low to very high, can affect any number of the senses, and sensory integration dysfunction can also be involved (Robertson and Simmons 2013). Some individuals find that stressful situations exacerbate their sensory challenges, and they can even experience changes to their tolerance levels over time (Grandin 1995). High sensory tolerance can lead to sensory seeking behaviours, for example high tolerance in the tactile sense may involve the individual seeking out particular textures to feel, and low tolerance can lead to avoidance behaviours, for example in the auditory sense the individual may become distressed by particular sounds such as the buzzing of a refrigerator (Robertson and Simmons 2013). Self-stimulatory behaviour, or ‘stimming’, such as hand-flapping and rocking can often be sensory seeking in function and comforting to the individual (American Psychiatric Association 2000, American Psychiatric Association 2013).

A particularly important example of a sensory intolerance which is commonly reported by those on the spectrum occurs when they are confronted with looking at another human's face. Many individuals with autism find that making eye contact is uncomfortable, even overwhelming (Jones et al. 2003). Existing research has found that when shown photographs of faces, individuals with autism display heightened activation in the regions of the brain responsible for processing threatening social and emotional cues when compared to their neurotypical peers (Dalton et al. 2005) particularly when the stimuli appears to be moving directly towards them (Crawford et al. 2016). This provides some insight into the challenges individuals with autism face when engaging in social interactions, and assists in guiding the development of the Social Tutor user interface and included learning materials.

2.1.5 Generalisation and Maintenance

Generalisation of skills to novel contexts and maintenance of those skills over time have been acknowledged as two of the most challenging aspects of developing interventions of any sort for individuals on the autism spectrum, and have been for decades (McCleery 2015, Radley et al. 2015, Neely et al. 2016). McCleery (2015) comments that the causes behind this are likely two-fold. First, some of the trouble faced transferring skills to new contexts is likely related to the core difficulties associated with autism itself, in particular rigidity of routines and the need for sameness, along with difficulties with communication and social interaction. Second, the issue may also be related to the particular strategies that have been used with this population failing to meet the needs of the learners, for example traditional approaches have focussed on replicating particular behaviours or vocabulary but without doing so in a way that is meaningful to the learner or fostering the understanding behind why the skill is important (McCleery 2015).

Generalisation can be categorised broadly into three groups, 'training effects' that occur only within the context of the training, 'near-transfer' that occurs when skills are used in similar computer-based tasks or standardised tests, and 'far-transfer' where skills are applied in real-world situations, such as when interacting with peers or in novel environments (Whyte et al. 2015). It has often been observed that individuals with autism will display marked improvements in the intervention environment, displaying training effects and even near-transfer generalisation, but then fail to achieve far-transfer generalisation to use their new skills in novel situations. Neely et al. (2016) recently conducted a review and meta-study of generalisation and maintenance in relation to functional living skills interventions and found that, despite these being known ongoing issues, many researchers are still failing to investigate whether their interventions lead to generalisation and maintenance of skills at all, and those who do often employ a "train and hope" strategy rather than explicitly including tactics to address these acknowledged issues.

In their seminal article, Stokes and Baer (1977) identified nine categories into which most techniques for assessing or programming generalisation can be categorised. A brief description of each category can be seen in Table 2. Stokes and Baer (1977) further classified generalisation into stimulus or response generalisation, and identified several dimensions generalisation can occur across, including person, setting or skills. Following their review of the existing literature at the time, Stokes and Baer (1977) found that over half the

Table 2: Categories of generalisation

Train and Hope	Do not explicitly teach for generalisation, but assess for it and hope that some occurs incidentally. The most common approach.
Sequential Modification	Assess for generalisation after an intervention, and if it is not present they implement procedures to produce it by systematically modifying every non-generalised condition.
Introduce to Natural Maintaining Contingencies	Link the desired behaviour to naturally occurring maintaining reinforcement in the participant's everyday environment. One of the most dependable approaches.
Train Sufficient Exemplars	Continue to teach exemplars of a target lesson until generalisation occurs sufficiently to satisfy the problem being addressed.
Train Loosely	Teach with relatively little control over the stimuli presented and the correct responses allowed, maximising sampling of relevant dimensions for transfer to other situations and other forms of the behaviour.
Use Indiscriminable Contingencies	Unpredictable schedules of reinforcement elicit high resistance to behaviour extinction and this approach can be applied to generalisation where reinforcement and non-reinforcement events are indiscriminable.
Program Common Stimuli	Where generalisation is expected to occur in the presence of appropriate stimuli, simply ensure that sufficient common stimuli are present in the training and generalisation environments. Suitable stimuli are often those with a function in the desired behaviour, e.g. a child's peers in a social skills intervention.
Mediate Generalization	Establish a response as part of the target skill that is likely to be shared across other problems, and will provide sufficient similarity between the existing skill and the new skill to lead to generalisation.
Train "To Generalize"	If generalisation is viewed as a skill in itself, it is seen that it can also be addressed and reinforced directly, e.g. explicitly telling learners to see a new situation as "the same" as a situation they have previously encountered. However, observation suggests that this is often unsuccessful.

existing research employed the "train and hope" approach where researchers probed for generalisation after the fact, rather than explicitly programming for it.

In the recent review by Neely et al. (2016) a similar trend was still present. Further, Neely et al. (2016) found that while assessing for maintenance was more common with almost all papers reviewed employing some type of data collection after the end of the intervention period, many of these only assessed for maintenance once, and very few extended that past three months post-intervention.

Stokes and Baer (1977) noted that in order to maximise generalisation, researchers need to program for generalisation and maintenance from the very beginning, and that a number of techniques could be employed to maximise outcomes. These techniques include teaching skills in the setting they would typically take place in as much and as soon as possible, linking this training to naturally occurring consequences, teaching across multiple stimuli and providing many opportunities to practice the target behaviour. They also noted the importance of collecting data across the intervention to allow researchers to assess the efficacy of their approach (Stokes and Baer 1977).

One recent example where researchers have paid particular attention to the difficulties with generalisation and directly employed strategies to combat them is the Superheroes Social Skills Program. Radley et al. (2015) have incorporated several of the generalisation technologies identified by Stokes and Osnes (1989),

directly harnessing the natural consequences of social skill use, diverse training through multiple channels, namely a social game and video models, and utilising functional behaviour mediators, in this case employing self-monitoring techniques. Consequently, evaluation of this program with two 11 and 12 year old boys indicated clear skill generalisation to novel settings and short-term maintenance of skills, although long-term maintenance was not monitored (Radley et al. 2015).

Another approach that has experienced success generalising social interaction skills is Say-Do Correspondence Training (Rosenberg et al. 2015). In this intervention, participants identified a peer they wished to talk to before recess, then received reinforcers after recess if they successfully talked to the chosen peer. They "say" who they are going to talk with, and then "do" the planned behaviour. This intervention takes place in the environment the skill is intended for use in, a key point highlighted by Stokes and Baer (1977), and is based on correspondence training which is an established technique for behavioural interventions with both adults and children, with and without disabilities (Rosenberg et al. 2015). Correspondence training is traditionally made up of a verbal "say" component intended to prompt the individual to undertake the nonverbal "do" component independently. In reality, the "say" and "do" components can both be verbal or nonverbal depending on the desired outcome and the particular learner's needs. It is often used in situations where the instructor can easily interact with the learner prior to the situation that the "do" component applies to, but either cannot easily access the learner during the context of the "do" component or it is desirable for the learner to engage in the "do" component independently. Rosenberg et al. (2015) suggests that for this reason, the Say-Do approach may be particularly suitable for situations where the instructor does not want to disrupt the natural and spontaneous social interactions between children, for example during recess at school. Rosenberg et al. (2015) found that the Say-Do approach led to increased social interactions during school recess times for all three children with autism in their case study, even after tangible reinforcements were removed. Existing work with the Say-Do approach in other contexts indicates that the target skills continue to be displayed in the target scenarios even once the Say-Do approach has ended.

Given that difficulties with both generalisation and long-term maintenance of new skills has been a serious and ongoing issue in the area of intervention research for autism, there is a clear need for all researchers in this field to explicitly consider how their intervention addresses and assesses these deficits. Wherever possible, researchers should overtly incorporate assessment of and programming for both generalisation and maintenance into their programs if the interventions they are developing are expected to provide real-world benefits for the individuals they target. This has been done in the current research, with the Social Tutor software employing several of the recommended techniques to maximise the chances of generalisation. The current research also addresses limitations in existing studies as identified by Neely et al. (2016) in that the evaluation process explicitly measures generalisation and maintenance at multiple time points until up to four months after software use has ended.

2.2 Existing Social Skills Interventions

To gain an understanding of the gaps and opportunities within the field and to inform the development of a successful software-based social skills tutor, an investigation of both traditional and technology-based social skills interventions was conducted. Technology-based interventions include hardware and software, with some incorporating virtual and augmented reality. Here, a brief overview of some of the more influential and novel interventions are given, and it is from these existing approaches that inspiration is drawn for the design and educational approach of the Social Tutor software developed here. As there is a wealth of fascinating, cutting edge interventions under development in this area, only a small selection of those that fit particularly well with the objectives of this research and that are well established in the field with experimental support, have practical elements that can be directly drawn from and implemented in the Social Tutor, or provide a unique insight to assist in this development are included here. Table 3 provides an overview of these highlights, with a more detailed discussion of each featured intervention following.

2.2.1 Traditional Interventions

Story and Comic Style interventions

One theory that attempts to explain some of the social difficulties that children with autism encounter suggests that they lack a fully developed 'Theory of Mind'. This means they have difficulty understanding that other people have separate thoughts and feelings to themselves. Carol Gray's Social Stories™ and Comic Strip Conversations (Bock et al. 2001, Gray 2001, Quirmbach et al. 2008) and the thought bubble approach used by Wellman et al (2002) aim to address this deficit.

One of the best known and most influential social skills interventions for children with autism are Carol Gray's Social Stories™ (2001). These are instructional stories that explain how to behave in particular social situations. They are written following a set of guidelines developed by Carol Gray, that state that sentences should be short and explicit and accompanied by simple, informative icons that support understanding. An extensive study evaluating the efficacy of Social Stories was conducted by Quirmbach et al (2008). It involved forty five children in a randomised control trial and examined the ability of the stories to elicit, maintain and generalise cooperative behaviours in a game. The results strongly supported the effectiveness of Social Stories™ for this purpose, as all children with average verbal skills or above, as measured by the Verbal Comprehension Index, made significant improvements. Social Stories™ are very visual and provide explicit instructions specific to the situation, suiting the typical learning style of children with autism. Several other smaller studies support these findings (Delano and Snell 2006, Sansosti and Powell-Smith 2008, Balakrishnan and Alias 2017), while other studies suggest that the effectiveness of Social Stories™ can be variable and reliant on a variety of factors, including the quality of the stories themselves and the behaviours they are being applied to, particularly given that untrained individuals are often the ones responsible for implementing the intervention (Lorimer et al. 2002, Reynhout and Carter 2006).

Table 3: Examples of existing interventions for children with autism

Intervention Title	Summary	Evaluation Outcomes
Traditional Interventions		
Applied Behaviour Analysis (Lovaas 1987, Schreibman 2000)	Therapist provides consequences for behaviours, e.g. objects, food and actions that the learner finds reinforcing when desirable behaviours occur.	Well established as a highly successful technique, particularly when used for early intervention. Time intensive and can be overwhelming for the participant (Sallows and Graupner 2005).
Social Stories™ (Gray 2001, Quirmbach et al. 2008, Balakrishnan and Alias 2017)	Carefully formatted instructional stories explain how to behave in particular social situations.	Extensive study by Quirmbach et al. (2008) indicates high effectiveness, however outcomes rely heavily on the quality of the stories used.
LEGO therapy and the Social Use of Language Programme (SULP) (Owens et al. 2008)	Compared LEGO building tasks as facilitator of social interaction against established peer group intervention SULP which involves stories, modelling and role-play.	LEGO therapy reduced maladaptive behaviours, SULP improved social and communication skills. LEGO found to be a natural motivator and helped children interact with peers.
Video Modelling (Marcus and Wilder 2009, Dowrick 2012)	Children watch videos of themselves (self-modelling) or others (peer-modelling) correctly performing desired behaviours.	Strong evidence of effectiveness, especially self-modelling, but may need to be used in conjunction with other techniques to maintain long term behavioural changes (Reichow and Volkmar 2010).
Hardware Interventions		
Robots (Robins et al. 2005, Scassellati 2005, Kozima et al. 2009, Huijnen et al. 2016)	Robots can be used as social facilitators for high functioning children with autism and for eliciting verbalisation from children who are rarely verbal.	Robots appear to be naturally engaging for children with autism, making them potentially effective for a range of applications. Most existing research is still exploratory.
SIDES Cooperative Table Top Game (Piper et al. 2006)	A touch-screen table top game designed to encourage cooperative skill building between up to four players.	Initial play testing indicated high engagement and excitement but more structure required to ensure fairness and enforce pro-social behaviours.
Emotion Bubbles (Madsen et al. 2008)	Wearable camera system includes automated facial expression detection software that lets users know what the expression on a detected face is via colourful bubbles displayed on screen.	Game-like activities were found to be very engaging for users and helped them critically analyse faces for unspoken social cues.
Software Interventions		
Virtual Reality (Parsons et al. 2005, Herrera et al. 2008, Ke and Im 2013, Cheng et al. 2015)	Virtual reality fully immerses the learner in a simulated environment, often using specialised helmets or other hardware.	Virtual environments can improve some daily living skills, however high functioning individuals do not always behave as they would in the real world making generalisation of skills uncertain for some audiences (Parsons et al. 2005).
Augmented Reality (Chen et al. 2015, Washington et al. 2016)	Augmented reality provides information overlayed on a view of the real external environment, often using wearable technology such as specialised glasses or a helmet.	Systems have been shown to increase the percentage of facial expressions correctly recognised and responded to. Generalisation expected to be supported given grounding in the real environment.
Group Therapy Linked Software (Beaumont and Sofronoff 2008, Whalen et al. 2010)	Software developed in conjunction with group therapy provides learners with opportunities for linked practice in the form of role-play with peers and real-world experience.	Both examples here resulted in improved test scores from pre- to post-intervention. Hybrid approaches show promise in addressing generalisation from intervention to real-world application (Whyte et al. 2015).
Standalone Software (Silver and Oakes 2001, Abirached et al. 2011, Hopkins et al. 2011, Sturm et al. 2016)	Standalone software requires no special hardware beyond a home computer making it highly accessible. It is cost effective and can be used often and independently by learners.	The examples here indicated high levels of user engagement and led to improvements in target skills, however generalisation to novel contexts was not evaluated.

In the same vein as Social Stories™, Carol Gray has developed Comic Strip Conversations, which are developed following similar rules to the stories, but in comic strip format. The use of these comic strips has lead to similarly positive results. Wellman et al. (2002) also use a pictorial approach, but start with a more concrete version and gradually work towards the more abstract images. Initially, Wellman et al. (2002) used dolls with cardboard cut-out thought bubbles above their heads, and gradually reduced the concrete supports. Wellman et al. (2002) concentrated on generalisation of skills to real social situations with peers and demonstrated increased performance with skill transfer to novel contexts.

In all of these interventions the focus is on providing visual supports to aid understanding along with clear and concise step-by-step information, as suits the typical learning style of individuals with autism. These same guidelines can also be incorporated into software developed for this user group, whenever visual, written or spoken information is presented.

Play Based and Peer Group Interventions

Peer group interventions are used extensively to help develop social skills in children with autism, often in conjunction with other methods and tools such as LEGO®, robots or software such as the Junior Detective game. Children with autism are less likely than their neurotypical peers to initiate social interaction, often play alongside rather than with peers, and typically engage in less sophisticated interaction behaviours. It is thought that an object of mutual interest, for example LEGO®, acts as a facilitator and can help children with autism interact more richly with peers. Peer group interventions range from short but frequent school based groups, often including neurotypical peers, to longer and less frequent clinical groups, all of which have evidence to support their effectiveness to varying degrees (Owens et al. 2008, Reichow and Volkmar 2010).

Less formal, naturalistic approaches that centre on activities and materials that are naturally motivating and reinforcing and occur in the everyday life of the children with autism have been shown to support generalisation of skills. One such intervention is LEGO® therapy. A study by Owens et al. (2008) contrasted two peer group therapies for 6 to 11 year olds, LEGO® therapy and the Social Use of Language Programme (SULP). In LEGO® therapy, children in small groups are given roles, and must work together following social rules to build a LEGO® construction. The small group can include neurotypical peers and adults as well. The construction task requires group members to use many social behaviours including joint attention, verbal and nonverbal communication, collaboration and problem solving skills. LEGO® is particularly suited to this learner group as it is predictable and systematic, fitting with their common preference for consistency. The SULP intervention is used by a number of schools and therapists and begins with stories, then adults modelling desired behaviours, followed by the children practicing these behaviours and playing games within the social group. Finally, activities are performed in new situations to encourage generalisation. Owens et al. (2008) found that the children involved in LEGO® therapy reduced their maladaptive behaviours, and those in the SULP group improved their social and communication skills, with both intervention groups outperforming those in the control group. Since the two therapies appear to target

different sets of social skills more research is required, however both did lead to improvements in social behaviour and both were relatively cost effective and easy to implement.

The aim of a virtual peer as an intervention is not to replace learning opportunities such as those experienced in real peer to peer play, but to provide a helpful first step in leading to the development of the sophisticated behaviours required for rich, everyday social interactions.

Applied Behaviour Analysis

Applied Behaviour Analysis (ABA) is one of the most widely known techniques for reducing undesirable behaviours and increasing preferred behaviours in children with autism (Lovaas 1987). Traditional ABA involves a therapist providing direct consequences, for example providing objects, food and actions that the learner finds reinforcing when desirable behaviours occur. While traditional ABA is very effective at teaching desirable behaviours to children with autism, problems with generalisation to novel contexts and self-initiation of behaviours were found (Schreibman 2000). Modern ABA aims to address these issues by incorporating more naturalistic behavioural approaches that use real-world settings and are more child-driven, and this has had demonstrated success (Schreibman 2000). ABA is highly effective when used as an early intervention technique, with Sallows and Graupner (2005) finding that approximately 48% of children under 5 years old who received the prescribed ABA intervention were successful in mainstream school classrooms by age 7, and many more made significant improvements to their language, intellectual and adaptive skills. There are some shortcomings to this approach, particularly its time consuming nature. ABA also relies heavily on trained professionals, which quickly becomes expensive. Additionally, it requires the child to interact in an intense fashion with another human being which can be very confronting, at least initially (Hailpern 2007). The ABA approach is known to be effective for many individuals and is widely used for a variety of different applications. Many programs similar to that of Lovaas have been developed and its principles, such as prompting and positive reinforcement, are used in a range of settings (Reichow and Volkmar 2010). These principles are likewise suitable for inclusion in the social tutoring software being developed here.

TEACCH Intervention

Panerai, Ferrante and Zingale (2002) investigated the effectiveness of the Benefits of the Treatment and Education of Autistic and Communication Handicapped Children (TEACCH) programme as compared to a control group who were in typical classrooms with support teachers. They found that students in the TEACCH program made significant gains across the duration of the evaluation. TEACCH provides continuous, structured intervention, has a strong focus on the use of visual aids to make abstract concepts more concrete, and provides for environmental adaptation and training in alternative communication (Panerai et al. 2002). As autism is a pervasive disorder, TEACCH is designed to be used in all aspects of the learner's life instead of being restricted to specific learning sessions. The use of visual aids, adaptations in the learner's environment, and the focus on providing more methods of communication are all important general

principles that are widely used in a variety of educational situations for children with autism, and provide valuable guidance for the educational approach implemented in the current research.

Video Modelling

Video modelling is a technique in which the learner is shown a video of someone, possibly a peer or themselves (self-modelling), performing an action that the learner is intended to acquire. Video modelling has many advantages including that minimal expertise or expense is required to implement the intervention, it is repeatable, it can be conducted in a standardised manner, and it is portable. A review by Reichow and Volkmar (2010) into best practices for social skills interventions found numerous studies supporting the effectiveness of video modelling, but suggest that video modelling alone may not be sufficient to maintain long term behavioural changes and state that more research is required into exactly what circumstances optimise the effectiveness of video modelling, for example the type of model, such as self, peer or adult. A more recent review by Wong et al. (2015) likewise indicated strong support for the use of video modelling as an evidence-based practice for teaching skills to individuals with autism generally, and work by Dowrick (2012) suggests that the reason self-modelling is successful is that it increases learners' ability to see their own potential in achieving the target behaviour.

Marcus and Wilder (2009) compared the effectiveness of self-video modelling and peer-video modelling with three children with autism, one four year old male, one nine year old male and one nine year old female. The acquisition task was for the children to learn the sounds and symbols for a set of Greek and Arabic letters. In the self-modelling condition, all three children reached the mastery condition whereas only one child did in the peer-modelling condition. Anecdotally, the authors reported that children enjoyed the self videos more and even wanted to watch them after the study was concluded. However, this study involved a textual task not a socially oriented one. Sherer et al (2001) compared self and video modelling for teaching conversation skills to five children, but found no significant difference between the two, with some learners performing better in one condition and some in the other. More recently Sng et al. (2014) reviewed video modelling and scripts for teaching conversation skills specifically, and found video modelling to be borderline between questionable and effective as an intervention for this purpose. Thus, video modelling has strong evidence of effectiveness for teaching many different types of skills to individuals with autism, including social skills, but more investigation is required for conversation-related skills specifically.

It is hoped that the human-like appearance and behaviour of the virtual characters in the software developed for the current research may capitalise on the same effects that cause video modelling to be so successful, thus enhancing educational outcomes for learners.

2.2.2 Technology Based Interventions

It is often said that individuals with autism have an affinity for computers and technology in general, and both the survey by Putnam and Chong (2008) and the recent investigation of technology usage patterns among adolescents with autism (MacMullin et al. 2016) support this. Consequently, any technology based

intervention is likely to be appealing to young people with autism. Combining this innate interest with educational content is hoped to prove very beneficial for them educationally. In a study by Jacklin and Farr (2005), the impact of computer use in general on the social interactions of children with moderate autism was investigated. The motivation behind this was that using a computer would provide an object of joint attention and would help to lower anxiety levels, making the social interaction more enjoyable and relaxing. When the children were focussed on their computer based tasks, fewer self-stimulatory behaviours were observed and they were more willing to interact with their teachers. Even more encouraging was that in a follow up case study better turn taking and on-task behaviour was observed, fewer maladaptive behaviours were present, and the participants displayed improved eye gaze and problem solving skills. Jacklin and Farr (2005) emphasise the importance of monitoring computer use to ensure that it does not reinforce any obsessive or repetitive behaviours or increase the learner's social isolation.

As there are many technology based interventions currently available, only a sample is given here. These are restricted to approaches that have been experimentally validated, provide a particularly novel approach, or are directly informative in the development of the Social Tutor and the choice of its platform.

Robots and Hardware

While robots are very appealing and motivating for many children with and without autism, extensive research into their efficacy is still somewhat lacking, with most existing research being exploratory in nature (Huijnen et al. 2016). Furthermore, robots can be quite expensive and present numerous drawbacks in terms of their usefulness as social skills interventions. Robots have a set appearance, not being customisable in this sense. This makes generalising any social skills that children with autism may develop while using the robot into a considerable challenge. Additionally, their appearance is typically very dissimilar to a real human. For children who find faces difficult to look at this may be an advantage, making the robot an anxiety free learning tool, but conversely it is likely to make generalisation of skills to a real person difficult. Thus, robots may not be best suited to the purpose of teaching 'social etiquette' between socially active individuals, however there is evidence of their potential as social facilitators, helping to break down barriers and make interacting with peers and adults easier for children with autism (Huijnen et al. 2016).

Research from the AuRoRA group has demonstrated that robots help to engage high functioning children in social interaction with adults and their peers, and help low functioning children engage in parallel play, an important first step towards socially interactive play (Werry et al. 2001, Robins et al. 2005). Another group of researchers have also investigated the notion of robots as social catalysts, with equally promising results. Scassellati (2005) found that by reacting to participants' actions, rather than simply following a set script, the number of social behaviours from the participants towards the robot was significantly higher. It was found that even a very simple robot following a set script was potentially useful for encouraging low functioning, rarely vocal children with autism to elicit vocalisations, generating excitement and many utterances from participants (Scassellati 2005). Another simple, commercially available robot is Keepon (Kozima et al. 2009). Keepon has been carefully designed to ensure that it conveys the potential for social agency and

emotional expression while being very simple in appearance, in line with its capabilities, and making it socially accessible for young children with autism. It can be used in both autonomous and authorable mode, and approximately 400 hours of interaction data has been collected over the course of four years. Interestingly, Keepon has been shown to elicit social actions from children on the spectrum, including spontaneous shared observation of Keepon's mental states with a third party such as a caregiver.

A more recent example of a social robot is Nao, who is also commercially available and able to act both autonomously and in an authorable 'Wizard of Oz' mode (Huskens et al. 2015, Warren et al. 2015). Nao has been used as a mediator in a LEGO-based intervention, and has also been used as part of an autonomous system designed to model social gestures to children, assess the quality of their imitations, and give feedback. It should be noted that outcomes so far have been mixed, with one suggested explanation being the limited repertoire of social responses Nao can produce. Still, Nao remains a very interesting platform for future research. For a more extensive list of existing social robots see Huijnen et al. (2016). While robots may not currently be suited for teaching rules of social etiquette, there is clearly potential for many other social and language skill benefits to be gained from their use.

A few novel hardware based interventions have also been developed, notably the SIDES cooperative table top game and the Emotion Bubbles portable system. The SIDES cooperative table top game was developed in close consultation with twelve high school students with autism. The goal was to develop a game that encouraged cooperative skill development without it feeling like an educational game (Piper et al. 2006). A sturdy touch screen big enough for four players to sit around and interact simultaneously is at the heart of the system. As many individuals with autism experience poor fine motor skills, a large touch screen makes it accessible to a wider range learners. The game itself enforces the rules, making it more predictable than a human ‘referee’ and helping to reduce anxiety while learners are having fun and developing confidence in their social skills. Initial play testing indicated that the system was very motivating and exciting, but perhaps too exciting as players often talked over each other and quieter players were left out. Increased built-in structure is required to encourage more pro-social behaviours (Piper et al. 2006).

The Emotion Bubbles system also mentioned combines a small portable computer and a software package that aims to help learners with autism to read facial expressions (Madsen et al. 2008). The computer’s camera can be pointed towards a person’s face, which is then analysed in the software and the ‘emotion bubbles’, represented on-screen as colourful circles, will grow or shrink to indicate which emotion is being represented on the tracked face, and to what extent. A pilot study involving three high school age males evaluated the potential of the system. The participants were asked to point the camera at their partner and try to get them to display particular emotions, using the system as a guide. The results suggest that this technology has much potential, as the boys quickly understood how to use the software and appeared to thoroughly enjoy the experience. The next stage of development for the Emotion Bubbles system is applying it to teach skills that can be used in real social contexts.

While interesting lessons can be learned from these robot and hardware-based systems, the need for specialist equipment and its cost can be a barrier to uptake for many families, and having a fixed configuration can be limiting. Thus, the Social Tutor is targeted for use on standard home computers.

Virtual Environments and Augmented Reality

Virtual and augmented environments are appealing and motivating for most learners, and are thought to promote generalisation to real-world contexts as the learner is either entirely immersed into a simulated environment or is interacting in the real environment with additional information overlayed on some form of display. These technologies provide learners with the opportunity to role-play scenarios in a realistic yet supported environment. However, these approaches also come with limitations. Like the robot and hardware-based approaches, virtual and augmented reality often require specialised, sometimes costly, equipment. For some learners, particularly those with sensory issues, having to wear equipment such as helmets or goggles can also be a major barrier, and while virtual and augmented reality applications can be deployed as three dimensional worlds on a typical computer or mobile device, the immersive effect is not as strong.

Perhaps more concerning, Parsons et al. (2005) found that teenagers with autism behaved differently in the simulated environment than they would in a real environment, and stated that because they knew the environment was not real, they did not feel the need to behave in their normal manner. This suggests that generalisation for this functioning level and age group may not be supported. However, virtual reality has been shown to lead to significant benefits for children with autism for other purposes, such as teaching life skills including finding a seat in a crowded cafe and safely crossing the road (Strickland 1998, Kerr 2002)

Herrera et al. (2008) developed a virtual environment that used a scaffolding approach to gradually take children from functional interaction to imaginative play. In this manner, abstract ideas can be made concrete and illustrated clearly. Through use of this virtual environment children improved their skills, with one participant even generalising their skills to another context. Children with autism have difficulty identifying their mistakes and the causes behind them and must be explicitly taught how to deal with new situations. In a situation with peers, making a social mistake can cause severe anxiety and discomfort for the child. Thus, collaborative virtual environments which facilitate role-play between real humans but in a controlled manner may provide a highly beneficial environment for learners with autism to practice their social skills in a less threatening context (Kerr 2002). Software-based learning opportunities make it easy to keep initial scenarios simple and gradually add distractions and complexities as the learner increases their confidence. Kerr emphasises that the purpose of virtual environments as autism interventions is as a valuable tool for developing learners' social skills repertoires, and must be accompanied by practice in real social situations. These same advantages and caveats apply to the development of the Social Tutor for this research.

Existing virtual worlds such as Second Life provide another interesting avenue for investigation, particularly since they are reasonably accessible to families, requiring only the use of a standard computer and not any specialised equipment. Ke and Im (2013) developed a set of social skills focussed Second Life tasks and

tested the efficacy of this approach with four primary school aged children. A group of adults with education backgrounds were also recruited for the study, their role being to control characters within Second Life, interacting with the children during their learning tasks as communication partners and facilitators. It was found that in general the participants improved their ability to initiate and maintain social behaviours, and also improved their dispositions towards developing peer friendships and engaging in social interactions with others (Ke and Im 2013). While promising, this approach required the involvement of adult mediators, thus learners were not fully self-sufficient, in contrast with the goal of the Social Tutor being developed here.

A more recent study by Cheng et al. (2015) involved development of a three-dimensional virtual environment to teach various aspects of social understanding, deployed using a head-mounted display. They conducted a preliminary study over six weeks with three participants on the autism spectrum, aged 10 to 13, and found that the target behaviours improved from baseline to intervention, and improvements were maintained at two, four and six weeks post-intervention. While Cheng et al. (2015) did not formally evaluate generalisation to everyday situations, anecdotal evidence suggests that some generalisation did occur, for example one participant increased their efforts to socialise with the researchers, use manners and raised their hand when the virtual character asked a question. While only a preliminary evaluation, it lends support to the idea of virtual environments as promising tools for improving social skills in children on the spectrum.

Augmented reality is another interesting technology gaining ground in the area of autism intervention, particularly when paired with wearable or otherwise mobile devices. The recent release of Google Glass has opened up new avenues for researchers, with Washington et al. (2016) harnessing the technology to create a prototype wearable social aid for children with autism. The system uses automated emotion recognition and provides social cues in real-time on the heads up display. The system can run in a casual mode, or wearers can engage in gamified activities that encourage them to develop their emotion recognition skills. The system also auto-records 'emotional moments' throughout the day that can be reviewed by parents and therapists via an Android application. An initial evaluation of the system has been conducted with twenty children with autism and twenty typically developing children, aged 6 to 17 years old. Children responded well to wearing Google Glass and enjoyed the gamified activities and feedback mechanisms. Interestingly, participants overwhelmingly preferred verbal cues over visual cues, finding the visual cues distracting (Washington et al. 2016). While still early days, the combination of augmented reality with wearable technology holds much potential for social skill development in children with autism. However, for the purposes of the current research, equipment that may present a barrier to uptake for families is undesirable.

General Software

A wide range of software targeting many of the difficulties associated with autism are available, for example software has been developed to encourage vocalisation in pre-vocal children (Hailpern et al. 2009) and to encourage development of spoken language in young children at the earlier stages of language acquisition (Lehman 1998), however many of these programs have not been experimentally validated. Some examples of software that have received positive experimental results and focus on social skills for higher functioning

individuals include The Junior Detective, Teach Town: Basics, FaceSay and Emotion Trainer (Silver and Oakes 2001, Beaumont and Sofronoff 2008, Whalen et al. 2010, Hopkins et al. 2011).

The Junior Detective and TeachTown: Basics are both computer assisted intervention programs that involve a software use component alongside opportunities to practice skills in real-world role-plays (Beaumont and Sofronoff 2008, Whalen et al. 2010, Jones et al. 2016). Hybrid approaches such as these appear to be promising in addressing the issue of generalisation from intervention to real-world application (Whyte et al. 2015). The Junior Detective computer game was evaluated as part of a sequence of social skills group therapy sessions, where students were given opportunities to role-play the skills taught in the game. It was found that this combination led to significant improvements in the participants' social skills and their ability to suggest strategies to manage their emotions and those of others. In a follow up session months later, participants had maintained their skills (Beaumont and Sofronoff 2008). The TeachTown: Basics software takes an ABA approach where learners are taught using a discrete trial format and correct responses are reinforced immediately with praise and graphics, and on a variable ratio also rewarded with short animated games (Whalen et al. 2010, Jones et al. 2016). The TeachTown Connection real-world activities aimed to generalise the skills taught in the software as well as teaching additional skills and utilise principles of Pivotal Response Training. Most students showed significant improvement from pre-test to post-test, including on standardised measures (Whalen et al. 2010). Both TeachTown: Basics and The Junior Detective demonstrate how software can be used as step in the scaffolding process that leads to the development and maintenance of sophisticated social behaviours and problem solving skills.

Emotion Trainer, FaceSay and LIFEisGAME are examples of software designed to teach children how to identify emotions based on the appearance of peoples' faces (Silver and Oakes 2001, Abirached et al. 2011, Hopkins et al. 2011). Silver and Oakes (2001) developed Emotion Trainer, which presents learners with an image or text description of an emotional face or scene and proves multiple choice buttons for learners to use to indicate which emotion is being depicted. Students are rewarded with a 'well done' message and a simple animation for a correct choice, and asked to 'try again' and given a direct cue for an incorrect choice. While there are five sections of increasing difficulty, the program does not adapt to the user. The Emotion Trainer was evaluated using a randomised control trial in which eleven pairs of children with autism matched by age, school grade and gender participated. One child in each pair used the software while the other child did not. All children who used the software improved their skills, but to varying degrees, compared to those who did not. Additionally, children were able to generalise their skills to a similar paper-based task, but their ability to apply their skills to real social situations was not investigated (Silver and Oakes 2001).

While Emotion Trainer offered one primary type of task and had one goal, the FaceSay software offers learners a range of games, with the overall aims being to increase their skills in emotion detection, face detection and social interaction (Hopkins et al. 2011). The games include identifying what object a face was looking at, matching the missing facial part to a given face, and matching the expression on a pair of faces. Again, the software did not adapt to the user. It was found that children classified as having 'low functioning'

autism improved on both emotion recognition and social interactions, while high functioning children improved in these target areas as well as facial recognition (Hopkins et al. 2011).

In contrast to FaceSay and Emotion Trainer, LIFEisGAME takes a unique approach to teaching emotion recognition and uses Active Appearance Models to have a virtual character directly mimic the user's own facial expression in real-time (Abirached et al. 2011). The pilot study presented users with a set of games ranging from observation and recognition, to matching a shown expression with their own face. Users responded well to these games, and the approach was found to be highly motivating.

In more recent research, the serious game eMot-iCan has been developed for mobile devices and is designed for teaching and assessing emotion recognition skills (Sturm et al. 2016). The authors suggest that atypical attention patterns may be behind many of the social and communication difficulties experienced by individuals with autism, and aim to explicitly teach users what elements to pay attention to in order to read facial expressions. Users are presented with a set of photo or cartoon images and must choose the correct match. Some additionally noteworthy features of this work include that administrators can customise the trials for individual users, and that being designed for a mobile device means that it can be taken to clinics and schools and used in a consistent manner across various environments (Sturm et al. 2016). Pilot results suggest that both administrators and children found their aspects of the software intuitive to use and engaging.

Well-designed software certainly appears to be a promising avenue for basic social skills development in children on the autism spectrum, with the added benefits of it not requiring any specialised equipment and typically not presenting any major barriers for individuals with sensory difficulties. Many lessons can be learned from the sample of technology-based interventions provided here, particularly around the importance of scaffolding and insights into the kinds of activities that learners find engaging and useful.

2.3 General Educational Considerations

When developing educational software for any user group, the educational process that takes place must be carefully considered so that the mechanisms put in place within the software support its ultimate goals. This is particularly true for individuals with autism given that generalisation of skills to novel contexts is a known difficulty, and that a mismatch between the techniques used in existing interventions and the needs of this learner group is thought to contribute to this (McCleery 2015).

The Adaptive Control of Thought – Rational (ACT-R) theory of human cognition is one of the best established theories of human cognition (Ritter et al. 2007, Crook and Sutherland 2017). ACT-R suggests that for educational materials to be most effective they must present concepts along with procedures so that students can understand what they are doing and why, new knowledge must build upon existing knowledge so that stronger and longer lasting connections can be made, and students must be presented with opportunities to learn and practice their skills that includes dedicated instruction, explorative experiences and

defined tasks. ACT-R further states that students' knowledge must be assessed regularly to ensure that the educational materials presented are focussing on what the individual student needs, rather than providing unnecessary instruction in areas where they are already confident, and relevant feedback must be provided (Ritter et al. 2007, Crook and Sutherland 2017). Thus, the ACT-R theory emphasises the importance of students building conceptual knowledge and cognitive skills, not just procedural skills and rote facts. Again, this provides strong indication that a scaffolding approach should be incorporated in the Social Tutor software, and highlights the benefits of explaining to learners why the skills they are being taught and the tasks they are presented with are important and worthy of their time and effort. This is particularly relevant for individuals with autism who may not intuitively recognise the value of particular social niceties.

One-on-one tutoring has been found to produce greater understanding and a higher level of motivation in students than traditional classroom situations, and students are often able to progress through topic content at a faster rate. The incorporation of virtual characters into the Social Tutor software is intended to echo this interaction style. Historically the average performance of students in a one-on-one tutoring situation was believed to be around two standard deviations above the average of students in a typical classroom situation (Graesser et al. 1999, Chi et al. 2001), however a more recent analysis of the literature found the average effect size over the included studies to be smaller, at only 0.79 standard deviations (VanLehn 2011). Interestingly, the average performance of the intelligent tutoring systems reviewed was 0.76 standard deviations, indicating performance close to that of an expert human one-on-one tutor (VanLehn 2011). In both cases, personalised instruction is shown to lead to improved outcomes for learners when compared to a traditional classroom, and this is something that social tutoring software can successfully provide.

2.3.1 What Human Tutors Do

In the case of a virtual tutor, understanding what makes human one-on-one tutoring effective is essential to developing a useful application. As the meta-analysis by Bowman-Perrott et al. (2013) demonstrates, peer tutoring is a well-established evidence-based practice that has been successfully used across a range of subjects, settings and age groups, and is shown to be effective for learners both with and without disabilities.

There are two main schools of thought about what makes tutoring effective, one being that it is the tutor's actions that result in positive learning outcomes, and the second suggesting that it is the student's ability to construct knowledge and build connections between concepts that results in learning, and that successful tutors facilitate this process (Chi et al. 2001, Hmelo-Silver and Barrows 2015). It was found that even when tutors were restricted from giving explanations and feedback and could only prompt the students, students learned just as effectively, however this approach relies on students having access to the information they need in a format that they can consume, and it is suggested that the effect is due to the students having to take more control of their own learning (Chi et al. 2001).

There are many hypotheses surrounding why a tutor's actions can lead to positive learning outcomes, some with more evidential support than others. One is that human tutors are thought to engage in continuous

diagnostic assessment of their tutee, identifying gaps in mastery, misconceptions and lacking skills. Unfortunately in practice human tutors rarely engage in this behaviour and often fail to ask questions that could help them unearth this information about their tutees (Putnam 1987, VanLehn 2011). While acknowledged as a core component of teaching 'best practice' (Ritter et al. 2007), this is an issue that has been acknowledged for some time. For example Putnam (1987) observed expert teachers working with students on simple mathematics problems and found that they only appeared to explicitly determine the nature of a difficulty before correcting it 7% of the time. Another hypothesis is that tasks can be individualised for the tutee, however again human tutors are often found to simply work from a curriculum script with only minor deviation, much the same as in a modern classroom where providing differentiated curriculum for learners is expected.

Learner control of dialogue and the broader domain knowledge of the tutor are two more areas where tutoring is hypothesised to be at an advantage, as students can ask as many questions as they need to achieve understanding and tutors can explain concepts in depth and in alternative ways, however it has been shown that students rarely take initiative outside of confirming that a statement they make or a behaviour they are performing is correct, and the broader domain knowledge is rarely utilised to advantage (VanLehn 2011). In all of these areas, computer tutoring systems can perform in a similar manner to how human tutors behave in practice, although there is a lot of scope for both human tutors and computer systems to increase the richness and personalisation of their teaching approaches here.

A tutor action hypothesis that appears more promising is that of immediate feedback and prompting. This process lets the student know they are on the right track and guides them towards correct understanding (Chi et al. 2001, Bowman-Perrott et al. 2013). In a one-on-one scenario, tutors typically allow the student to continue at their own pace until they get stuck or make a mistake, they then intervene to resolve the issue so the student can continue without losing momentum (VanLehn 2011). In a classroom scenario, the student's error may not get identified straight away or they may not receive assistance immediately, causing them to stall and lose this momentum, become unnecessarily frustrated or confused, and possibly resulting in the need to backtrack and re-do work. Tutors also typically encourage students to 'think out loud' and explain their reasoning as they go, making it easier to identify misunderstandings and facilitating students to become actively engaged with their learning, rather than passive recipients of information (Merrill et al. 1992, Chi and Wylie 2014). There are a number of techniques that can be used in software-based tutoring systems to emulate this behaviour to benefit learners, such as the previously mentioned scaffolding approach, specifically breaking large tasks into smaller subtasks that can be individually assessed with feedback provided to assist the learner in future attempts.

As is evident from its wide implementation in the social skills interventions already discussed, scaffolding can be a very powerful learning tool, particularly for individuals with autism who need concepts to be explicitly taught. Specifically, scaffolding often involves decomposing complex concepts and tasks into simpler, more manageable subtasks, and within a lesson sequence this often takes on the broad format of

introduction, demonstration, then practice, where demonstration for social skills in an autonomous tutoring application could involve multiple agents acting out the interaction to be practiced (Jackson et al. 2010a, van de Pol et al. 2015). Tutors encourage learning by guiding students through this process, helping learners to master the subtasks, and gradually working towards the final goal. As opposed to a classroom setting where the learner often passively receives information, a tutoring session encourages students to interact with their new knowledge through predicting, justifying, criticising and otherwise engaging with the material (Chi et al. 2001, Chi and Wylie 2014).

Interesting to note is that most tutors lack formal training, and yet tutoring is a very effective educational tool even when feedback is provided by a peer or other non-expert (Chi et al. 2001, Hamer et al. 2015). This suggests that even an imperfect tutor can provide great benefit, and thus indicates that an imperfect virtual tutor can also still be valuable to students. Tutors typically follow a set pattern when working with learners. First, the tutor asks a question, to which the student provides an answer. The tutor provides feedback, and performs scaffolding across a number of turns with the student in order to help the learner develop their understanding. Finally the tutor assesses the learner's comprehension of the taught content (Chi et al. 2001). This same pattern can be performed by a virtual tutor. Throughout this process, the tutor monitors the learner for confusion and frustration, as deeper learning is achieved when learner misconceptions are addressed immediately.

For a virtual tutor to identify when learners are struggling or have misunderstood, several methods can be employed. Basic approaches such as tracking student performance across and within tasks, analysing the mistakes made or tracking the number of times students engage with the same task, are relatively straightforward to integrate into a software tutor. Another approach with some potential is detection of frustration, confusion or boredom through facial expression recognition, emotion in speech, or use of external sensors. While emotion detection is outside of the scope of the current project as it relies on the availability of particular hardware, such as a webcam, microphone or biometric sensors, understanding the emotional aspects of learning and the impact different affective states have on learning outcomes is important when designing educational activities.

2.3.2 Emotional Aspects of Learning

Emotion can have a strong impact on learning, so to maximise educational outcomes it is important to understand this relationship. Students experience a wide range of emotions while they are learning, from confusion, frustration, dejection and boredom, to satisfaction, enthusiasm and excitement. Typically individuals who are anxious, angry or depressed do not retain information effectively or perform well in learning tasks, so it is the role of the tutor to guide learners through these states and into affective states more conducive to learning, as expert human teachers naturally do (Kort et al. 2001, Storbeck et al. 2015).

Emotions can be viewed has having an evolutionary function, where even slightly stressful situations and negative emotions can trigger a flight or fight reflex. This results in many physiological changes including an

increase in heart rate and blood pressure, and adrenaline being released which causes the brain to switch into a reactive mode rather than a reflective, problem-solving mode (O'Regan 2003, Wolfe 2006, Storbeck et al. 2015). While memory is enhanced at this time, it is not typically an ideal situation in which to be learning new concepts or making new connections (Wolfe 2006). Further, research indicates that positive emotions during learning can reduce cognitive effort and increase working memory (Storbeck et al. 2015) and thus providing students with learning opportunities that are inherently pleasant can also result in strong retention. Gamification is one approach gaining much attention of late, with a recent review showing that in an educational context inclusion of game-like aspects or embedding the learning within a game can, when done mindfully, lead to increased motivation, engagement and enjoyment (Hamari et al. 2014).

It should be acknowledged that a small degree of frustration or uncertainty can be constructive and may even indicate that a learner is in their zone of proximal development (Vygotsky 1978). The zone of proximal development is defined in the seminal article of Vygotsky (1978) as "the distance between [a learner's] actual developmental level as determined through independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" and is considered the 'sweet spot' where the balance between what an individual knows and what they need to know is ideal for making a new connection. Some frustration is natural while learning a new skill or concept, for example when a student recognises that they are close to succeeding in a task and becomes motivated to persevere until they do. It is a fine line to tread, with too much frustration being counterproductive and disengaging, however human tutors intuitively step in at the right time to support the learner. Emulating this behaviour remains an active research area for intelligent tutoring software.

Emotions can also be seen from a behaviourist viewpoint where the emotions themselves act as reward and punishment, and therefore influence the choices an individual makes. In this context, negative emotions like anxiety behave as punishment and cause the individual to avoid the situation that triggered the negative emotion (O'Regan 2003). The goal is therefore to minimise the occurrence of unnecessary negative emotions and increase the positive emotions. Acknowledging and celebrating learners achievements, using positive reinforcement, and providing inherently enjoyable tasks are important ways to maximise the positive emotions associated with learning, while detecting negative emotions and dealing with them constructively, for example noticing frustration and providing guidance, are essential in minimising the negative emotions.

2.4 Advantages of Virtual Tutors

There are a multitude of reasons why virtual tutors, or pedagogical agents as they are often known, are well suited for use with individuals with autism, not least of which is the widely acknowledged affinity that many individuals report having with computers and technology (Putnam and Chong 2008, Baron-Cohen et al. 2009). A technology based intervention for social skills is appealing for students with autism for this reason, and also allows the learner to work through material at their own pace rather than being subjected to the constraints of a classroom or therapy group. It is openly acknowledged that nothing should aim to replace

genuine interaction with peers and others when learning about social interaction; however, an independent learning tool, such as one incorporating a virtual tutor, can provide a valuable first step in developing these complex social skills, as well as giving learners a way to revisit and consolidate their skills as needed. Furthermore, the technology required to use virtual tutoring software is becoming more and more accessible and affordable for families, with tablet and mobile computing in particular developing at a rapid pace in recent years (Ploog et al. 2013, Meder and Wegner 2015).

Autonomous virtual tutors allow learners to practice their developing skills independently, relieving some pressure from caregivers, teachers, and other professionals who work with the student, and can be used to complement other interventions. This also allows those working with the learner to focus on the more complex and specific aspects of the learner's education, while many routine and general points are covered by the virtual tutor. Virtual tutors can provide a stress-free learning opportunity as the anxiety connected with interacting with real humans is removed, and the tutor can be programmed to ensure that it only provides positive and guiding feedback, rather than criticism that other humans may provide. Additionally, using a virtual tutor means that the learner can practice their skills without interfering with others or learning inappropriate responses from incidental people in the learning environment (Kerr 2002, Gay et al. 2016). Another advantage is that the virtual tutor will never get tired or impatient, unlike even the most patient human teacher (Massaro 2004). Additionally, a virtual tutor can be available for practice at any time of the day, which may particularly suit those individuals who experience abnormal sleep patterns (Limoges et al. 2005). Virtual tutors provide consistent feedback and behaviours, which can help control anxiety in those who feel more at ease in predictable situations (Parsons et al. 2000, South and Rodgers 2017).

Software-based virtual tutors are highly customisable and can be tailored to suit the individual learner's needs, an important consideration for any technology used with children on the spectrum (Ploog et al. 2013). For example, for a learner who finds looking at faces uncomfortable the virtual human's appearance could start out very cartoon-like and, as the learner becomes accustomed to it and their confidence grows, the realism and complexity can gradually be increased. Similarly, the lesson content can be modified to meet the individual's current level of interest and need. For example, generic images can be replaced with those that have special significance to the learner, and any taught phrases can be updated to match those being learned in school or therapist-based interventions the learner is participating in. Being software-based, many different media can be incorporated into the learning material, including line drawings, photos, videos, animations and more, and training with a variety of stimuli is one method that can help to support generalisation of skills to novel situations (Stokes and Baer 1977, McCleery 2015). Further, the software can adapt to the learner's needs in a dynamic way, only offering complex lessons once simpler prerequisite lessons have been successfully attempted.

In the context of developing nonverbal skills, animated virtual tutors can be particularly useful as they can model behaviours for the learner, for example facial expressions, body language and gaze behaviours. This is akin to the video modelling technique which has had success with many individuals with autism (Ploog et al.

2013). Furthermore, multiple tutor 'personas' with unique appearances and voices can be used to model target behaviours in an effort to improve the likelihood of generalisation to multiple people and situations. The flexibility and customisation that virtual tutors offer make them a cost effective and potentially highly beneficial intervention tool.

2.5 Existing Virtual Tutors

The development of pedagogical agents and associated technologies is a very active research area, and inspiration for development of the Social Tutor software in this research is drawn from the examples listed here. Many novel proposals and agents exist in the early stage of development, however only those that have been developed to a functional stage with their effectiveness investigated and that align particularly closely with the goals of the Social Tutor are discussed here. Pedagogical agents are used in a range of educational applications, from on the job training for adults to teaching basic skills to young children. As this research project focuses on tutoring software for school aged children, only applications aimed at this audience are discussed, and to be classified as 'tutoring' software rather than 'general' software, the program must respond to the user in a manner that at least partially resembles how a human tutor would respond. As previously discussed, some examples of 'general' software are included in Subsection 2.2.2.

Virtual tutors are gaining traction in a multitude of areas, including showing promise for application in the complex area of social skills development for children on the autism spectrum. Embodied pedagogical agents can be categorised into two groups, the first where the agent only appears when it is required, either requested or unrequested, and the second, referred to as a peer learning agent or virtual peer, where the agent is always present either as a learning partner or opponent in a task (Sklar and Richards 2010). Table 4 provides an overview of the tutoring applications discussed in detail here.

While a diverse set of application areas are discussed here, they provide evidence of the efficacy of virtual tutors for teaching children a variety of skills in different contexts, and have led to inspiration for many features of the Social Tutor software developed for the current project, ranging from approaches towards user interaction, underlying pedagogical frameworks, virtual agent presentation and responsiveness, and approaches towards personalisation and automated assessment.

2.5.1 Language and Reading Tutors

A number of virtual tutors are available for improving reading and other language skills in children. Project LISTEN, a research project at Carnegie Mellon University, has resulted in the development, deployment and evaluation of an automated reading tutor. This reading tutor, which has been used extensively by several primary schools across America, uses the Sphinx-II speech recogniser to listen as children read aloud, and is able to detect errors and provide spoken and visual feedback immediately. Research has shown the software to be very effective, with children using the software improving their reading significantly faster than peers

in a typical classroom setting (Mostow 2005). While it does not incorporate a visually embodied virtual agent, it does harness speech recognition technology to respond to users in a human-like way.

Table 4: Examples of existing tutoring software

Application Title	Participant Description	Embodied Agent(s)?	Autonomous?	Evaluation Outcomes
Language and Reading Tutors				
Project LISTEN (Mostow 2005)	Primary school students (neurotypical)	No	Yes	Significantly faster improvements in reading vs. typical classroom
iSTART, iSTART-2 (McNamara et al. 2004, Jacovina et al. 2016)	Secondary school students (neurotypical)	Yes - multiple	Yes	Improved comprehension skills vs. no training
Baldi and Timo (Bosseler and Massaro 2003)	Primary school students (autism, hearing impaired)	Yes - single	Yes	Significantly increased vocabulary, maintenance and generalisation to other contexts
Sight Word Pedagogical Agent (Saadatzi et al. 2017)	Young adults (autism, intellectual disability)	Yes - single	Yes	Significantly improved reading of target sight words, maintenance and generalisation to other contexts
Mathematics and Science Tutors				
AutoTutor (Graesser et al. 2005)	Secondary school students (neurotypical)	Yes – single	Yes	Improvement of up to one letter grade in physics assessment
Wayang Outpost (Woolf et al. 2010)	Secondary school students (neurotypical – focus on low achievers)	Yes – multiple	Yes	Reduced anxiety and frustration, improvements in mathematics assessment
Cognitive Tutor (Ritter et al. 2007)	Secondary school students (neurotypical – focus on low achievers)	No	Yes	Significantly higher improvements in Mathematics assessment vs. typical classroom
Betty's Brain (Blair et al. 2007, Biswas et al. 2009)	Primary school students (neurotypical)	Yes – single	Yes	Students who taught the system and received prompts outperformed those without prompts and those building concept maps with system coaching. All conditions lead to learning gains.
ISLA (Mondragon et al. 2016)	Primary school students (autism)	Yes - single	Yes	Students supported by an affective tutor showed better emotion management and higher learning gains than those in a control group.
Social Skills Tutors				
Sam (Tartaro and Cassell 2008)	Primary school students (autism)	Yes - single	No	Improved gaze and turn-taking behaviour, skills generalised to context with human peers, improved scores on Test of Early Language Development
ECHOES (Bernardini et al. 2014)	Primary school students (autism)	Yes - single	Yes	Children enjoyed the system and showed improved social behaviours with the virtual peer. No standardised measures and generalisation not evaluated.

While aimed at high school students rather than primary school students, the University of Memphis has also developed a tutor focussed on teaching strategies for reading comprehension. The reading tutor, Interactive Strategy Training for Active Reading and Thinking (iSTART), includes multiple pedagogical agents who interact with the student to teach a modified version of the Self-Explanation Reading Training (SERT) technique for reading comprehension. They encourage the student to use strategies such as paraphrasing, predicting and elaborating to develop their understanding of the text. In a controlled study it was found that students using iSTART did improve their comprehension skills (McNamara et al. 2004), however after extended software use it was found that student motivation decreased (Jackson et al. 2010b). The research group has since developed iSTART-ME and iSTART-2 which incorporate more game-like elements to enhance long term motivation, and which have demonstrated effectiveness with middle school, high school and college aged students (Snow et al. 2016). This continues to be used as an active research platform for investigating effective design and teaching strategies for pedagogical agents.

Baldi and his successor Timo are animated virtual tutors developed by Bosseler and Massaro (2003) who are designed to improve the vocabulary of children with hearing impairments. Baldi and Timo have anatomically correct facial muscles, allowing their faces to make highly realistic movements during speech. The software in which Baldi and Timo are embedded provides learners with opportunities to interact with the taught words in different ways, using the words in appropriate contexts, representing them as pictures, typing the words, and other activities. The goal of this is to reinforce the sound and meaning of the words and thus encourage a deeper understanding and higher retention of the learned content. It was found that by watching Baldi as he spoke, children improved their vocabulary significantly more than when they only listened. Following the success of Baldi with hearing impaired children, he was trialled with children with autism for the same purpose. The same gains in vocabulary were found, and a month later the children retained 85% of the learned words. Most significantly, these children with autism were able to use the words they learned in everyday situations, providing evidence that virtual tutors can lead to generalisation and use of learned skills in novel contexts for this learner group.

In more recent work Saadatzi et al. (2017) developed a desktop application for teaching sight words to young adults with autism. Three males aged 19-20 with mild to moderate intellectual disability in addition to their diagnosis of autism were involved in the study. The desktop application incorporated a full-bodied pedagogical agent in a virtual classroom and both text-to-speech and automatic speech recognition so that users could engage with the software via the more natural mode of voice rather than typing or tapping. Participants were taught four target words using the software, and were trained until they completed three consecutive sessions with 100% success, then their performance without reinforcement was checked via an assessment phase. 8 weeks after the assessment phase it was found that two participants could remember the target words at 100% accuracy, while no data could be collected from the third. Further, it was found that these participants were able to generalise their use of the four target words to a novel environment in the form of their classroom, and a novel stimuli as written words on paper, and the novel change agent of their teacher. No control group was included in the study so comparisons cannot be made to other methods of

instruction, however it is very encouraging that the learners were able to improve their performance and apply it to contexts outside of the original software environment.

2.5.2 Mathematics and Science Tutors

Along with the iSTART reading tutors, the University of Memphis has developed AutoTutor and its open-source counterpart GnuTutor, which have been used to teach physics concepts to high school students. They incorporate a virtual tutor that asks students questions requiring an answer in sentence format. Natural language processing techniques, such as latent semantic analysis (LSA), are performed on students' written input to determine if the student understands the content or whether more probing questions or hints are required to assess their understanding. The virtual tutor uses speech synthesis and animation to appear lifelike and interact with the student (Graesser et al. 2005). GnuTutor is an open source version of AutoTutor that includes the majority of the functionality, however it is highly reliant on the student possessing good written language skills and is therefore unlikely to be appropriate for use with students who experience communication difficulties, such as children with autism.

Developed by the University of Massachusetts Amherst to help high school students improve their mathematics skills for the Scholastic Aptitude Test (SAT), Wayang Outpost is an online Flash game that includes a number of virtual peer tutors (Woolf et al. 2010). Many mathematical problems are presented in the context of the game's story, and are presented using interactivity and animations. Wayang Outpost stores information about the student's interaction with the system, including hints requested, answers chosen and time taken to answer, and builds this into an individual student model which is in turn used to guide how the virtual tutors react to the student in question. This includes the student's spatial aptitude, which in turn affects whether visual hints or arithmetic hints are more likely to be given by the tutoring system. Even without the intelligence component activated, it was found that Wayang Outpost resulted in improved results for learners, while including the intelligence and decision making component boosted these gains even further (Woolf et al. 2010). Wayang Outpost continues to be used as a research platform for investigating student interactions with intelligent tutoring systems, including investigating underlying factors relating to student affect and motivation (Rai et al. 2013) and provides much inspiration to the current project in terms of engagement strategies and student modelling.

Another successful high school mathematics tutoring program is Carnegie Mellon University's Cognitive Tutor (Ritter et al. 2007). In a study comparing a group of students learning by traditional classroom methods with a group of students using the Cognitive Tutor, it was found that those using the Cognitive Tutor comfortably outperformed their classmates in terms of grades and standard testing. It has also been found that using Cognitive Tutor, student attitudes towards learning mathematics improved and that disadvantaged populations also gained significant benefits. While Cognitive Tutor does not use an animated virtual tutor, it does embody a wide range of relevant technologies including monitoring student knowledge and their interactions with the system in order to adjust content accordingly, guiding students in the right direction, and ensuring students only progress when they have sufficiently mastered the prerequisite skills (Ritter et al.

2007). Cognitive Tutor is now commercially available and also continues to be used in research, for example Fancsali et al. (2016) analysed data from Cognitive Tutor to explore how different aspects of the learning environment can impact student outcomes when using intelligent tutoring systems in a classroom setting, finding that the human teacher also has an important role in ensuring their students engage mindfully and purposefully with such software if positive outcomes are to be achieved.

A novel approach to pedagogical agents that is gaining interest is that of a teachable pedagogical agent, one that the student must teach concepts to as a means of learning the concepts themselves. This idea is motivated by the observation that many teachers find that they have a better understanding of a concept after they have taught it. In this scenario, the student takes more responsibility for their own learning, a valuable life skill, and tests their understanding by trying to pass on their knowledge to a virtual agent. Betty's Brain is an example of such an agent (Blair et al. 2007). Using a concept map style interface, students teach Betty concepts by adding nodes and connections between the nodes. Betty can then answer questions using the concept map, and can tell students when she detects missing information. Betty's Brain has been incorporated into a number of appealing game-like fronts, including a quiz where students put their virtual agents against one another to see which has learnt the concepts best (Blair et al. 2007). The Betty's Brain system was tested with fifth grade students on a task requiring them to develop concept maps about river ecosystems, and then eight weeks later used the same systems but applied to a new topic, the land-based nitrogen cycle. Three versions of Betty's Brain were tested, one in which students taught the system, one in which they taught the system and received prompts from Betty, and one in which they built a concept map for themselves but with coaching from the system. It was found that students in the first two conditions performed better than in the last, providing evidence that learning by teaching is a valuable technique (Biswas et al. 2009). Further work in the 'teachable agents' domain has shown that students complete more tasks and are more motivated when these agents have their own intrinsic motivation and behave in a more friendly and human way (Borjigin et al. 2015), which mirrors the findings of related work with non-teachable pedagogical agents.

Interest in affective tutoring systems is also growing, where the pedagogical agent detects and responds to users' emotional state. One particularly relevant example of such a system is ISLA, developed by Mondragon et al. (2016) to provide emotional support to students with autism while teaching them mathematics skills. While still in the prototype stage, a small evaluation of ISLA was conducted with 12 children with autism, aged 6 to 12 years old, who were randomly allocated into an experimental group with affective support or a control group with no affective support. It was found that the affective support improved participants' levels of encouragement and decreased frustration and anxiety. Mondragon et al. (2016) intend to conduct a longer term and larger scale evaluation given these promising preliminary results.

2.5.3 Social Skills Tutors

The domain knowledge required for the topics discussed previously, such as mathematics and reading, is relatively clear cut and typically has right and wrong answers with set facts and rules that can be followed to reach these outcomes. In contrast, the domain of social interaction presents a bigger challenge. Different

cultural backgrounds, locations, and situations can call for a different ‘answer’ in terms of the social actions required. Also, neurotypical individuals learn social skills almost intuitively through their interactions with parents, peers and others throughout their development, making it hard for educators to know exactly what to teach and how to teach it when a student requires explicit instruction in this area. Understandably, there are few functional pedagogical agents in existence that cater to social skills teaching.

One such example is Sam, a life-size animated virtual peer, designed to be gender ambiguous so that both boys and girls will relate to it (Tartaro and Cassell 2006). All agents described thus far are considered to be autonomous, in that they require no outside input in order to respond to the learner and are embedded in a standalone program. In contrast, Sam is an authorable virtual peer and requires the researcher to observe the student and choose Sam's actions from a set of pre-recorded speech segments and scripted gestures. Sam was designed specifically for children with autism, and engages them in collaborative story telling. Through this Sam models positive social behaviours including turn-taking, gaze and questioning, and helps learners recognise when they are being given an opportunity to contribute to the dialogue (Tartaro and Cassell 2008). It was found that through interaction with Sam learners were able to significantly improve their Test of Early Language Development scores, and even displayed enhanced social behaviours such as improved gaze, which they then used with their peers. This evidence of generalisation is particularly encouraging, indicating that virtual agents can be beneficial in improving the social skills of children with autism.

Another example of a virtual peer for social skills teaching is Andy of the ECHOES program (Bernardini et al. 2014). ECHOES provides an exploratory environment where Andy is presented as an autonomous play partner in a virtual world that the learner can interact with through a large touch screen and eye gaze tracking system. ECHOES is based on strong theoretical underpinnings and best practice principles, with activities based on encouraging cooperation, joint attention and initiation of social behaviours. ECHOES was evaluated via deployment to five school sites across the UK, where a total of 29 children with autism aged 4 to 14 years old interacted with the software over a six week period. The 19 children who had the most exposure to ECHOES and completed all pre- and post-tests were used in the evaluation. While some positive trends were observed in this preliminary evaluation, no significant conclusions could be drawn. Anecdotally some very promising events occurred, for example one child who was initially thought to be non-communicative by his teachers and practitioners waved and said 'Hi Andy!' in a later session, while others began spontaneously greeting their teacher after they had been practicing with Andy, something they had not previously done. The authors reported that teachers and practitioners also expressed enthusiasm about the platform. ECHOES appears to have promise as a tool for practitioners and specialised classrooms, however due to the requirement for specialised equipment it may not be appropriate for home use yet.

Virtual tutors are gaining traction in a multitude of areas, with strong evidence of their efficacy in the more clear cut domains of mathematics and reading, and resulting in some systems such as the Cognitive Tutor being not only well-established experimentally but also now commercially available. Virtual tutors are

likewise showing much promise in the more complex domain of social skills development for children with autism, with Sam and Andy being two prominent examples in this space.

2.6 Using Virtual Tutors as Interventions

Clearly, virtual tutors have been used successfully for many different applications, both with neurotypical children and those with autism. Learning from the experience of others, there are a number of recommendations to be taken into account when designing educational software for learners with autism, along with some particular limitations that come hand-in-hand with developing educational software targeting an area as complex as social skills. These guidelines and known challenges are presented here.

2.6.1 General Development Guidelines

A very broad set of well established guidelines for designers in a range of disciplines are the Principles of Universal Design (Connell et al. 1997), with a more recent iteration of these being the notion of Inclusive Design (University of Cambridge 2017). Both of these philosophies emphasise the importance of designing products and services that are accessible to as many users as possible, regardless of any physical or other challenges they may have. The nature of a virtual tutor delivered via a standard desktop computer means that some of the product-focussed principles, such as requirements for sufficient size of and space around equipment and the possibility for presenting information in a tactile manner, are not as relevant in this context, however most are highly applicable. These include considering equitable and flexible use, which leads to providing multiple modes of interaction. In the case of the Social Tutor software, this could encompass allowing the user to input responses via the keyboard, mouse, touch screen and speech depending on the task and user preference. These guidelines, along with the principle of perceptible information, suggest that not only should input be multimodal, but output should also be provided in a variety of ways, such as visually, textually and aurally. An example of this may be when asking a question the system provides the written text, an informative icon and reads it aloud. In the case of individuals with autism, providing multiple options for input and output and letting the learner choose which to utilise is important considering that some learners may have abnormally high or low sensory tolerance towards some options, and this notion is supported by outcomes from a recent small scale survey conducted by Fletcher-Watson et al. (2016). The Principles of Universal Design also stresses the importance of being tolerant to errors (Connell et al. 1997). For example in a virtual tutoring system, preventing users from accidentally clicking an irrelevant button by disabling it when not needed or allowing them to undo their last action and provide a different response can help increase tolerance to errors and reduce their likelihood in the first place.

Helal, Mokhtari and Abdulrazak (2008) provide guidelines for developing virtual companions, most of which can be applied successfully to developing pedagogical agents. They cite the following as important requirements: adaptability towards the environment and user, availability of multiple options for completing tasks, provision of useful and accurate solutions, proactive offering of services to the user, being tolerant to faults and unexpected inputs and having the ability to adapt its goals and behaviours to suit the user's needs.

Many of these points overlap with those suggested by the Principles of Universal Design (Connell et al. 1997) and Inclusive Design (University of Cambridge 2017). The first point, adaptability, is an interesting one as often virtual agents are considered adaptable if they learn from their user and alter their behaviour accordingly. Caution must be taken when applying this principle to individuals with autism who may display abnormal social and verbal behaviours. As the goal of the pedagogical agent in this case is to help minimise undesirable behaviours and maximise desirable ones, having the tutor adapt its behaviour to the user in this way is counterproductive. The more appropriate form of adaptability here is to gather data during interactions with the user and use this to determine their current needs, in turn allowing the system to present lesson content appropriate to the learner's current level and to cater to their learning style and sensory needs.

Tartaro and Cassell (2006) and Silver and Oakes (2001) provide guidelines specifically for developing software for individuals with autism, derived from their own experiences doing so. Both stress the importance of scaffolding. This involves providing the learner with very simple and straightforward learning experiences and tasks initially and, as the learner increases in competence, gradually adding complexity and distractions. Silver and Oakes (2001) note the importance of providing opportunities to repeat tasks in order to reinforce the concepts within them, and stress the need to provide timely and accurate feedback so that learners understand where they went wrong, why and what to do next time. Children with autism have difficulty learning from their own mistakes without explicit support, so such feedback is vital. Both research teams state that providing tasks that are inherently reinforcing and rewarding leads to the richest outcomes, and Tartaro and Cassell (2006) further this by stating that generalisation of behaviours to real situations and novel contexts must also be considered and supported as much as possible. Further, Tartaro and Cassell (2006) add that social skills interventions should provide a safe environment for children to practice their skills in, and that the use of roles can help children understand the dynamics and social conventions involved in social situations. Finally, they highlight that children with autism, just like their neurotypical peers, are all individuals and thus interventions should be customisable to their personal needs and skills.

More recently Fletcher-Watson et al. (2016) elicited recommendations directly from young children with autism themselves, their caregivers, and educators, prior to developing an iPad game. Their findings further support the inclusion of customisable features, minimising unnecessary images and background music to avoid fixation and distraction, inclusion of a reward token system, and having the system make no response when incorrect answers are given rather than a negative response. Input into the appearance and nature of characters and other visual elements was also covered, with some conflicting ideas around whether photo-realism to promote generalisation or cartoon animations to promote engagement would ultimately be more beneficial (Fletcher-Watson et al. 2016). After evaluation of the designed game the researchers found that individuals had differing reward preferences, and that while very young children were happy to continue playing the game even when it was very repetitive, more able children lost interest unless continually challenged. Families reported positive perceptions of the software generally (Fletcher-Watson et al. 2016).

In the case of virtual tutors, many of these recommendations can be easily implemented and much can be personalised and adapted to the child including, but not limited to, the appearance of the virtual tutor, input and output modes, and the content and format of the lessons provided.

2.6.2 Autism Specific Considerations

When designing for learners with autism, the traits and needs particular to these individuals must be thoughtfully catered for. As introduced in Section 2.1, communication challenges and sensory tolerance and integration issues are particularly important to take into consideration, as is the need to support generalisation of skills to other contexts.

Sensory Difficulties

As discussed earlier, sensory integration and tolerance issues have a major impact on individuals with autism (Robertson and Simmons 2013), and must be taken into consideration when designing interventions of any form for this learner group. Sensory overload is of particular concern, as individuals with low sensory tolerance may require only minimal exposure to particular stimuli before registering a strong response. To minimise the risk of sensory overload and thus make the tutoring software accessible for a wider range of learners, it is recommended to omit unnecessary material and avoid developing software that is aurally or visually 'noisy' (Clark and Choi 2005, Fletcher-Watson et al. 2016). In practical terms, this means avoiding the use of animations, bright colours, or sound effects unless they add significant educational value. Doing so helps keep the interface simple, making it easier for the learner to understand what is required, while minimising possible distractions or fixation (Davis et al. 2005, Fletcher-Watson et al. 2016). Some individuals with autism have low tactile tolerance which may make mouse and keyboard use challenging, some may find reading from the screen difficult due to literacy difficulties resulting from the communication deficit associated with autism, while others may have low aural tolerance, making a speech-recognition and text-to-speech interface confronting. Clearly, multiple input and output modes should be offered where possible so users can choose which best suits their needs, thus catering for this wide range of challenges and preferences. Additionally, many individuals with autism find looking at faces difficult, and one cause of this is thought to be the sheer amount of information, visual and social, that is contained in the human face (Jones et al. 2003). An animated pedagogical agent can be advantageous here, as it can be given a very simple and cartoonish appearance initially, and as the learner increases in confidence, the complexity can be increased. This is an example of scaffolding, which in a study by Parsons and Mitchell (2002) was shown to support generalisation of skills to real social situations.

Communication Impairment

Impairment in communication skills is typical for children with autism, and must be catered for when designing educational software for individuals on the spectrum. To support those with low reading ability resulting from this communication deficit, it is recommended to provide a visual prompt such as an icon along with any verbal or written information or instructions given (Quill 1997, Shane et al. 2009, Knight et

al. 2015). Icons should be simple and clear, and used wherever they add meaning, without being used excessively and contributing to sensory overload. Provision of multiple input and output modes is also important in the context of the communication difficulties, not just for sensory reasons. For example, expecting a learner with communication difficulties to write or speak full, grammatically correct sentences when their language skills are not the central focus of the lesson may serve to put learners off and draw attention away from the social skills concepts being taught. Instead, point and click interfaces and other simple interaction modes may be better suited to allowing students to express and explore their knowledge without exacerbating communication barriers.

Following this same line of thought, complex or lengthy sentences should be avoided in favour of short, concise sentences. Learners with autism often miss subtle cues and can become confused or distressed by ambiguity, so instructions should be presented in simple, clear steps and scaffolding used to move learners from simpler concepts to more complex ones as their skills improve (Parsons et al. 2000, Brown et al. 2001, Silver and Oakes 2001, Quirmbach et al. 2008). Self-paced lessons are ideal as they give the learner a sense of control and ownership of the learning process, lowering anxiety and helping with content retention.

Generalisation to Novel Contexts

As discussed earlier, difficulty generalising skills and knowledge to new situations is thought to be connected at least in part to the tendency towards repetitive behaviours and interests that is part of a diagnosis of autism (McCleery 2015). It is not unusual for participants to improve their skills at an intervention task, but fail to exhibit these same improvements in other situations.

Following on from the theoretical discussion of Subsection 2.1.5 including recommendations from the seminal article of Stokes and Baer (1977) and more recent review by McCleery (2015), a number of practical steps can be taken to encourage generalisation. First is to ensure that intervention tasks are embedded in real-world experiences and situations, so that their social value can be understood by the learner. Another is to expose the learner to a wide range of situations and media within the intervention tasks, the idea being that if they are able to generalise across these different situations and see the similarities and common cues between them, it will help with generalisation to contexts outside of the intervention environment. In the case of a software-based intervention, this may mean including videos, line drawings, animations, photos and other varieties of media rather than only exposing the learner to one media type. Additionally, with a virtual tutor who models facial expression and other gestures, the ability to change the appearance and voice of the tutor may be beneficial. This is similar to having a student role-play situations with a variety of peers instead of just their favoured peer. This approach may improve the chances of generalisation as it avoids having the learner associate the task with the single tutor presenting it, and helps them identify common elements across multiple appearances and situations. Additionally, presenting predictable tasks should not mean identical tasks, as learners with autism need to be gently encouraged to be flexible in their thinking. Instead, tasks should follow a predictable pattern, warn the learner before significant changes to the expected occur, but present some differences and alternatives each time they undergo the task.

2.6.3 Socially Responsive Agents

Studies have shown that deeper and more meaningful learning occurs in social contexts rather than when working alone (Krämer and Bente 2010). To take advantage of this finding, it is imperative to create a virtual tutor that is sufficiently realistic and relatable so that the student engages with it in a social manner, constructing knowledge collaboratively as with a human peer or tutor (Krämer and Bente 2010) and building rapport with the user (Zhao et al. 2016). It is likewise important to ensure that the virtual tutor does not impede or disrupt learning, for example interrupting with unwanted hints, as this can negate any positive effects the presence of a virtual tutor may have, and ultimately makes the educational experience much less enjoyable (Conati and Manske 2009).

Studies have shown that it is primarily the speech capability of a virtual agent that is responsible for motivating the learner and increasing the quality of their learning and problem solving skills (Krämer and Bente 2010). It was found that agents that could speak led to students having fewer difficulties and being more able to apply their skills to other contexts than when they used a text-based agent. One explanation for this is that it enables students with weak reading skills to engage with the content more easily, allowing them to focus their thinking on the concepts involved in the task rather than the mechanics of reading the material. Research regarding whether individuals with autism have a preference for computer generated voices or pre-recorded voice actors is mixed, with some suggestion that verbal children find computer generated voices "too synthetic" (Williams et al. 2004), others finding that non-verbal children actually perform better with computer generated voices than pre-recorded voice actors (Herring 2015), and some finding no significant difference between the two (Ramdoss et al. 2011). While pedagogical agents alone were not found to have a general impact on student motivation or learning, likeable agents did lead to some improvements. Interestingly, Tsiourti et al. (2016) found that for the older neurotypical adults in their study, having the virtual human mirror the emotional facial expressions of the user led to it being perceived as more likeable and persuasive. Whether this holds true for individuals with autism is unknown.

As Tsiourti et al. (2016) demonstrated, while speech is clearly important, an agent's nonverbal communication also has an impact on engagement. Schilbach et al (2006) demonstrated that virtual characters displaying social facial expressions, for example raising their eyebrows or smiling at the participant, caused the same brain regions to activate as they do in human-human interaction. In human-human interacion, nonverbal behaviour has several functions which may also be helpful in a virtual learning environment. For example, teachers model tasks for students, use illustrative gestures such as pointing, use gestures to emphasise important points and guide learner focus, as well as engage in dialogue management and turn-taking cues (Allmendinger 2010, Krämer and Bente 2010). It has been shown that smiling and other feedback cues affect student interest, motivation and learning outcomes, for example encouraging students to continue down a particular train of thought by smiling and nodding assists them to know they are progressing well (Allmendinger 2010, Krämer and Bente 2010). However, it is imperative that nonverbal behaviours appear sufficiently natural, as in human-human interaction they are processed automatically by the limbic system, and this may fail to occur if the behaviours appear odd (Krämer and Bente 2010). It is unknown if

these nonverbal cues will impact the target population of the current study given that difficulties with social interaction form a core part of a diagnosis of autism. While research into measuring rapport with a virtual tutor exists, it focusses on neurotypical users and the techniques have not been validated with individuals with autism, making them unsuitable for use in the current study.

2.6.4 Assistive Technologies

Following on from the notion that social contexts lead to more effective learning (Krämer and Bente 2010), and the prediction that therefore socially responsive agents will be more effective at teaching social skills, an investigation into just how to create such an agent is necessary. While there are a range of technologies which may potentially be of benefit when integrated into a pedagogical agent for teaching social skills to children with autism, to ensure the software is accessible to as many families as possible the Social Tutor must be able to be deployed on the technology already present in homes and schools. For this reason, gesture and speech recognition have some potential as they only require a web camera or microphone, both of which are commonly built into domestic laptops and desktops. These additional technologies could also be used to mitigate some of the other difficulties associated with autism, such as poor fine motor skills and sensory difficulties, as can careful design of the user interface.

Speech Recognition

Individuals on the autism spectrum often display poor motor skills, both fine and gross, making it difficult to perform clear and coordinated gestures (Noterdaeme et al. 2010). This can make using a mouse or keyboard quite challenging, since fine motor control is essential. Along with touch screen technology, speech recognition is also a possibility to assist with combating this. Speech recognition technology can also be used to provide novel socially-focussed educational activities, such as practicing speaking greetings aloud or directly interacting with a virtual human to practice conversational turn taking in a more realistic manner. While speech recognition potentially has a lot to offer a virtual tutoring system, it also brings with it many challenges in terms of implementation.

Speech recognition systems rely heavily on making predictions about what is likely to be said next. However, children with autism experience communication difficulties and may not follow social conventions, therefore saying things seemingly out of context. This provides an additional challenge for speech recognition systems that rely on matching sound input to a limited selection of expected patterns. One possibility to combat this is to structure activities in such a way that it is possible to present users with a range of answers to select from or where expected responses are very restricted, such as the approach taken in teaching sight words to children with autism by Saadatzi et al. (2017). This approach is only likely to be appropriate in some circumstances, and doing this reduces the realism of tasks such as practicing conversations, limiting its usefulness and applicability to many real-world situations.

Another challenge when using speech recognition software with individuals on the spectrum is that they often exhibit atypical speech properties due to their communication difficulties. A study by Hoque (2008)

which collected and analysed 100 minutes of one-on-one conversational speech between individuals with autism, down syndrome and their neurotypical teachers, found that there were notable differences between the groups. Examples include distinct differences in pitch, intensity and energy of utterances, with individuals with autism displaying less use of appropriate intonation. However, it is not just differences between these groups but also between the speech of children and adults, for example Jokisch et al. (2009) demonstrated differences in speech characteristics across age groups ranging from very young to the elderly.

Clearly the atypical speech characteristics of individuals with autism provide a challenge in terms of appropriately training and developing a model for speech recognition in such a tutoring system. One solution is to use audiovisual speech recognition instead of relying solely on audio information. Navarathna et al. (2010) developed a speaker independent automatic speech recognition system for use with GPS based navigators inside cars and found that in a noisy environment, the audio-visual approach provided higher accuracy and robustness compared to an audio-only approach, supporting the use of audio-visual data over audio-only data for speech recognition. The downside is that both a microphone and camera are then required to implement this technology.

Traditionally, the accuracy of a speech recognition system is based on the percentage of words that it correctly interprets. However, Hirschberg et al. (2004) suggest that for a tutoring application, this is inappropriate. Instead, the conceptual understanding of the speech recognition system is important, for example interpreting the utterance 'show me the trains' as 'show me trains' should not be considered an error. Hirschberg et al. (2004) suggest that inclusion of pragmatic, semantic, lexical and conceptual features may be used to provide more relevant accuracy measures. Rotaru and Litman (2006) investigated the impact of emotional speech on word recognition rates and found that emotional speech often leads to a high error rate, which in turn increases frustration and emotional content of the user's speech, causing a cycle. Rotaru and Litman (2006) suggest that being able to detect emotional speech and correct for it can help reduce speech recognition errors. Additionally, the authors suggest that in a tutoring application, a low threshold for speech recognition errors is likely to result in better outcomes for learners than a high threshold, as long as the tutoring systems implements a 'guiding' rather than 'punishing' approach to incorrect and ambiguous answers. In any tutoring application, special care must be taken to balance the risk of giving incorrect feedback versus the risk of unduly frustrating learners and reducing their engagement and motivation.

Clearly developing a speech recognition system for this purpose and learner population brings with it many additional challenges, on top of the well acknowledged existing challenges of dealing with multiple speakers, noise conditions, slang and filler words, and other scenarios likely to be encountered in a real-world application of the technology (Zeng et al. 2009). While speech recognition can be used successfully in some tasks with this learner group (Saadatzi et al. 2017) and the reliability of the technology is improving, it does not offer enough benefit to be a priority for implementation in the current research.

Gesture Recognition

Another possible input mode is gesture recognition. Gesture recognition brings with it its own challenges, and given the fine and gross motor skill difficulties common in children on the autism spectrum it is unlikely to be of sufficient benefit for the current application (Noterdaeme et al. 2010). Individuals with autism can find gestures and body language confusing as it is, and introducing a new set of gestures needed to interact with a tutoring system is likely to be an unnecessary challenge for learners. An additional consideration for alternative input modes is the environment in which the learner will be located. If the computer being used is located in a communal area or classroom, the learner is unlikely to feel comfortable interacting with the computer in a way that draws attention to themselves, particularly given the social difficulties many learners on the spectrum already experience. These alternative input modes may also be distracting to other students working in the same area. However, if the social skills being taught in the tutoring software are gesture-rich, then providing opportunities for learners to practice these skills and gain feedback from the system would be highly beneficial in assisting the learner to apply their skills to real-world situations successfully. Recent work has shown that gesture-based games such as those using a Kinect have the potential to support social skills development by providing opportunities for cooperative competition and emotional experiences (Ge and Fan 2017), however these are not goals of the current research.

User Interface Design

One simple way to cater for those with poor motor skills without drawing undue attention or distracting others working in the same area is simply through thoughtful user interface design. Avoiding interaction that requires typing or otherwise excessive keyboard use, and ensuring on-screen buttons are large enough to be forgiving when tapped, clicked or dragged, can go a long way to making the software accessible. Using this approach means the software can easily be paired with a touch screen or track pad if the user prefers, but remains comfortably usable when using a typical desktop computer with a standard mouse. While this alone does not harness the additional educational benefits of speech and gesture recognition, it does assist in making the software more widely accessible, as per the design principles already discussed (Connell et al. 1997, University of Cambridge 2017).

2.7 Existing Social Skills Curricula

Developing a curriculum for any purpose is a colossal task involving multiple steps including specification of scope and requirements, content development, testing, and validation of learning outcomes. As the focus of the current work is specifically to develop software, it was determined that also developing a curriculum would not be viable. Given this, it was necessary to evaluate existing curricula and select those that could be best adapted to a software context. Here a discussion of existing social skills curricula is presented, drawing on the preceding discussion regarding educational goals for the current application and known technological limitations, culminating in the selection of a small set of complimentary curricula which are implemented in the Social Tutor software.

2.7.1 Requirements for Software-Based Implementation

In order for a curriculum to be suitable for implementation in the Social Tutor, there are several features that are especially desirable and some elements that are difficult to translate into a software context. Inclusion of recommended scripts for the teacher or characters to speak, explicit step-by-step instructions relating to how particular skills should be performed, and a high level of visual material and digital media content such as worksheets, demonstration videos and songs, are all likely to make for engaging content and are straight-forward to translate into a software context with high fidelity, ensuring the validity of the curriculum is maintained. Depending on the level of detail given and the approach taken by the curriculum, role-play can also be a valuable feature that, given the unique nature of software including virtual humans, can to an extent be implemented in this context. At the very least, role-plays can be modelled between the virtual characters and then the learner encouraged to analyse and respond to what they observed. As discussed in Subsection 2.2.1, video modelling has demonstrated benefits for improving social skills in learners on the autism spectrum, and it is anticipated that observing virtual humans will be analogous to this. Likewise, curricula that implement and draw on other techniques that have demonstrated success for learners with autism are viewed favourably in the curricula selection process.

There are a number of features that are commonplace in existing social skills curricula but difficult to reliably implement in a software context given the limitations of current technology at the time of development. The main examples of such features include anything reliant on the educator observing and responding to the learner's behaviour in a natural setting, and open ended discussion or role-play. In a software context it is imperative to ensure that all possible interaction pathways can be responded to in a meaningful, or at the very least non-confusing and non-counterproductive, manner. While human-computer interaction technologies that could potentially be implemented to simulate these teacher facilitated activities are available, as discussed in Subsection 2.6.4 the technology in many cases is not yet reliable enough to ensure that the student is always provided with correct feedback, and therefore could unintentionally lead to reinforcement of undesirable social behaviours. Therefore, for the current research curricula that relied heavily on these techniques were avoided.

In addition to the practical considerations discussed above, some preference was given to curricula developed in Australia. It is anticipated that Australian content will be more likely to be in line with local teaching and therapeutic practices and wider curriculum content, and thus more likely to connect with and reinforce content that learners recruited for the software evaluation have already been exposed to, maximising the potential benefit for them.

2.7.2 Considered Curricula

Many curricula were assessed for their suitability for implementation in the social tutoring software, with recommendations from KidsMatter (2016a) guiding the initial search. Here some noteworthy examples are discussed, however the range of social skills curricula available is vast and ever growing, and the examples here represent only a small subset of the possibilities. Curricula that appeared promising but were ultimately

decided against due to their reliance on techniques that are difficult to implement in a software context or otherwise not aligning directly with the goals of this study include the 'Friendly Kids, Friendly Classrooms' curriculum (McGrath and Francey 1991), the 'I Can Problem Solve' curriculum (Shure 1993) and the 'SOS Social Skills in Our Schools' curriculum (Dunn 2005). In addition to these, the 'PATHS curriculum' (Kusché and Greenberg 1994) and the 'Think Social!' curriculum (Winner 2005) were identified as being potentially suitable for implementation in the Social Tutor, however due to a variety of barriers permission could not be obtained for their use at this time.

The 'Friendly Kids, Friendly Classrooms' curriculum (McGrath and Francey 1991) is an evidence-based program used in many Australian primary schools that aligns well with the desired themes to be addressed in the Social Tutor, covering cooperation and conversation skills amongst other pro-social behaviours. While recommended by both the Tasmanian and Western Australian education departments, at the time of curriculum selection there was only one study that formally investigated its effectiveness (KidsMatter 2016b). While that study did reveal positive results and continued use by schools reflects well on the program, the lack of formal evaluation coupled with it being designed for neurotypical children and not being evaluated specifically with children on the autism spectrum, made it unsuitable for the current study. Additionally, being designed for classroom use, it relies heavily on natural and guided interactions between students and is therefore challenging to implement electronically.

'I Can Problem Solve' (Shure 1993) is another curriculum widely used in Australian schools, although it was originally developed in the United States. It is based on the principles of Cognitive Behaviour Therapy and there is significant evidence of its effectiveness, including assessment after one year and two years of instruction (Boyle and Hassett-Walker 2008, KidsMatter 2016c). While the content does address social interactions and problem solving, it did not align closely enough with the specific topics to be addressed in the software, it did not include any explicit methods for assessing if the skills taught had been mastered by learners, and it has not been assessed with children on the autism spectrum.

The 'SOS Social Skills in Our Schools' program (Dunn 2005) was developed in the United States specifically for primary school aged children with pervasive developmental disorders, of which autism is a part, to participate in alongside their neurotypical peers and the wider school community. SOS provides detailed lesson plans, although some steps may be too broad for individuals on the spectrum without some prerequisite teaching first, for example suggesting a student "interrupt appropriately" is a skill in itself that must be taught. It should be noted that this is unlikely to be an issue in the intended classroom and community setting, but must be specifically addressed if implemented in a closed software setting. The pilot study for SOS was very promising, with children increasing their appropriate social behaviours after engaging with the program (Dunn 2005), however at the time of curriculum selection no independent evaluations of the program had been conducted, making it a promising but not yet established curriculum. Between this and the emphasis on techniques such as peer mentoring, group discussion and role-play that are

difficult to implement in a software context, it was determined that the SOS curriculum was not suitable for the current study.

The PATHS curriculum (Kusché and Greenberg 1994) is an evidence based program developed in the United States for primary school aged neurotypical children and also has demonstrated effectiveness for learners with special needs, including children with hearing impairment, behavioural difficulties, learning difficulties and gifted individuals. The program is underpinned by five conceptual models which address optimal developmental growth for each individual, the environmental impact that the teacher and classroom atmosphere have on development, learner neurobiology, psychodynamic education and emotional intelligence. There is strong evidence of its effectiveness with diverse learner groups, multiple settings and over long-term use of the program (Kam et al. 2004, KidsMatter 2016e). The curriculum content is organised into flexible, dynamic modules, and clear guidelines for educators are provided along with worksheets and visual materials, all of which make the curriculum suitable for potential adaptation to electronic format. Despite approach through multiple channels, permission to use the curriculum could not be obtained.

'Think Social!: A Social Thinking Curriculum for School-age Students' (Winner 2005) was developed in the United States for children ranging in grade from Kindergarten up to Grade 5. The program has a strong theoretical underpinning and multiple small-scale evaluations demonstrating positive outcomes for learners (Crooke et al. 2007, Kathy et al. 2009). The complementary comic-book style 'Superflex' materials present a narrative where social difficulties are 'bad guys' that can be 'defeated', making the program visually very engaging for young children and conceptually easy to follow. With newer materials also incorporating digital media aspects, the program is attractive for implementation in the social tutoring software. Unfortunately, permission could not be obtained to do so at this time.

2.7.3 Selected Curricula

In order to provide sufficient depth and breadth of content to learners, it was determined that a combination of sources was required, and ultimately three curricula were selected for this purpose. The 'Playing and Learning to Socialise' (PALS) curriculum (Cooper et al. 2003a) is aimed at kindergarten aged children and provides fundamental skills, while the 'Skillstreaming' (McGinnis and Goldstein 2012) and 'Social Decision Making/Social Problem Solving' (SDM/SPS) curricula (Butler and Poedubicky 2006) build on this foundation and provide more advanced instruction. A discussion of each chosen curriculum is provided here.

PALS: Playing and Learning to Socialise

The PALS: Playing and Learning to Socialise curriculum (Cooper et al. 2003a) is aimed at neurotypical kindergarten aged children and covers a set of highly relevant basic skills including greeting, listening, and taking turns. Skills are taught in an explicit way which is appropriate for learners with autism, and there are multiple studies supporting its effectiveness both with neurotypical children and those on the autism spectrum (Cooper et al. 2002, Teague 2005, James and Mellor 2007, Jones 2010, KidsMatter 2016d). The curriculum includes engaging elements such as puppets with scripts depicting role-playing between these

characters, explicit teacher scripts, digital media elements including modelling videos and songs, and suggestions for follow up activities, that makes it very suitable for implementation in an electronic context. Initial evaluation showed reduced problem behaviours and increased use of appropriate social skills in the children who participated in the 10 week PALS program, while there was no change in the control group (Cooper et al. 2002). The experiment was repeated using different social skills measures in a follow up study by both the original authors (Cooper et al. 2003b) and a separate research group (James and Mellor 2007), and in both cases the positive results were replicated. Additionally, in two separate research projects PALS has been used as part of a wider social skills curriculum, the Pathways to Prevention Project (Teague 2005) and the Learning, Enjoying, Growing and Support (LEGS) model (Hourihan and Hoban 2004). Teague (2005) conducted a large scale study with over 300 Australian preschool students and found that learners improved their self-regulation skills, social information processing skills and social competency after engaging with the program and that this was maintained 12 months later. Hourihan and Hoban (2004) specifically targeted rural families with children at risk of conduct disorder and combined PALS with both a parenting program, Incredible Years: Kids Challenge and Change, and a transition to school program. Amongst other benefits, it was found that children improved their social skills particularly in the areas of cooperation, interaction and independence, while also exhibiting a reduction in internalising behaviour. In more recent research in Ireland, PALS was evaluated with 90 children across 15 schools (Jones 2010). Parents and teachers were interviewed to evaluate perceived effectiveness of the program, and the children participated in pre- and post-intervention assessments. PALS was found to reduce some targeted negative behaviours and increase pro-social interactions in the children in the experimental group. Identification of feelings and development of empathy were areas identified as lacking in PALS (Jones 2010), however these are not currently a focus area for the social tutoring software. Overall, the PALS program is well supported experimentally and provides content that is suitable for adaptation to a software context, making it an appropriate choice for inclusion in the Social Tutor.

Skillstreaming

Skillstreaming is an approach that has been used in the United States with diverse purposes and populations for over twenty years (McGinnis and Goldstein 2012). Given its history, it is unsurprising that this program has a strong theoretical basis and is well established experimentally. Skillstreaming utilises modelling, role-playing, feedback and transfer, and draws on foundations of Vygotsky's Sociocultural Theory which suggests that, with the assistance of a more skilled adult or peer, students learn tasks within their own zone of proximal development, as previously discussed in Subsection 2.3.2 (Dowling 2010, McGinnis and Goldstein 2012). The version deemed most appropriate for implementation here, given that it is intended to follow on from the PALS program aimed at 4 to 6 year olds, is that aimed at 6 to 12 year olds titled 'Skillstreaming the Elementary School Child' (McGinnis and Goldstein 2012). This Skillstreaming manual contains explicit lesson plans including step by step instructions for a range of everyday social behaviours, such as the steps involved in beginning and ending conversations and introducing yourself to someone new. The manual allows lessons to be mixed and matched rather than needing to be done sequentially, which is well suited to

implementation in the social tutoring software. Modelling and role-play are core elements of the curriculum but are approached in such a way that they can be implemented using the virtual humans. Given its twenty year history many evaluation studies exist, however particularly noteworthy due to their inclusion of individuals with autism are the works by Kamps et al. (1992), Lopata et al. (2006) and Dowling (2010).

Kamps et al. (1992) implemented a combination of Skillstreaming and the ACCEPTS curriculum with three students on the autism spectrum with the goal of increasing peer interactions. They found that students increased their frequency of interaction, time they were engaged in peer interactions, duration of sustained interactions and their responsiveness to each other. The students had maintained these gains one month later during follow-up. While only a small-scale study, this is a very promising result as maintenance of skills is a known challenge for individuals with autism. Lopata et al. (2006) likewise used Skillstreaming as part of a larger curriculum, combined with cognitive behavioural therapy. In a preliminary study conducted with 21 males aged 6 to 13 years old, it was found that participants made significant improvements across the course of the intervention. Interestingly, parents reported significant improvement of their children's adaptability in changing tasks, sharing, adapting to altered routines and changes in the environment, and a decrease in odd behaviours, however teachers did not report any of these observations, possibly as they do not spend as much time with the participants in one-on-one settings in order to notice subtle differences and idiosyncrasies (Lopata et al. 2006). Dowling (2010) investigated the efficacy of Skillstreaming with children on the autism spectrum aged 7 to 12 years old. While significant improvements in social skills measures were achieved only for children aged 7 to 9 years old, observations of treatment group children of all ages revealed that they spent more time interacting with peers and less time in isolation post-intervention, along with improved eye contact, turn taking and cooperation, indicating that positive outcomes were achieved for a large percentage of children participating in the program (Dowling 2010). Overall, Skillstreaming is well established for use with the general population, has demonstrated effectiveness for children on the autism spectrum and provides appropriately structured content making it suitable for inclusion in the social tutoring program.

Social Decision Making/Social Problem Solving

Similar to Skillstreaming, the Social Decision Making/Social Problem Solving (SDM/SPS) approach has been around for many years, having been first developed in 1979, and is widely used in the United States for a variety of purposes and with individuals from a range of backgrounds including both neurotypical children and children with special needs (Elias and Butler 2005). SDM/SPS takes a constructivist approach, but also emphasises the needs for explicit skills training, a supportive educational environment, and draws on social-emotional learning theory (KidsMatter 2016f). It was determined that the manual aimed at Grades 2 to 3 would be most appropriate for the target learner group, with the content complementing both Skillstreaming and PALS, and addressing the desired topics. The SDM/SPS manual provides clear lesson plans with explicit instructions for teaching and, like Skillstreaming, activities are organised into self-contained topics that can be flexibly mix and matched, making it appropriate for adaptation to the Social Tutor. There are two large scale studies by the authors supporting the effectiveness of SDM/SPS. The first study was conducted with 109 boys aged 7 to 15 years old from special education classes, which found improvements in self-control, as

rated by teachers, along with reduced social isolation at two month follow up (Elias 1983). The second study was conducted with 158 students from grade 5 transitioning into middle school, and found that students who received SDM/SPS training were much better equipped to deal with the stressors that accompanied the transition, with 11 of the 28 stressors being reported as significantly less of a problem for the experimental group than for the children in the control group (Elias et al. 1986). It should be noted that both of these evaluations occurred over thirty years ago, both were limited to New Jersey and in both cases the curriculum developer was heavily involved in the study (Promising Practices Network 2016). For this reason, PALS and Skillstreaming are the primary sources of the social tutoring software content, with SDM/SPS being used to complement these and provide added depth and breadth.

2.8 Assessment Tools and Techniques

Assessment is a core feature of any educational program, be it a traditional classroom approach or within educational software such as the Social Tutor being developed here. To best meet the needs of learners assessment should be targeted, purposeful and ongoing, with assessment outcomes explicitly used to inform future learning activities. Here a range of assessment tools and techniques are discussed with two key purposes in mind: first, for use directly in the Social Tutor itself; and second, for use during the software evaluation phase. The software itself incorporates an ongoing, automated assessment and dynamic lesson sequencing system, so relevant techniques from existing virtual tutors and other computer-aided learning software are discussed. Additionally, the evaluation phase of the current research seeks to determine if behavioural changes occurred following use of the software, and as such a suitable behavioural assessment tool must be identified for this purpose. Tools and techniques for both purposes are discussed here.

2.8.1 Assessment within Software

In order to successfully meet the academic needs of learners, first what they already know must be accurately assessed and this information used to make an informed decision about what to teach them next. In the Social Tutor, this process must also be automated so it can be performed continually by the software. Four broad applications of assessment are addressed here – determining the topic sequence, i.e. what large-scale skills need to be taught; determining the method of instruction, i.e. how to teach these skills based on the individual's needs; determining the lesson content on a smaller scale, i.e. what tasks to present to the learner to help them improve at the current skill; and finally providing effective feedback.

Determining Broad Topic Sequence

The Social Tutor developed for the current research is based on a small set of established, experimentally supported social skills curriculum, and their content guides the higher level topic sequence that learners are provided with. The software is aimed at individual use on home and school desktop computers and therefore must provide self-contained practice and assessment opportunities that do not require a peer or parent to be present, so in this context observation of the child in natural social situations is not viable. Given this, one

option is for the learner to self-report, however while this has been shown to be accurate for purposes such as assessing anxiety and depression (Ozsivadjian et al. 2014) great care must be taken when using self-reports for assessing social skills, as discrepancies can exist between what a student knows they should do and what they actually do, and whether a difficulty stems from a skill deficit or a performance deficit greatly influences the educational tasks required to overcome it (Bellack 1983). Instead, opportunities for observing, interacting with and responding to virtual role-plays can be used to help increase the level of realism and consequently make differences between skill and performance deficits more easily detectable.

Another viable option is to allow educators and caregivers the ability to complete a social skills assessment for the learner in question when a new account in the software is created, and potentially even allow them to input updated information as it becomes available over time. There are many established social skills assessment tools in existence, with two of the most commonly used and recommended being the Matson Evaluation of Social Skills with Youngsters (MESSY) and the Social Skills Rating System (SSRS) (Wilkins 2010). In both MESSY and SSRS, evaluation items are presented in Likert-style scale and forms exist for the individual, parent and teachers to respond. This process can be automated, and thus incorporated into an autonomous social tutor. MESSY has sound psychometric properties and has been validated for use with individuals with autism, whereas SSRS exhibits some inconsistencies from test to retest and lower inter-rater reliability, thus use of the MESSY assessment tool appears preferable (Wilkins 2010). Both MESSY and SSRS are valid for use with primary and secondary school age children, and thus are applicable here.

The Behavioural Assessment of Social Interaction in Young Children (BASYC) is another tool that may be useful for high level assessment in the social tutoring application, as it is designed to be easy for teachers to administer and thus does not require psychology training to perform, and can be used for goal planning and monitoring existing social skills treatments (Gillis et al. 2010). It has been developed to meet the need for an objective measurement system for social behaviours and to inform intervention planning and monitoring, thus its goals marry with the requirements of this study on several levels. BASYC provides a list of interactions as a guide and a checklist of behaviours, so the influence of examiner subjectivity is minimised and the task of automating assessment is simplified. However, completing this assessment requires behavioural observation in naturalistic, semi-structured settings. While there is potential for BASYC to be adapted to a software environment where a virtual peer behaves as the examiner, experimental evaluation would be required to determine if the assessment maintained its validity in this context.

More recently Social Skills Q-Sort (SSQ) has been developed to screen for autism in a school based setting (Locke et al. 2013). It is designed for use by paraprofessionals and involves sorting a set of one hundred items into nine piles according to those that describe the child most to least. While a unique approach and appropriate for use by non-clinicians, its purpose is more about overall diagnostics and less about developing a profile of strengths and difficulties that could be used to inform the topic sequence of an intervention.

In the current implementation of the Social Tutor the selected curricula themselves are used to provide the broad topic sequence, and students simply choose which topic they feel will be most beneficial to them at the

time. In future iterations of the software when more of the curricula content is implemented, a standard test such as the MESSY or SSRS may be incorporated to ensure that students begin at the right difficulty level within the content, and to periodically check their progress and adjust learning activities accordingly.

Determining Method of Instruction

Educational experiences that mesh well with the current knowledge and learning style of the student are acknowledged to improve the processing of new knowledge, facilitate a deeper understanding of the content and generally expedite the learning process (Truong 2016). A number of factors influence the student learning style at any given moment, including their pre-existing preferences and their level of experience with the current concept (Truong 2016). For example, it has been shown that inexperienced and experienced learners display different needs, with inexperienced learners gaining more from following worked examples and experienced learners benefitting more from solving problems (Wittwer et al. 2010). It should be noted that care must be taken when considering the use of exploratory educational games such as the one used by Robison et al. (2009) as it has been shown that these typically only benefit students who already have the skills to gain knowledge from this style of task, whereas less skilled learners need more structure (Conati 2002). It is anticipated that learners with autism may need more structure due to their preference for sameness and difficulty learning without explicit explanations (Rapin and Tuchman 2008). Similarly, in terms of learning preferences it is often found that individuals with autism fare best with visual information over spoken instruction (Shane et al. 2009, Knight et al. 2015). While all learners are individuals, this aligns with the communication difficulties that are a core deficit of autism. An automated mechanism that can detect and implement the appropriate method of instruction for the student's current situation, much like human educators unconsciously do, would be a valuable component for an autonomous tutoring system.

Shute and Towle (2003) present a generic framework for intelligent tutoring systems that takes into account individual learner differences, the learner's current state of knowledge and best practices for instruction of the learner. It is based on Dick Snow's aptitude-treatment interaction (ATI) research, which aims to quantify and predict diverse learner profiles to allow for lesson presentation and content to be adapted to the learner's needs. Content presentation can range from step-by-step, highly structured instruction to exploratory presentation where the student has nearly complete control over the lesson sequence, with different presentation styles suiting different learning strategies. The three elements presented in the framework of Shute and Towle (2003) are the content model, learner model and instructional model, with these elements being used by an adaptive engine to determine what and how content should be presented. Shute and Towle (2003) then propose the use of Learning Objects (LOs) to facilitate flexible content presentation. LOs are small, self-contained and reusable components that can be combined into lesson sequences. Each LO should be limited to one of the three types of knowledge: basic knowledge, which includes facts and formulas; procedural knowledge, such as steps and skills; or conceptual knowledge, which covers understanding and theory. Sets of LOs that comprehensively teach a particular skill or knowledge set can then be defined, with relationships between LOs influencing the sequence that tasks are presented in. This framework provides much flexibility and provides inspiration for the structure of the Social Tutor developed here.

Determining Lesson Sequences

In order to present students with learning tasks suited to their current needs, it is essential to continually assess their state of knowledge. Research suggests that in order to be most effective, assessments should be integrated into the overall learning sequence rather than viewed as a separate activity, and used to continually inform and adjust the activities presented to learners (Black 2015). It is often seen that students learn how to complete a task or pass a topic without gaining any deep understanding of the topic material covered (Conati 2002). Providing opportunities for reflection on the processes and concepts involved, for example in self-explanation tasks, ensuring that any reward activities do not distract from the desired lesson outcomes, and implementing robust methods of assessing student knowledge all contribute to combating this issue.

Shute and Towle (2003) state that common methods of evaluating student mastery are insufficient, for example simply getting a particular percentage or a certain number of consecutive assessment tasks correct. Instead, Shute and Towle (2003) suggest the use of Bayesian inference networks (BINs) or student mental modelling to provide probabilistic values which can be used to determine gaps or misunderstandings in the learners' knowledge map. While a promising technique, unfortunately to successfully implement this requires tasks with open-ended or flexible answers, which in a software environment typically translates to writing paragraph-style answers. This is highly challenging for individuals with autism who experience both language and communication difficulties, making it an unsuitable approach for the current study. However, more recently de Klerk et al. (2016) have investigated the use of multimedia performance-based assessment, where users interact with a virtual lesson and the interaction data is fed into a BIN for assessment. This work emerged after the Social Tutor for the current research was developed but may be promising for incorporation in future iterations of the software.

In many autonomous tutoring applications, a common approach to judging students knowledge is to use latent semantic analysis (LSA) techniques to judge the semantic similarity of student responses to a provided 'ideal' response. This is the approach taken in the successful iSTART tutoring system discussed earlier (Jackson et al. 2010a). Hu and Xia (2010) also use latent semantic techniques in their automated assessment system and found no significant difference between the grades provided by their system and those provided by teachers, suggesting that this is an educationally valid technique. Latent semantic techniques such as these rely on comparisons between blocks of text, so again given that autism is characterised in part by impairment in communication and language skills, it is unreasonable to expect children using the system to be able to provide sufficiently lengthy and coherent written responses for LSA techniques to be applicable, making this unsuitable for the current study.

Meyer and Land (2010) recommend the use of speak aloud self-explanations as a reflective practice. Meta-cognitive skills and reflective practice, such as self-explanations, have been demonstrated to lead to better problem solving skills and the construction of deeper, more meaningful conceptual connections (Mitrovic 2001, Amico et al. 2015). Such meta-cognitive skills can be nurtured in students to help them improve their ability to learn. Amico et al. (2015) reviewed a year-long drama therapy course for developing social skills in

adolescents with autism and emphasised the benefits of reflective practice and having students explore the perspectives of other characters. Nicholas et al. (2015) found that for neurotypical young adults, recall of events that occurred in a virtual world could be enhanced by the use of highly detailed reminiscing involving open-ended questioning by a virtual partner. Mitrovic (2001) conducted a study with university level computer science students to evaluate their self-assessment capabilities. It was found that more able students displayed better understanding of their own educational needs, while less able students abandoned many more practice questions, often citing that the problem was too easy even when evidence suggested otherwise. This suggests that a system that prompts students to consider more carefully the reasons for their difficulties may help to nurture meta-cognitive skills and improve educational outcomes. Drawing on this, the Social Tutor includes two part 'homework' tasks where the first part asks students to plan their homework, and the second part asks them to reflect on how they went and why.

Black and William (2009) also emphasise the importance of reflective practice for deep and long term learning. They suggest that reflection can assist students to make the processes they unconsciously use explicit and concrete, making them easier to understand and implement in future. It is suggested that discussion with peers and others improves the outcomes of reflective practice, in following with Vygotsky's principle that ideas are initially constructed in social interactions, and then internalised by the learner (Black and Wiliam 2009). Additionally, challenging students to identify other situations where they can use the same thinking processes, to compare and contrast ideas, and to critically analyse their ideas, can help learners improve their problem-solving and cognitive skills in general and to apply their skills to other areas. While social learning may appear in conflict with the development of a social tutoring program to be used individually, the virtual agent can play the role of a peer and activate these same learning gains, as is attempted in the 'homework' activities of the Social Tutor described previously.

In addition to speak aloud self-explanation, Meyer and Land (2010) recommend conceptual mapping as a method of making misunderstandings and barriers to knowledge observable and hence manageable for educators, and a recent meta-review further supports the use of various graphic organisers for supporting individuals with autism to organise and express their knowledge effectively (Finnegan and Mazin 2016). Concept maps are particularly applicable to autonomous tutoring software as they can be automatically assessed, and existing work has shown that paper-based graphic organisers can lead to strong learning gains in children with high functioning autism (Roberts and Joiner 2007, Finnegan and Mazin 2016). When used in conjunction with peer group instruction, conceptual mapping has been shown to lead to improvements in social skills (Laushey et al. 2009), believed to be due to it being a very visual medium and thus making clear otherwise abstract ideas. Existing research by Kinchin et al (2000) suggests that concept maps allow educators to discover what students really know and how their knowledge is interconnected, rather than trying to make judgements and informed guesses, and emphasises the importance of synthesising and integrating ideas and concepts rather than simply repeating isolated facts. In light of this evidence, several concept map activities have been incorporated into the Social Tutor for assessment purposes.

As part of the learning process and formative assessment, concept maps can be created collaboratively between peers or between the learner and educator. For example, in the Social Tutor the virtual teacher can assist the student on request by providing a hint about a missing or incorrect link. Several concept map types exist, and the type used must be considered carefully in relation to the desired outcome and the target content, as no single dominant method currently exists (Park and Calvo 2008, Watson et al. 2016). Spontaneous maps can be challenging to automatically assess, as students are free to use any terms and connections they wish, however the richness of assessment can be highly beneficial, with map hierarchy indicating knowledge depth and interconnectedness of ideas (Kinchin et al. 2000, Park and Calvo 2008). The most simple concept maps may be in the form of 'fill in the blanks', and if terms to fit the blanks are provided, the task of assessment is further simplified (Park and Calvo 2008, Cline et al. 2010). Concept map format can fit anywhere between these two extremes, however care must be taken to ensure the task is sufficiently complex that the outcome is representative of the students' actual knowledge and not just 'good guessing' and yet assessable in a consistent and valid way. Depending on the map type, measures may include raw and weighted counts of connections, node and proposition matching, and measures of congruence and salience, i.e. proportion of valid student propositions over all criterion propositions and over all student propositions in the population, respectively (Park and Calvo 2008).

Cline et al (2010) developed an automated system for constructing and assessing concept maps known as the Concept Mapping Tool (CMT). The CMT is a web-based tool including GUI front ends for teachers to build criterion concept maps and for students to build their own maps, in the form of directed graphs, which are then compared to produce a grade. CMT uses a rule-based evaluation system to compare the nodes, direction of connections between nodes and other aspects of the map to determine a final grade. The system performs rapidly and thus students are given immediate feedback, which has been repeatedly demonstrated to be beneficial to the learning process (Cline et al. 2010). Students are presented with the central concept, concept nodes and distractor nodes based on the criterion map provided by the teacher, and are required to use these to demonstrate their knowledge by providing connections between appropriate concept nodes. This is highly structured, as students cannot provide their own terms for concepts, however it is also flexible as no hint is given to the student regarding the connections between the concept nodes and distractor nodes must be dealt with correctly as well. This is the approach mirrored in concept map activities within the Social Tutor, with students provided with all needed nodes plus a small number of distractor nodes, and their task is simply to connect them in a logical manner.

Theory of Mind (ToM) techniques are another suggestion for evaluating social awareness in virtual role-plays. In evaluation of the ‘Fear Not!’ educational program for constructively dealing with bullying, Hall et al. (2009) evaluated neurotypical children's social awareness through ToM questions. Children were presented with bullying scenarios acted out by virtual characters and following this were asked by the 'victim' character for advice. At the conclusion of the program, children were provided with a questionnaire asking them to judge how various characters felt at different points throughout the story. Questions were devised by experts in the field and asked learners to make inferences about mental states, emotions and

intentions of the characters. Students were asked a combination of short answer and multiple choice style questions, which were accompanied by visual prompts, such as screen shots, to help them remember the role-play. Hall et al. (2009) found this technique provided valuable insight into the children's social awareness of the presented situations, however application of this insight was not discussed. Assessing social awareness is a challenge as socially competent adults still often disagree on the interpretation of a social situation, thus there is often no definite distinction between 'right' and 'wrong' answers, rather answers fall on a continuum from less to more probable. This makes it particularly challenging to implement robustly in an automated manner as is required in this Social Tutor, and makes the implementation of a variety of techniques combined using heuristic rules and scoring thresholds a more viable approach for the current research.

Providing Effective Feedback

Feedback is an essential element of learning in any context. It has been shown that immediate feedback while a student is undertaking a task provides the most benefit and avoids situations where the student solidifies misconceptions rather than accurate understandings, presumably because the student is still engaged in thinking about the concepts and processes at hand (Stuart 2004, Bowman-Perrott et al. 2013, Crook and Sutherland 2017). However, determining how to provide feedback and what kind of feedback to give is of great importance. It was found that having a pedagogical agent interrupt students to provide hints provided no benefit, with experimental data indicating that students did not read the provided hints at all in these situations (Conati and Manske 2009). The content of feedback is likewise essential, as shown in the study by Hattie and Timperley (2007). It was found that simply providing praise, reward or punishment only had a small influence, while feedback suggesting how to perform a task better or containing information about the task lead to very significant gains. Feedback must cater to the student's immediate needs, with task-level feedback addressing misunderstandings about the task or the outcome, and process-related feedback assisting students to use their own error-detection strategies and to choose appropriate strategies to implement, and finally self-regulation feedback, helping students to monitor, determine and review their own practices (Hattie and Timperley 2007).

Black and William (2009) emphasise the need for ongoing assessment, as it provides three key functions: establishing what students know now, ascertaining what they need to know, and determining what to do to reach these goals. If this is done regularly, the educational process is managed such that the chances of misunderstandings, repetition of already mastered content, and other difficulties are minimised. Accurately assessing student needs means accurately determining the cause of difficulties that students are encountering. This could be for a range of reasons, including misunderstandings of the language used, the purpose of the task, or the task itself, being misled by an unimportant element of the task, using ineffective strategies, or simply not providing a clear or sufficiently detailed response (Black and Wiliam 2009). In many of these situations it is possible that the student does in fact have the targeted skills or knowledge mastered, but simply misunderstood what was required of them. By implementing ongoing assessment and feedback these difficulties can be detected and rectified in a timely manner, ensuring students do not waste time or inadvertently consolidate inaccurate knowledge or skills.

2.8.2 Assessment for Evaluation of Software

In addition to automatically analysing student progress during software use, for the purpose of this study it is also necessary to assess changes in student behaviour across the experimental period to evaluate what effect the software may be having on the student in their everyday life, and if they are able to generalise what they have learned to real-world contexts. Not only does student baseline behaviour and their behaviour immediately following the end of software use need to be assessed, but since both skill generalisation and maintenance are known difficulties for many existing social skills interventions, it is also vital to determine whether they maintain any behavioural changes over time.

Several factors need to be considered when selecting an appropriate tool for this purpose. First, the tool needs to be sensitive enough to detect subtle changes in the social skills of individuals with autism. Tools designed to assess social skills in neurotypical individuals or to diagnose autism without a fine-grained assessment of severity may not be sufficient, particularly given the short time period of software use, the small set of skills being targeted, and the intention of the software to be a supportive, complementary tool to work in conjunction with other interventions that the child may be participating in, rather than a standalone intervention. It is reasonable to expect that even if individuals benefit significantly from using the software, they will still exhibit similar severity of autism before and after software use. Other pragmatic considerations include that the behavioural assessment tool needs to be one that can be successfully utilised by the research team who have diverse backgrounds and not specifically for use by trained psychiatrists, ideally one that can be delivered electronically to reduce the burden on caregivers who must complete the assessment four times over the experimental period, and for this same reason, one where only the areas targeted by the software need to be administered to achieve a score for comparison. A number of standardised tests were considered, guided by recommendations from experienced colleagues in appropriate fields and drawn from those used in related published works. The assessment tools used in existing studies related to the three curricula selected for inclusion in the software were also considered, however many were restricted to use by psychiatrists only and thus were not suitable for this study. Discussion of some of the more widely used tools follows.

As noted in Subsection 2.8.1, two of the most commonly used social skills assessment tools are the Social Skills Rating System (SSRS) and the Matson Evaluation of Social Skills with Youngsters (MESSY). However, both the SRSS and MESSY were developed with neurotypical individuals in mind, rather than being specifically developed for individuals with autism (White et al. 2007, Matson et al. 2013). White et al. (2007) noted that because of this, the SSRS measures developing social skills in a broad sense and is not sensitive enough to pick up the subtle nuances in behaviour that are required for the purposes of the current study. Additionally, the SSRS has been shown to have some inconsistencies between raters and between test periods. While also developed for neurotypical individuals, the MESSY has been assessed for use with those on the spectrum and been shown to have high internal consistency, however when assessing test-retest consistency only a small sample size was used and, while results are promising, the authors note that the evaluation should be repeated with a larger sample (Matson et al. 2013). Furthermore, the MESSY caregiver report form has 64 items covering social skills as a broad area, which does not allow for fine grain evaluation

of the target areas in this study, namely greeting, conversation skills, listening and turn-taking, thus making it unsuitable for the current evaluation (Matson et al. 2013). The Behavioural Assessment of Social Inclusion in Young Children (BAYSC), also discussed in Subsection 2.8.1, is unsuitable for the same reason, as it has only 20 questions and is therefore better suited for broad behavioural assessments (Gillis et al. 2010). The Autism Social Skills Profile (ASSP) is another tool that appears promising, being a straightforward checklist that can be completed by parents and addressing many of the target areas this software aims to address (Bellini 2006). However, the ASSP has not had wide uptake by other researchers and assessments of the ASSP's reliability and validity have only been conducted by the original authors (Bellini 2007). While promising for future use, it was determined that a more established tool was required.

The assessment tool found to be most appropriate for use in the current software evaluation is the Vineland Adaptive Behaviour Scales, Second Edition also known as Vineland-II (Sparrow 2011). Vineland-II is derived from the first standardised adaptive behaviour test, the Vineland Social Maturity Scale (VSMS) developed in 1935, and is one of the most widely used adaptive behaviour assessments in the world (Sparrow 2011). It is intended to be very flexible in content addressed, being organised into the domains of Communication, Daily Living Skills, Socialization, Motor Skills, and Maladaptive Behaviour, and each domain being further broken down by subdomain (Sparrow et al. 2005a). For the purposes of this study, this means that the fine-grained assessment required to detect subtle differences in participant behaviour before and after software use is present, and it is also possible to be mindful of the burden placed on the caregivers in the study and administer only the domains containing content targeted by the software. Furthermore, Vineland-II is flexible in delivery with multiple surveys and rating forms targeted at both caregivers and teachers, allowing researchers and clinicians to select the tools most relevant to their needs (Sparrow et al. 2005b, Sparrow 2011). Vineland-II has been standardised on 3,695 individuals from birth to over 90 years of age, and has also been stratified based on socioeconomic background, ethnicity, and educational level. Importantly for the current study, data has also been collected for several clinical populations including individuals with autism, and has been found to be valid for use with this population (Carter et al. 1998, Perry et al. 2009, Sparrow 2011). Thus, Vineland-II is the behavioural assessment tool of choice for this research. After examining the individual items within the Vineland-II and aligning them with the content of the Social Tutor software, it was decided that the complete Socialization and Maladaptive Behaviours domains would be administered, along with the Receptive and Expressive subdomains of the Communication domain.

2.9 Conclusion

People living with autism face many challenges, not least of which is engaging in successful social interactions. The human social world can be very complex and unpredictable, with many unspoken rules, and can be difficult for individuals with autism to navigate due to their challenges with communication and social skills, as well as their strong preference for sameness. Furthermore, many individuals with autism report an enjoyment of technology, in part due to its predictability. The development of a virtual-human based social skills tutoring software hopes to harness this enjoyment to provide a non-threatening and

motivating educational stepping stone, assisting individuals with autism to improve their social understanding and engage in more successful social interactions with their peers and the world around them.

A range of factors contribute to the difficulties that individuals with autism have engaging in social situations, including challenges both using nonverbal communication themselves and understanding the nonverbal cues of others (Rapin and Tuchman 2008). Failure to develop a typical 'theory of mind' has also been noted as a possible factor in the difficulties those with autism face, with 'theory of mind' referring to the ability to recognise that other people have thoughts and feelings that are different from your own (Leslie 1987). Without this, understanding and predicting other people's motivations and responses can be incredibly challenging, leading to difficulties making friends and developing other appropriate relationships. This in turn can impact on individuals' wellbeing, leading to loneliness, isolation, social anxiety and depression (Bauminger and Kasari 2000). Where their neurotypical peers generally learn these social skills through their everyday experiences, individuals with autism often need to be explicitly taught. Thus, developing appropriate social skills is very important, with conversation skills, reading and responding to nonverbal cues, regulating and expressing emotions, and developing coping strategies for stressful situations being identified as being of particular importance (Rubin 2007).

Some of the current interventions commonly used to address these issues include Applied Behaviour Analysis (ABA) therapy (Lovaas 1987), Social Stories™ and Comic Strip Conversations (Gray 2001, Quirmbach et al. 2008), and video modelling (Marcus and Wilder 2009). There is strong evidence of efficacy for all of these, and many lessons can be learned from these interventions and applied in the Social Tutor software. However, these interventions also require significant time input by caregivers and health professionals and are not typically used independently by individuals with autism themselves. Software and hardware focussed interventions are gaining interest not only because of the affinity individuals with autism appear to have for technology, but also because they can often be used independently by learners. Examples include the use of robots (Huijnen et al. 2016), specialised wearable technology (Madsen et al. 2008) and specific social skills software including both virtual and augmented reality (Herrera et al. 2008, Cheng et al. 2015, Washington et al. 2016). These technologies are promising, with a number displaying evidence of generalisation from the intervention condition to real-world social interaction, which is a known challenge for interventions targeting individuals with autism (McCleery 2015), however many of these technologies still require further evaluation beyond initial pilot testing, or they require specialised equipment that may not be easily accessible to educators and caregivers.

Virtual humans appear to be a promising avenue that draws upon the benefits of technology-based interventions and possibly even video modelling, while not requiring any specialised equipment beyond a home computer or mobile device to interact with. Existing studies by Tartaro and Cassell (2008) and Bosseler and Massaro (2003) have provided the inspiration for the current work, as they have demonstrated educational gains and generalisation of skills to other contexts when using virtual agent-based software with children on the autism spectrum. Tartaro and Cassell (2008) focussed on social skills, however their virtual

peer requires input by the researcher for it to interact, and thus cannot be used independently by the learner. Bosseler and Massaro (2003) used an autonomous peer, but with a focus on increasing vocabulary and language skills. Combining the autonomy from the work of Bosseler and Massaro (2003) and the social skills educational experience from Tartaro and Cassell (2008), the ultimate goal of this work is to develop an autonomous virtual tutor for developing social skills in children with autism which can complement existing interventions and relieve some pressure from caregivers and educators, while providing a motivating and enjoyable educational experience.

To achieve this goal, first a small set of complementary social skills curricula based on valid psychological and educational research and appropriate for adaptation to a software context have been identified, namely PALS: Playing and Learning to Socialise (Cooper et al. 2003a), Skillstreaming (McGinnis and Goldstein 2012) and Social Decision Making/Social Problem Solving (Elias and Butler 2005). To dynamically respond to users, the system must make judgements about learner needs and contain mechanisms for providing the best sequence of learning experiences based on these judgements. Thus, valid and automatable knowledge assessment strategies have been reviewed, with particular consideration given to conceptual mapping. Existing research in computer science has provided guidelines for automatic assessment of concept maps and educational research has demonstrated improved social skill knowledge when conceptual mapping techniques are used with students with autism (Kinchin et al. 2000, Laushey et al. 2009, Cline et al. 2010). For this reason, conceptual mapping activities were recommended for inclusion in the Social Tutor. Similarly, drawing on the success of video modelling for teaching social skills to learners with autism (Marcus and Wilder 2009), the virtual characters have also been programmed to engage in role-plays, both demonstrative and interactive, to assist learners to identify both problem and pro-social behaviours.

Careful thought has likewise been given to the overall design of the software itself, with consideration given to minimising the impact of sensory difficulties (Clark and Choi 2005, Davis et al. 2005) and communication challenges (Shane et al. 2009). This is achieved by keeping the interface simple and including visual cues and reading supports. Increasing the likelihood of generalisation to novel contexts is also critical (McCleery 2015) such as by ensuring tasks are embedded in real-world situations and including variety in the tasks, images, videos and other media learners are exposed to. Finally, careful attention has been given to selection of the assessment approaches and tools used both within the software itself and during the evaluation of the software, with a simple heuristic-based automated assessment system being implemented in the Social Tutor software and the Vineland-II (Sparrow 2011) selected for use in the evaluation due to its flexibility, sufficiently fine-grained level of assessment, and alignment with the focus content of the Social Tutor.

CHAPTER 3. RESEARCH AIMS

Existing research suggests that autonomous (self-directed) virtual humans can be used successfully with children on the autism spectrum to improve their language skills, and that authorable (researcher controlled) virtual humans can be used to improve their social skills (Bosseler and Massaro 2003, Tartaro and Cassell 2008). This research explores the combination of these ideas to investigate the use of autonomous virtual agents in teaching the social skills required for taking part in successful conversations, building on an earlier study with similar goals but a smaller scale (Milne et al. 2009).

The core objectives of this research are as follows:

1. Design and implement an evidence-based Social Tutor software program that can be used by children with autism
2. Determine if knowledge of targeted social skills changes due to interaction with the Social Tutor
3. Determine if behaviour of targeted social skills changes due to interaction with the Social Tutor
4. Determine if any changes in knowledge or behaviour are maintained after software use ends
5. Determine participants' and caregivers' perceptions of the software

To meet objective one of this project, software has been developed that aims to facilitate the improvement of conversation-related social skills in children with autism. This software utilises virtual humans to explain and model pro-social behaviours including greeting, conversational turn-taking, and starting and ending conversations. It features a large collection of short, interactive lessons and dynamically responds to learner's educational needs by presenting lessons that are at an appropriate level for the learner, determined according to their past interactions with the software. Learners work at their own pace and typically complete two to three activities in a single session, with each activity taking approximately five minutes each. A snapshot of this software is shown in Figure 1, with further details about the software itself in Chapter 4.

To address objectives two to five a software evaluation was undertaken. Reviews on the efficacy of existing social skills interventions for children with autism identified a range of areas where experimental methodologies for evaluating these interventions were lacking, and one point highlighted was the absence of control groups (Rao et al. 2008, Neely et al. 2016). To address objectives two and three and enable changes to knowledge or behaviour that occurred following software use to be attributed to the Social Tutor, this evaluation included both an experimental group who received the genuine content designed to explicitly teach social skills and a control group who received game-like content containing no explicit teaching of social skills. Details of the differences between the software presented to the experimental group and the control group can be seen in Section 5.3. All participants were asked to use the software for the same amount of time and complete the same data collection tasks.

During the software evaluation, participants and caregivers completed a pre-test, and then participants were asked to use the software for 10-15 minutes a day, 3-5 days a week, for three weeks. Immediately at the end

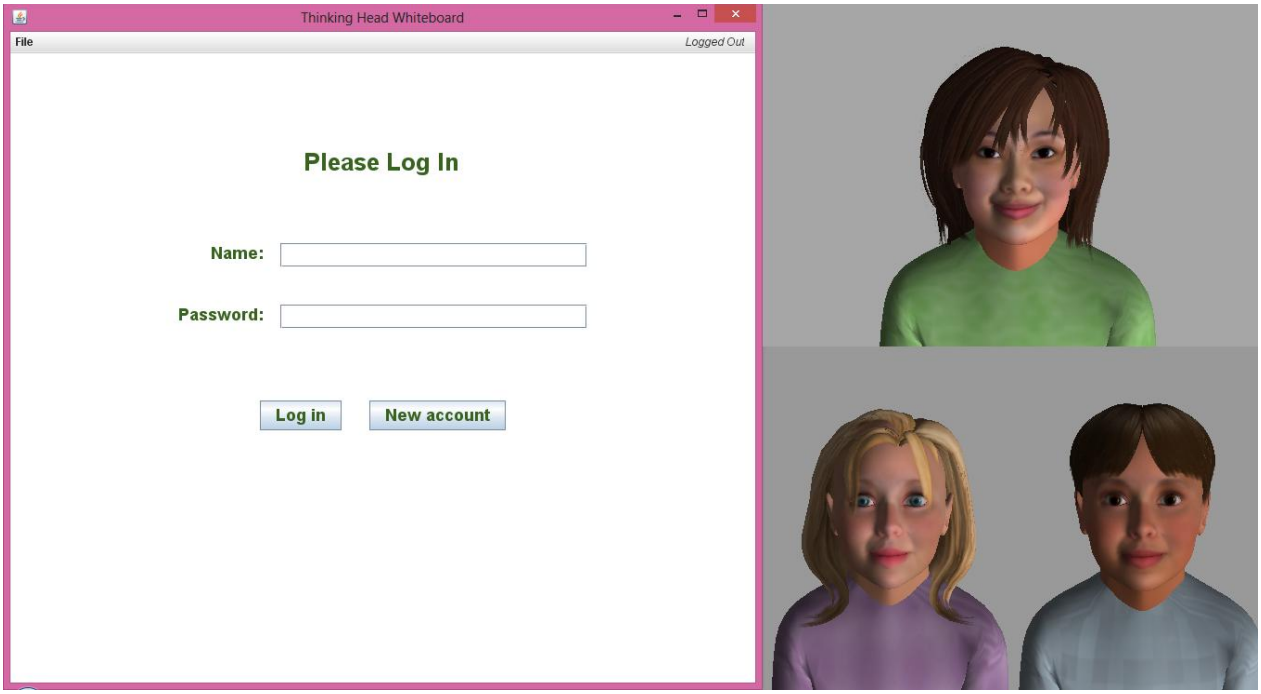


Figure 1: Social Tutor login screen
showing the virtual teacher "Kate" and two virtual students "Anna" and "Jack"

of the software use period, participants and caregivers were asked to complete a post-test. Participants were then asked to stop using the software. Finally, follow up post-tests were conducted at two and four months after the end of software use. Once the study had ended participants were free to recommence using the software if they desired, and were also able to access the version of the software used by the group they were not allocated to. A detailed discussion of the software evaluation procedure can be seen in Chapter 5.

To address objective two, at both pre-test and immediate post-test participants completed a content quiz directly assessing the social concepts being taught via the Social Tutor software. This quiz allowed participants to demonstrate their knowledge of the steps they had been taught for certain social skills, as well as demonstrate their knowledge of expected and acceptable behaviours in a range of relevant scenarios. The content quiz questions can be seen in Appendix A.

To address objective three, participants' caregivers were asked to complete the Socialization and Maladaptive Behaviours domains and the Receptive and Expressive subdomains from the Communication domain of the Vineland Adaptive Behaviour Scales (Vineland-II) to report on their child's current social behaviour. The Vineland-II was used to evaluate if caregivers observed any changes in their children's behaviour following software usage, i.e. if their knowledge generalised to other situations, a recognised challenge for autism interventions.

To address objective four, at both two and four months following the end of software use participants were asked to repeat the content quiz and their caregivers were asked to repeat the Vineland-II. These follow up

post-tests were used to assess if any changes present after using the software were maintained over time, a known challenge with interventions for individuals on the autism spectrum.

To address objective five, at pre-test participants were asked to complete a very brief questionnaire assessing their previous experience with computers and software, and their expectations of the Social Tutor. The pre-test questionnaire can be seen in Appendix B. At immediate post-test participants and caregivers were asked to complete another questionnaire rating and commenting on their experience with the Social Tutor and their suggestions for future development. The post-test questionnaire can be seen in Appendix C.

It was anticipated that participants would show improvements in content quiz scores from pre-test to immediate post-test, similar to that observed in the earlier study by Milne et al. (2009). It was also hypothesised that these gains would by and large be maintained at both the two and four month follow up post-tests, although some reduction was expected due to the lack of ongoing reinforcement. Further, as much attention has been given to promoting generalisation of the skills learned in the software to real-world situations, it was hypothesised that a modest improvement in social behaviour would be detected by the Vineland-II between pre-test and immediate post-test. Similarly to the content quiz, it is expected that these gains would reduce over time without ongoing reinforcement, but still be detectable at both the two and four month follow up post-tests. Given that this is the first time the novel Social Tutor software is being evaluated, the questionnaires are primarily exploratory in nature. However, effort has been made to ensure that the software is enjoyable and intuitive to use, and thus it is hypothesised that caregivers and participants from both groups will report a positive experience overall.

This multidisciplinary research is anticipated to contribute to the body of knowledge in both the autism and computer science fields. Broadly, it is anticipated that this research will contribute to the growing body of evidence regarding the use of automated tutoring systems for children on the autism spectrum, as well as for the use of these systems in teaching skills and concepts that do not have clear cut answers to compare with during automated assessment. More specifically, this research is expected to assist future software developers by providing insight into which features and aspects of interface design are effective and engaging for this population, particularly given the specific needs, strengths, and sensory challenges experienced by individuals on the autism spectrum. This research is also intended to provide ideas for future research regarding which features best support maintenance and generalisation given that these are known challenges when designing interventions of any kind of children with autism.

A detailed presentation of the results obtained and their analysis can be found in Chapter 6, with an in-depth discussion of these following in Chapter 7. Finally, a discussion of the implications of this research on developing educational software for children with autism and future recommendations for this area of investigation are addressed in Chapter 8.

CHAPTER 4. RESEARCH OBJECTIVE 1 - SOFTWARE IMPLEMENTATION

To address the first core objective of this research, namely to "design and implement an evidence-based Social Tutor software program that can be used by children with autism", desktop software involving three virtual humans and an interactive 'Whiteboard' component displaying content to users was developed. The interface was designed to be simple and intuitive so that study participants could use it without outside assistance. In this chapter an overview of the software architecture and the design choices behind its development are presented.

As discussed in Subsection 2.1.4, individuals with autism often experience sensory difficulties, be it difficulties with sensory integration or atypical sensory tolerance. For this reason the social tutoring software was designed to be visually simple with low sensory demand, as shown previously in Figure 1. As with most software, it was also designed to be intuitive to use so that learners could engage immediately with their learning. In the case of more complex elements or less obvious features, the virtual humans are available to guide students and take them through the necessary processes step by step. The software also has many unique features behind the scenes including automated assessment, dynamic lesson sequencing, a three tier rewards system, and XML lesson and curriculum authoring capabilities.

The Thinking Head Whiteboard software has been designed to support content development in a variety of topic areas, with particular focus on social skills for the current research, as well as literacy and language, and was designed for use by non-programmers and programmers alike. The discussion of the Whiteboard software architecture that follows is expanded from the previous publication by Milne et al. (2013).

4.1 Curriculum Implementation

As discussed in Subsection 2.7.3, three curricula that met the inclusion criteria and were deemed to complement and progress from one another were implemented in the social tutoring software. The 'Playing and Learning to Socialise' (PALS) curriculum (Cooper et al. 2003a), aimed at kindergarten aged learners, provides the basic starting points, while 'Skillstreaming' (McGinnis and Goldstein 2012) and 'Social Decision Making/Social Problem Solving' (SDM/SPS) provide more in-depth explanations and advanced activities that follow on from the PALS foundation. PALS also provides the basis for the virtual characters, with their puppet characters being used as the inspiration for the three virtual humans in the Social Tutor software, see Figure 2: PALS characters. In both PALS and the software developed for the current study, Kate is the teacher, Jack is a child with strong social skills, and Anna is a child with developing social skills.

Three core topics were addressed in the software: Greeting Others, Listening and Turn Taking, and Beginning, Ending and Maintaining Conversations. These topics were chosen because they are known issues for children on the autism spectrum, the behaviours required to perform these skills mesh well with the capabilities of the virtual humans being utilised and could be demonstrated and role-played suitably in the

software, and there was enough appropriate content across the three curricula to ensure that these topics could be effectively taught using the Social Tutor system. For 'Greeting Others' basic elements were drawn from PALS and advanced elements from the Skillstreaming curriculum; for 'Listening and Turn Taking' basic elements were again drawn from PALS and advanced elements from the SDM/SPS curriculum, and for the advanced topic 'Beginning, Ending and Maintaining Conversations' elements were drawn from both Skillstreaming and SDM/SPS. The implemented curricula provide the structure and core content of the topics to be taught, the steps within each skill being targeted, and the overall manner of teaching.

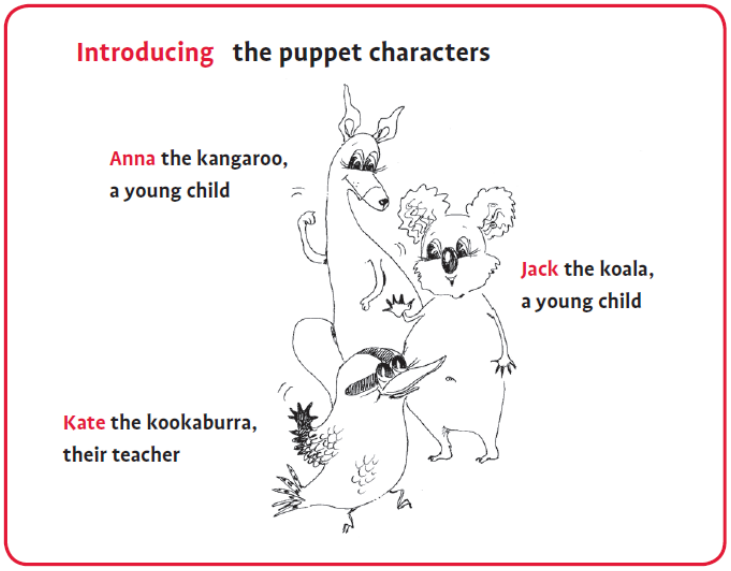


Figure 2: PALS characters (Cooper et al. 2003a)

The activities from the selected curricula were adapted into a software context with the goal of maintaining as much of the original appearance, structure, and wording as possible to maintain fidelity. This was relatively straightforward for worksheet-style activities as wording and images could be scanned or copied from the original documents and used directly, however as previously discussed, some teaching techniques such as group work, traditional role-play, observation and interaction with class peers, are not possible in this software context. Given this, to ensure students are offered enough repetition and practice, the materials drawn directly from the curricula have also been supplemented with a variety of activities created by the researcher that reinforce the target skills and steps. The one-off activities supplied in the curricula do not provide enough repetition for students to master the content, so multiple activities following the same template but with slight variations were also developed. The variations were partly to ensure learner engagement was maintained, but even more importantly to support generalisation of skills to novel contexts.

Additionally, there were a few areas where clarification and expansion were deemed appropriate. For example, in the "Asking Someone's Name" skill, how should a student respond if they ask someone their name and receive an unexpected response? The curricula do not explicitly cover this, and in a traditional learning context the student could simply ask the teacher or practice using peer role-play, however in the closed system of the Social Tutor software these areas need to be addressed in the included content. In cases such as this where information is not explicitly covered in the included curricula, more activities and

explanation have been added to proactively address any predicted issues. Additionally, students with autism benefit from having the purpose of particular skills and behaviours explained to them, for example why it is important to take turns, greet people or use good listening skills, therefore these explanations have been included not only as explicit short "objectives" lessons but also reinforced in other interactive tasks throughout each topic.

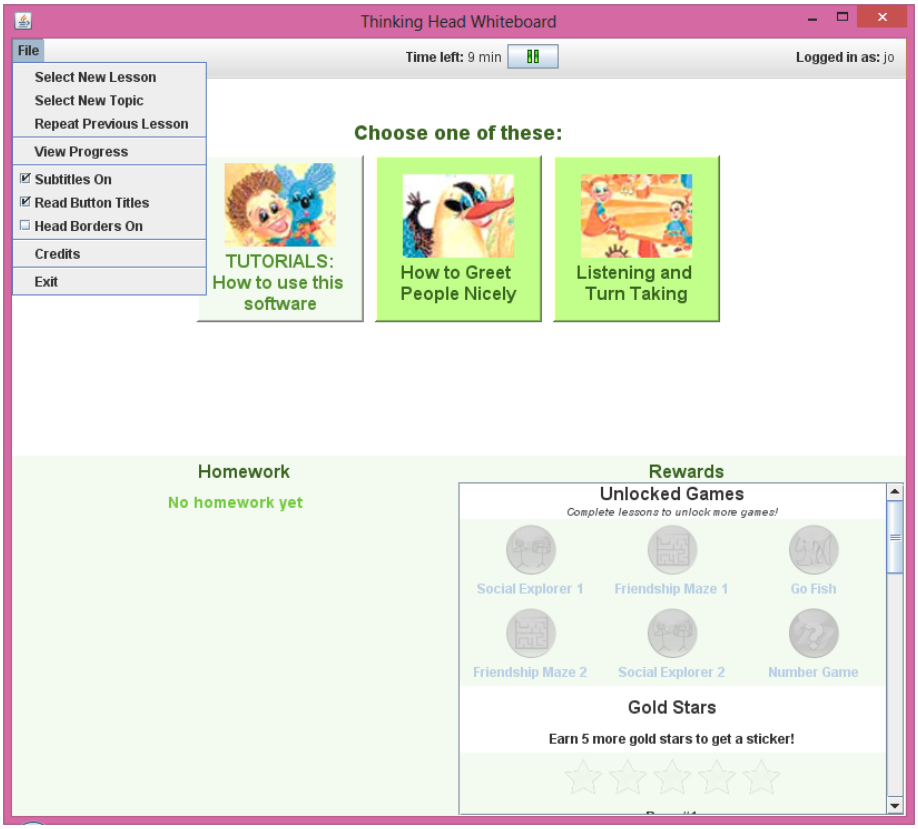


Figure 3: Screenshot of the Whiteboard topic selection screen for a new account

To ensure that the software is able to adapt to the learner's needs and provide them with choice in their learning, all content is divided into a hierarchical structure where the three topics of 'Greeting Others', 'Listening and Turn Taking', and 'Beginning, Ending and Maintaining Conversations' are further broken down first into objectives and then into short lessons. Students choose from the available topics when they open the software, and are then presented with up to three lessons within that topic to choose from. A screenshot from the Whiteboard topic selection screen can be seen in Figure 3. Each lesson is designed to take only a few minutes to complete so that learners can complete approximately three or four lessons each time they use the software. An overview of the in-software curriculum structure can be seen in Table 5. The number of lessons in each topic primarily reflects the content available from the three incorporated curricula, however where topics included skills considered to be more complex, additional lessons were added to provide learners' with sufficient explanation of the skills and greater opportunity to practice if needed. For example in the "Advanced Greeting" topic additional lessons have been created to provide students with instruction around what to do if your conversation partner responds in an unexpected way, which was not included in the original curricula but is a relatively common occurrence in real-world social interaction.

Table 5: Overview of activity structure indicating number of lessons in each category

Topic	Greeting Others		Listening and Turn Taking			Beginning, Ending and Maintaining Conversations *		
Objective	Basic Greeting	Advanced Greeting	Basic Listening	Advanced Listening	Turn Taking	Starting Conversations	Ending Conversations	Continuing Conversations
Core	10	26	13	12	20	11	4	14
Extra	5	10	5	6	9	1	1	8
Homework	1	4	1	1	0	2	0	2
Reward	2		2			2		

** Must complete "Listening and Turn Taking" topic to unlock this topic*

Topics and lessons can have prerequisite tasks specified and will not be available to the learner until those prerequisites are complete, for example students cannot access the topic on having good conversations until they have completed the topic on listening and taking turns. Each lesson has a minimum average correctness and accuracy that must be met in order for the lesson to be considered complete. This varies according to the lesson but is typically 80% correctness and 60% accuracy. Likewise, each objective has a minimum average correctness and accuracy that must be met, in this case all objectives require 80% for both, and all objectives must be complete for the topic to be considered complete overall.

Correctness and accuracy are calculated differently where necessary according to the lesson type, with correctness representing completeness of an activity (e.g. percentage of a video that was watched) or number of correct answers out of total answers (e.g. how many responses were correctly sorted into the right categories) and accuracy indicating of how many incorrect answers were tried before the task reached completion (e.g. how many responses were placed into the wrong category before they were moved into the correct one). Most lessons are designed to guide the student towards reaching 100% correctness before they leave a task, and accuracy is not applicable for some activities such as watching a video or observing the virtual people model a social skill, therefore combining correctness and accuracy provides a better picture of the learner's mastery of a particular task. For more detail regarding lesson sequencing see Subsection 4.6.3 Automated Assessment.

As can be seen in Table 5: Overview of activity structure, lessons are classified into several categories. Core lessons must be completed for the objective and topic they are associated with to be considered complete. In contrast, 'extra' lessons do not even need to be attempted but are provided for students who need additional support or repetition in a particular area. 'Homework' lessons are likewise optional and do not need to be completed or even attempted, but are encouraged via the virtual teacher Kate reminding students that homework exists just before they log out, and encouraging students to complete their homework reports when they log in. Homework is based on the format and content from the Skillstreaming curriculum and is designed to encourage students to practice their newly learned skills in the real world, be it with family or peers at school.

Finally, each topic has two 'reward' lessons associated with it - the first is unlocked and made accessible to the student when they reach 50% completion of a topic overall, and the second is unlocked at 100% completion. Reward lessons are designed to reinforce the social content being taught in the main lessons but in a more game-like manner. In addition to the lesson counts listed in the Table 5 there are a set of prerequisite lessons that are not explicitly listed but are offered to students and must be completed before the core lessons are presented. The majority of these are informational, for example explaining the objectives of a topic and the purpose of the skills being taught. Again, for more information regarding the lesson sequencing process see Subsection 4.6.3 Automated Assessment.

4.2 Architecture Overview

The social tutoring software is composed of two standalone programs, the virtual human software ‘Head X’ and the lesson interaction and display software 'Thinking Head Whiteboard', as shown in Figure 1 previously. Three instances of Head X are used in the Social Tutor, one to display the teacher character 'Kate' and one each to display the child characters 'Anna' and 'Jack'. As mentioned previously, the three characters are based on the puppet characters used in the PALS curriculum.

One challenge with existing virtual character software is that it is typically limited to a single persona and is not easily customisable; however, Head X was developed specifically to allow for a wide range of customisation in both the appearance of the virtual human and the associated synthetic voice. Head X was developed prior to the current study and was utilised as-is in the current research, for further details of its technical implementation see Luerssen and Lewis (2009) and Milne et al. (2011). The Thinking Head Whiteboard was developed specifically for the current research and controls the associated Head X instances via means of the Synapse interface that accompanies Head X. Synapse provides the ability to use memory-mapped files to efficiently synchronise and share data between multiple processes, and is designed to allow external programs written in Java, such as the Whiteboard component of the Social Tutor, as well as C-family languages, to communicate with and control Head X instances. An overview of the Social Tutor software architecture can be seen in Figure 4.

In addition to controlling the Head X instances, the Whiteboard software reads in and interprets XML curriculum and lesson files and external media files in order to display the interactive content, and then tracks learner interactions with the system so that their progress can be automatically assessed and an appropriate lesson sequence dynamically presented. Learner progress is then stored in XML log files which, for the purpose of data collection for the current study, are then automatically uploaded to a secure server. An example of a summary log file is shown in Figure 5.

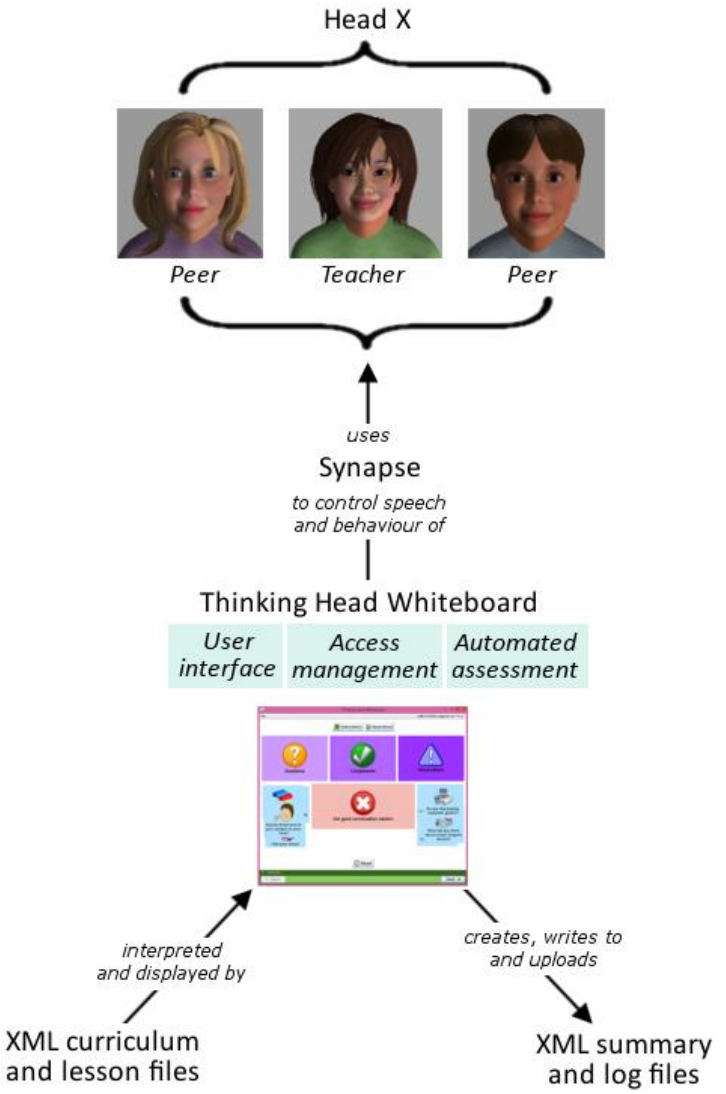


Figure 4: Social Tutor architecture overview

4.3 Head X Virtual Agent

The virtual human software Head X can generate dynamic, realistic speech on the fly and model realistic facial expressions, as demonstrated in Figure 6. This ability to model facial expressions and emotions makes it particularly useful for teaching social skills to children with autism and other social impairments. The capacity of Head X to model realistic facial shapes during speech similar, to Baldi and Timo (Bosseler and Massaro 2003), may help support individuals with the auditory processing difficulties that can be a co-morbid condition of autism, and may even have potential to benefit children with hearing impairments.

While a single virtual person can behave as a tutor and guide learners through their activities, it is also possible to display multiple Head X instances simultaneously, enabling educators to write lessons demonstrating interactions between individuals. In the case of the Social Tutor software this is invaluable, as it allows ideal pro-social behaviours to be demonstrated for learners and facilitates 'interactive role-plays' where students choose the behaviour that one of the virtual characters will display, and then observe the

response it elicits from the other characters. In this way learners can explore both desirable and non-desirable social behaviours and gain a better understanding of not only what behaviours are appropriate in certain situations, but also why they are important, and what impact their behaviour can have on other people.

The Head X software can be set to display subtitles, allowing for it to be used effectively even if the computer sound output is turned off as is sometimes the case for classroom computers, and catering to learners both with low sensory tolerance for sound and with auditory processing difficulties, along with those learners who simply do not enjoy listening to the synthetic voices. For those with low sensory tolerance towards visual input or those who do not have a sufficiently high reading ability to make them useful, subtitles can be turned off to simplify the display. As can be seen in Figure 3, learners can toggle subtitles using the File menu of the Social Tutor software. Their preference will be saved to their student file and remembered when they next log into the software.

```
<curriculum_summary id="PALS">
    <lesson_interaction accuracy="100%" correctness="100%"
        duration="0:0:0:32" id="0_Welcome"
        origin="action-parser"/>

    <lesson_interaction accuracy="100%" correctness="100%"
        duration="0:0:2:9" id="1-1_Greeting_Story1-1"
        origin="action-parser"/>

    <lesson_interaction accuracy="80%" correctness="80%"
        duration="0:0:2:6" id="1_Greeting_Activity1"
        origin="action-parser"/>

    <lesson_interaction accuracy="50%" correctness="50%"
        duration="0:0:1:44" id="1_Greeting_Face"
        origin="display-lesson-options"/>
</curriculum_summary>
```

Figure 5: An extract from a user's summary XML file

For further customisation, Head X can utilise head, hair, shoulder and other models created using external modelling programs such as FaceGen or Blender. While not implemented in the current research, for individuals with low tolerance towards visual input, the software could be customised to start with a very basic and less realistic virtual human face and gradually increase the realism and complexity as the student becomes comfortable and confident with the simpler versions. Head X can also be customised in terms of the appearance of hair shape and colour, and other accessory models including eye glasses and shoulders. All of these can be created in modelling programs such as Blender then placed and controlled via the Head X configuration files. Head X also comes with a range of pre-existing models for hair and accessories, making it possible for educators to create their own unique characters for use in the Social Tutor even without having 3D modelling skills themselves.

Figure 6: Head X can model many facial expressions including happiness, sadness, anger, surprise and disgust

For this research the three characters shown in Figure 3 were created and maintained for consistency throughout the study. The character models were built using FaceGen Modeller and are intended to be human versions of the three characters used in the PALS curriculum. The models were also intended to be both relatable for the target audience in terms of age and somewhat culturally diverse, however it should be noted that FaceGen is designed to produce adult faces and so creating suitable child faces was somewhat challenging and limited in scope, particularly given the importance of creating virtual characters' whose appearance is sufficiently realistic to support generalisation and potentially harness the benefits of video modelling, without falling prey to the long recognised phenomenon of the Uncanny Valley (Mori 1970). The Uncanny Valley applies to any anthropomorphic character, common examples being robots, computer animated characters in movies, and virtual avatars. It has been shown that as these characters become more human-like, observers respond to them increasingly positively, up to a point where a sudden drop occurs and the characters are instead perceived as both too human and not human enough, instead eliciting feelings of eeriness and discomfort. Mori (1970) suggested that this may be a biological response connected to human perceptions of death and disease, with a corpse falling into the deepest pit of the Uncanny Valley. Work on identifying precisely what factors elicit these reactions is ongoing, with the goal of supporting the development of humanoid robots and virtual characters that are well accepted and perceived positively by users (Ho and MacDorman 2017).

Some samples from the face creation process can be seen in Table 6. As can be seen in Table 6, attempts were made to create more explicitly culturally diverse characters however there were some difficulties sourcing suitable accompanying hair models and the final character models shown were deemed to be sufficiently representative of the local target audience for the immediate study. Allowing students to choose or customise their own virtual characters would be a desirable feature to add to the software in future, as it has the potential to increase engagement, and existing research has shown that having control over the appearance of a virtual avatar not only increases motivation and improves the user experience, but can also lead to improvements in skill performance (Mei et al. 2015).

4.4 Lesson Interaction and Display

The activity display component of the Thinking Head Whiteboard is written primarily in Java, with a small number of libraries written using C++. The first of these libraries enables basic speech recognition functionality to be used within the software, taking advantage of the speech recognition system built into the Windows operating system. The second library allows the Java component to control displayed instances of Head X via the Synapse interface as previously mentioned. All resources required for displaying an activity, including the XML curriculum and lesson activity definition files as well as additional resources such as images and videos, are organised in a strictly structured directory hierarchy, as can be seen in Figure 7.

This allows content developers to easily place and locate resource files and reference them within their XML lesson definition files. Multiple curricula can be supported simultaneously simply by placing the curriculum file for each in the parent 'input' directory, then creating a subdirectory with the same name to contain all the associated XML lesson definition files. This allows multiple learning areas to be accessible at once, for example both a literacy program and a social skills program could be loaded and the learner would be able to log in to separate accounts depending on which area they wish to focus on in a given session (see 4.6 for more on the curriculum access process). For the current study, this means that all participants were given a copy of both the experimental group activities and the control group activities but were only given access details for their assigned group initially, with access details for the alternative curriculum provided once data collection for the current study ended.

Table 6: Current, original and sample FaceGen models

Current faces	
Original faces	
Samples of other culturally diverse child faces	

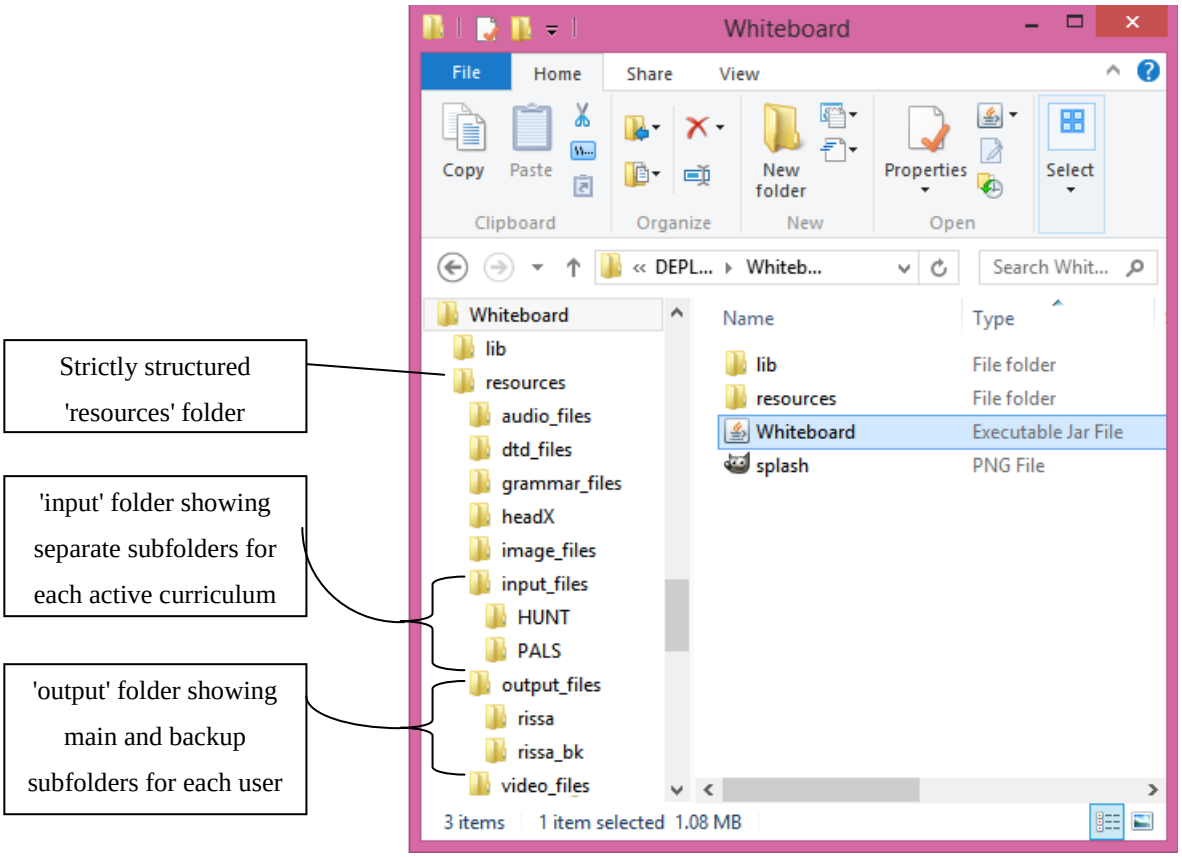


Figure 7: Directory structure of the Social Tutoring software

4.4.1 Lesson Authoring Process

Initial development of the Whiteboard was undertaken in close consultation with an educator with extensive teaching experience but no experience in software development. This guided ongoing development and assisted us to make the software usable for other educators with similar backgrounds, while also simplifying and speeding up the lesson authoring process for the current study. Lesson files and curriculum sequences are defined using relatively simple XML files which are read in and interpreted by the Whiteboard software. To facilitate this, lesson and curriculum definition files must adhere to the structure specified in the document type definition provided with the Thinking Head Whiteboard.

While this may appear daunting at first for non-programmers, most elements have usable default values so only very simple XML files are necessary for basic functionality, such as that seen in Figure 8. This XML demonstrates a lesson with a single page, a simple introductory action where the Head X teacher speaks to the learner, and a basic drag and drop activity, the output of which can be seen in Figure 9. While this example has only a single page, any number of pages each containing any of a variety of different activity types can be specified. To facilitate simple and convenient creation, each activity type has its own XML category. A list of currently supported activity types can be seen in Section 4.5 Lesson Types. A high level of control over the appearance and behaviour of the lesson activity and the virtual humans is possible, including customising what should occur when the learner completes an activity or performs part of a task correctly or incorrectly, although naturally this requires more complex XML lesson definition files to be created by the educator.

```
<lesson title="Drag and Drop Demo" id="test" prerequisites="" icon="">

    <page id="1">
        <action>
            <head_speak>
                Here is an activity for you to complete, [student_name]!
            </head_speak>
        </action>

        <drag_and_drop>
            <draggable_object id="obj" colour="yellow">
                <text>Object</text>
                <location>centre, 10%</location>
                <size>100, 100</size>
            </draggable_object>

            <drop_box id="myBox" colour="green">
                <draggables>obj</draggables>
                <text>Drop Box</text>
                <location>220, 250</location>
                <size>300, 300</size>
            </drop_box>
        </drag_and_drop>
    </page>

</lesson>
```

Figure 8: A basic XML lesson definition file

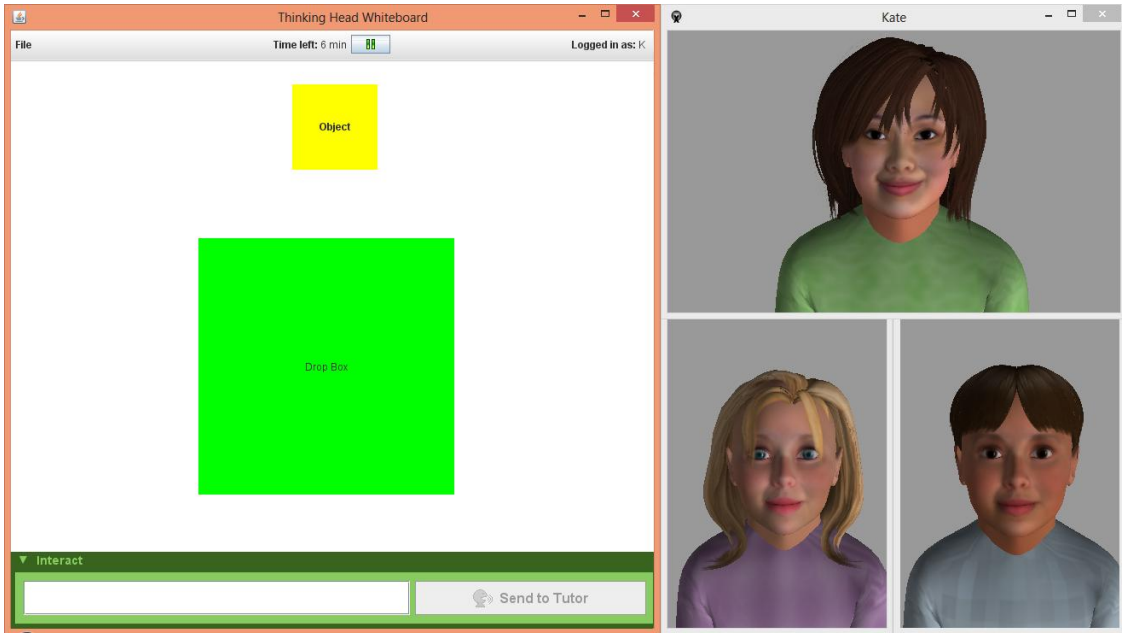


Figure 9: The output of the definition file shown in Figure 8

Topic and lesson buttons such as those seen in Figure 3 are automatically generated from the curriculum XML file and the contents of the 'input' folder. The activity types that can be specified in these lesson files are designed to have robust default behaviour while also allowing for a large range of customisation options, making them both simple and powerful to use. Lesson file content is interpreted and displayed alongside the virtual humans as interactive activities, while the curriculum sequences are used to guide the automated assessment system, helping to determine which tasks get presented to learners and when. While not utilised in the current study, the system makes it possible for caregivers and educators to modify the provided XML files, for example replacing default images with custom images or photographs that will be more engaging for their particular learner, modifying content to more closely align with strategies that their learners are addressing in other school or therapist based interventions, or write their own activities entirely.

Within the XML lesson files, educators are able to specify which virtual human should respond to particular events and what they will say or do, for example when a user completes a task correctly the educator may wish to override the default response with something more specific or meaningful for their learner. The Thinking Head Whiteboard reads these commands in from the XML files and then controls the actions of the displayed virtual humans via the Synapse interface. Educators can define what the virtual people say, their facial expressions, and a range of basic actions such as turning to face each other or the Whiteboard window or 'walking' off screen.

To further assist in the process of making the virtual character dialogue feel more natural, a set of placeholders have been set aside that are dynamically replaced with the correct value by the Whiteboard software. These can be seen in Table 7. Anecdotally, it has been observed in previous work that having the virtual character refer to the learner by name, even with a simple greeting such as "Hi Ben!", can leave a powerful first impression and have a desirable effect on the user's willingness to engage with the virtual characters (Milne et al. 2009). Similarly, if the virtual characters become too repetitive they seem less human and less likable. These placeholders aim to address these areas. It should be noted that these placeholders can be used not only in speech commands for the virtual characters, but also in lesson titles and topic titles.

4.4.2 Lesson Design Guidelines

The Thinking Head Whiteboard allows for free form lesson design, however there are some recommendations that educators using the software may like to consider, particularly if developing lessons for children with additional needs or sensory challenges. While these recommendations draw considerably from existing literature outlining the experience of previous research teams, they have been compiled by the researcher specifically around the features and capabilities of the Social Tutor software. These guidelines have been adhered to throughout the development of lesson activities for the current study.

Table 7: Placeholders and their purposes

Placeholder	Purpose
[student_name]	Replaced with the name of the learner. The name is derived from the alphabetical component of the account name, for example if the account name was "Josie", "Josie13" or "Jo1sie3" the virtual characters would refer to the user as "Josie".
[student1_name] [student2_name] [teacher_name]	Replaced with the names of the virtual characters. To allow for customisation of the virtual characters' names, hard coding their names into lesson files is discouraged.
[positivemessage] [oopsmessage]	Replaced with a random selection of either positive or negative utterances, such as "Great!" "Well done!" or "Awesome!" for the positive message or "Oops!" "Uh oh!" or "Sorry!" for the negative message.
[this_label] [this_box] [correct_box]	Replaced with the name of the current 'draggable' label, the box it has just been dropped on, or the box it should have been dropped on. Specific to activity types with a drag and drop element, such as word grids, concept maps and drag and drops.

The most basic recommendation is to take advantage of the built in functionality supporting the addition of images and icons. Most elements, from labels and buttons to the ‘draggable labels’ in the drag and drop activity demonstrated in Figure 9, support the display of icons in addition to or in place of text. It is also possible to set an icon for each lesson and topic, as shown in Figure 3. Making lessons highly visual is not only intended to make them more appealing, but also to make it easier for young pre-readers and children with developmental language deficits to understand what their choices are and the meanings behind each element. Without this it could be difficult for these learners to use the software as independently as intended. Along the same lines, it is highly recommended that instructions are kept concise and use simple language. This benefits all learners as clear instructions reduce confusion and frustration. In particular, individuals with autism can struggle to process and follow complex instructions (Silver and Oakes 2001), making this recommendation particularly important for this learner group. Additionally, some people find that listening to synthetic voices can be less enjoyable than listening to real human voices and so keeping blocks of speech short helps to both combat this and make any idiosyncrasies of the synthetic voice less obvious.

While the presentation of activities is important to ensure both understanding and engagement, assessment tasks should also be carefully chosen. As discussed in Subsection 2.8.1, assessment tasks should match the content being taught to ensure that learner mastery can be adequately determined. Much research supports the use of concept maps for this purpose, and hence they have been included in the Thinking Head Whiteboard activity types. Concept maps are considered to be valuable as they allow learners to demonstrate their understanding of the connections between concepts, rather than simply demonstrating recall of remembered facts (Park and Calvo 2008). With the aim of assessing learners' deeper understanding of the material being taught, every objective in the social tutoring curriculum contains a minimum of one concept map style activity, with most objectives containing two or three. The format of the implemented concept map activity type takes inspiration from Cline et al. (2010), where nodes are provided for learners to move around and connect together, including some distractor nodes, and visual hints as to the node hierarchy are also

provided via the default colouring of the nodes and the font on central nodes being bigger and bolder than end nodes.

For individuals with autism the ability to generalise newly learned skills and knowledge to novel contexts is a recognised challenge (McCleery 2015). To support these learners in particular, but to the benefit of all learners, it is recommended that a variety of activities and scenarios be presented so that learners can practice the skills being taught in a variety of contexts. This is anticipated to support generalisation and help avoid learners associating what they are learning with a single activity type. Throughout content development for the Social Tutor care was taken to use a variety of images including simple cartoons, screenshots of the virtual humans, and photos of real humans within and across activities, as well as offering a range of instructional lesson types including stories, videos, songs and virtual human demonstrations, and a variety of practice lesson types including various sorting activities, interactive role-plays and other activities. Doing this is intended to assist students to learn the appropriate cues and skills in a generalised way that will allow them to apply their skills to novel, real-world situations.

The final key suggestions relate to the structure of the curriculum and take inspiration from Shute and Towle (2003) and their Learning Objects approach to content development as discussed in Subsection 2.8.1. In order to keep lesson sequences flexible and able to adapt to both proficient and remedial learners' needs, it is recommended to create a large number of lessons that are short, concise and each focus on a single concept rather than a smaller number of long lessons. Shute and Towle (2003) further recommend that each lesson be restricted to either basic knowledge, such as facts and formulas, procedural knowledge, or conceptual knowledge. Lessons structured in this way can then be easily grouped together to form cohesive objectives and topics that effectively teach their target skill. Furthermore, in order to best promote generalisation to novel contexts, it is recommended that many different lessons on the same topic area are created, each of which presenting the key ideas using different wording, different activities, and different digital media such as videos, photos and clip art. In this way students are supported to identify the key concepts and skills intended, for example learning how to identify what "bored" looks like in a wider sense, not just what Anna looks like when she is bored.

Following on from this, while multiple pages can be defined within a single lesson, it is typically suggested that only a single page with a single activity be used. This has the added benefit of creating a sense of progress, allowing learners to complete several short lessons in a single sitting. Most people can relate to feeling frustrated and disheartened when a learning task takes them a long time to complete, so this approach is intended to assist learners to maintain engagement and motivation. It has also been shown that inexperienced and experienced learners display different learning styles, for example inexperienced learners benefit more from demonstrations and worked examples and experienced learners benefit more from problem solving style tasks (Wittwer et al. 2010), therefore consideration should also be given to the complexity of the tasks, aiming to provide simpler tasks at the start of a topic and getting more complex as students progress.

4.5 Lesson Types

As noted previously, when performing social skills individuals with autism can exhibit a disconnect between what they know they should do and what they actually do. In this case they are experiencing a performance deficit rather than a skills deficit, and in both cases different learning activities are needed for the student to overcome their difficulty (Bellack 1983). For this reason, the Whiteboard software provides an extensive set of activity types and other elements that can be used when writing lesson activity files, a list with descriptions can be seen in Table 8. Some of these best facilitate skill-focussed learning, such as sorting, answering questions and watching videos, and others aim to facilitate performance-based learning, such as interactive role-plays and simple speech recognition activities. Additionally, while many of these lesson types are general in application, some were developed for the use of other educators and are not used in the content for the current study at all, namely the web browser, flash player and cloze activity. Some of the more visual activity types are relied on heavily in the Social Tutor content, such as the drag and drop activity, word grid, concept map and videos, while others are used sparingly, for example the paint panel.

The concept map activity type is particularly noteworthy following the discussion in Subsection 2.8.1. As discussed, many common assessment techniques truly only assess information recall, however evidence suggests that concept maps can assess deeper understanding of a concept (Iacobelli and Cassell 2007). In the Thinking Head Whiteboard, the concept map activity type allows educators to define multiple nodes and what to display on those nodes. They can then define which nodes should be connected and whether that connection is non-directional or directional, for example in a food chain the direction matters and 'cow → eats → grass' would be correct but 'grass → eats → cow' would not. Finally, educators can define the percentage similarity required between the learner's concept map and the defined solution to consider the learner's map complete. This is useful, as in complex maps multiple correct solutions are often possible.

By developing the Thinking Head Whiteboard using this modular approach, it is intended that developers with the right set of tools and knowledge of the Java programming language will find it simple to add additional activity types to the Whiteboard, further expanding its usefulness.

4.6 Other Whiteboard Functions

Within the Java source code there exists not only the activity classes previously discussed, but also an extensive set of 'manager' classes defining all of the 'behind the scenes' functionality. The manager classes cover all manner of processes from interpreting the XML input files and writing the XML output files, managing the virtual human instances, keeping track of the learner's progress, undertaking automated assessment and dynamic lesson sequencing, and otherwise controlling the overall flow of events. An overview of each manager class is provided below.

Table 8: Lesson activity types available in the Thinking Head Whiteboard software

Activity Type and XML tag(s)	Description
Drag and Drop **Tags:** drag_and_drop, word_grid	Multiple 'draggable objects' and 'drop boxes' can be defined, suiting sorting activities. Both a 'freestyle' version as shown in Figure 9 and a grid version are available.
Cloze **Tags:** cloze	A 'fill in the blank' style activity where a block of text is displayed with some parts replaced with empty boxes that the learner must complete. Both typed text and 'drag & drop' input modes are available.
Question List **Tags:** question_list, likert	One or more questions can be displayed simultaneously, with Likert scale and yes/no style responses both available.
Flexible Face **Tags:** flexible_face, multi_flexible_face	A cartoon face that can either be used to display particular expressions, or be interactive with learners trying to achieve a 'target' expression by moving control points on the eyes and lips. Multiple faces can also be displayed with yes/no style buttons beneath each face, ideal for sorting activities.
Highlighter Text **Tags:** highligher	Two modes available. In demonstration mode, the virtual human reads the block of text aloud and certain phrases are highlighted as they are spoken. In interactive mode, the learner must highlight key words or phrases within the block of text with their mouse.
Concept Map **Tags:** mind_map	Multiple nodes and the expected connections between the nodes are defined. Concept maps can be specified as 'directional' or 'non-directional' and a minimum expected similarity between the student map and the defined answer can be specified. The learner assembles the concept map by dragging nodes into position and clicking from one node to the next to connect them.
Paint **Tags:** paint_panel	A basic painting panel allowing learners to draw a picture using a variety of brushes and colours and then save it as a bitmap file.
Video **Tags:** video	Displays a media player within the Thinking Head Whiteboard software. The video controls can be hidden or displayed.
Browser **Tags:** url, flash	Displays a webpage or Flash file from within the Thinking Head Whiteboard software. Can display HTML or Flash from both the 'resources' folder and the Internet.
Social Explorer **Tags:** social_explorer	A simple game with one user-controlled cartoon character. Any number of other characters and items can be defined, and the characters' emotional states change depending on what interactions are chosen by the user. Mazes for the characters to move through can be automatically generated. Alternatively, obstacles can be manually defined and placed in the XML.
Other **Tags:** image, label, button, action	Static images, labels and buttons can be combined to create custom activity types. Combined with an action tag, buttons become interactive. Actions that can be defined include having a specific virtual human speak or act, moving to another page, exiting the lesson, automatically playing a video, or simple speech recognition.

Whiteboard Manager

This is the main entry point for the program. It builds the main Whiteboard window and contains a helper class to place interactive pages onto the frame and make them visible. The Whiteboard Manager also performs appropriate actions when the Whiteboard program is closed including checking if there is new homework to prompt the student about, ensuring student progress is fully saved, backing up output and initiating file upload when appropriate.

Input Parser

The Input Parser initiates the display of lesson options and builds the interactive display object to be placed on the Whiteboard frame when a lesson page is chosen. This class builds any general components itself, for example the 'tutor interaction panel' seen at the bottom of Figure 9, and delegates activity-specific construction to the appropriate display class. The Input Parser also ensures that each page's introductory action, for example the virtual human reading instructions aloud for the learner, is triggered correctly.

Action Parser

The Action Parser interprets XML 'action' elements and triggers the appropriate actions to occur in the appropriate sequence, such as movement between pages of a lesson, returning to the lesson selection screen on activity completion, initiating virtual human speech and behaviour, triggering sounds and videos to play, displaying pop ups, and managing simple speech recognition activities.

Output Manager

The Output Manager controls the creation of the student, summary, quiz results and log XML files as well as reading and writing to these files. Key actions include inserting new records into the log file each time the Whiteboard is opened or a new lesson is shown, recording user interaction in the log file, and inserting or updating records in the summary file when lessons are attempted or completed. For the current study only, it also saves the results of the content quiz into a separate 'quiz results' file for convenience.

Assessment Manager

The Assessment Manager determines what activity choices to present to the learner next. See 'Automated Assessment' below for a detailed explanation of this process.

Assessment Viewer

The Assessment Viewer builds the 'View Progress' display that allows the user to self-monitor their progress either by lesson or by objective. This can also be printed out, which is useful for educators and caregivers who wish to keep a record of their learners' progress and identify points of strength and weakness. A screenshot of the 'View Progress' window can be seen in Figure 10.

Head Manager

The Head Manager launches and coordinates up to four virtual humans simultaneously, each represented by its own Head class. The Head Manager controls the positioning of each Head X instance on screen, loads the user's saved preferences for Head X related customisation such as subtitle display, and calls the speech and command methods of the correct Head object, which in turn sends instructions to the corresponding Head X instance, controlling its behaviour.

XML Parser

The XML Parser manages the lower level reading and writing of XML input and output files. It is also used to check the validity of XML input files against their document definition files and other validation rules when the Whiteboard is first loaded.

File Upload Manager

File upload only occurs for user accounts that are actively engaged in the current study; no files are uploaded for 'unlocked' accounts. The File Upload Manager connects to the researcher's secure server and uploads any updated or new student files for data collection. It also checks for new error logs and uploads these to assist with debugging. While primarily uploading very small text files, file upload can sometimes be a lengthy process, and so this runs in the background even once the visual elements of the Whiteboard have been closed.

Session Timer

The Session Timer is only displayed for user accounts that are actively engaged in the current study. Students are expected to spend ten to fifteen minutes using the software every session, to facilitate this the timer provides a visual count down, gives the student both a spoken and pop up alert when ten minutes is up, and then when fifteen minutes is up the timer waits until the student exits their current lesson, and then lets them know time is up, says goodbye, and closes the software.

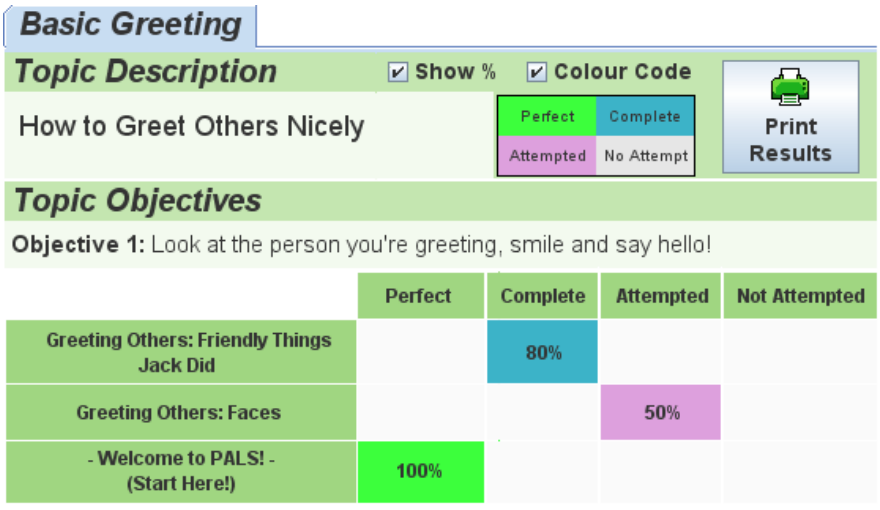


Figure 10: Condensed screenshot of the View Progress screen

Login Manager

The current implementation of the Login Manager is specific to the current research and requires some modification to make it appropriate for wider use. The Login Manager controls both account creation and set up, and the login process for existing accounts. Once a user's login credentials are validated the Login Manager triggers the process whereby the student is presented with the correct curriculum content. Students who are allocated passwords containing a number that is a multiple of three are directed towards the experimental social tutoring content; everyone else is directed toward the control group maze content. There are also two reserved passwords that can be used to create 'unlocked' accounts for both conditions; apple33 for the social content and orange88 for mazes.

Access Manager

The Access Manager is specific to the current research and should be disabled before distribution for general use. The Access Manager checks the number of days since the user created their account and what quizzes they have already completed, then based on this controls if the user is presented with a content quiz for data collection, with lesson activities, or if they are locked out of the software entirely. Once the user has completed their final content quiz their account automatically becomes 'unlocked'. This means they can access their lesson content without having to complete any more content quizzes or having their software use limited by the session timer.

4.6.2 Data Logging and Use

User interactions with the system are automatically and continually saved into an XML log file, while user progress through lesson tasks and topics is stored into an XML summary file, an extract of which is shown in Figure 5. The Output Manager provides helper functions that the display classes use to trigger these processes. The log files provide the means to investigate low-level user behaviour, down to the clicks, drags and other interactions in lesson activity screens, while the summary file is used as part of the automated assessment process to determine which tasks to present to the learner next. The contents of the summary file can also be viewed and printed by the student using the View Progress function of the Thinking Head Whiteboard, as shown in Figure 10. While caregivers and educators are likely to be interested in using the View Progress screen to observe their children's progress, it is expected that learners themselves are more likely to be motivated by an integrated progress and rewards system. For that reason a three stage rewards system has been implemented, as discussed in Subsection 4.6.4 below.

4.6.3 Automated Assessment

The automated assessment functionality of the Thinking Head Whiteboard is a core feature, however the current implementation is very basic. Following the discussion in Subsection 2.8.1, natural language processing and latent semantic analysis technqiues have been found to be effective in many existing tutoring systems and may be applied in some form in the future, however their need for written answers and the

incompatibility of that with the language and communication difficulties that accompany autism means that a simple heuristic-based algorithm has been implemented for the current iteration of the Social Tutor.

As discussed in Section 4.1, educators can create a curriculum definition XML file that specifies core lessons that must be completed by the learner, and can optionally specify 'extra lessons' that can be presented to the learner for additional practice if required. From this, the system determines exactly what activities to present to the learner and in what order, based on lesson prerequisites, previously completed activities, and learner proficiency. As discussed in Subsection 2.8.1, research recommends that for assessment to be most effective, it should be integrated into the learning activities rather than being a separate activity, and it should continually inform the learning activities offered to the student (Black 2015). The automated assessment system endeavours to do exactly that, with every task that the student does being assessed by the system, and thus influencing what activities they are presented with next. This also avoids any situations where the assessment applied is mismatched from the content that was actually taught, another issue that can occur when assessment and learning tasks are taken separately (Black 2015).

```
<curriculum desc="My Sample Curriculum" id="MySample">

    <topic id="MyTopic"
           desc="This is a sample topic"
           prerequisites="">

        <objective id="MyObjective"
                   desc="This is a sample objective"
                   required_correctness="0.8"
                   required_accuracy="0.8">

            <core>
                <lesson id="MyCoreLesson"
                        required_correctness="0.7"
                        required_accuracy="0.7"/>
            </core>

            <extra>
                <lesson id="MyExtraLesson"
                        required_correctness="1.0"
                        required_accuracy="1.0"/>

                <lesson id="MySecondExtraLesson"
                        required_correctness="1.0"
                        required_accuracy="1.0"/>
            </extra>

        </objective>
    </topic>
</curriculum>
```

Figure 11: A sample curriculum definition XML file

As previously discussed, curriculum definition files can contain multiple topics, with each topic containing one or more objectives. Each objective must contain at least one 'core' lesson, and all objectives must be finished successfully for the topic to be considered complete. As can be seen in Figure 11, which shows a sample curriculum definition file, educators can set prerequisites for topics by listing the IDs of other topics in the 'prerequisites' attribute of the topic tag. This means that no lessons from the given topic will be displayed to the learner until all earlier topics are completed first, allowing the educator to ensure the learner has gained sufficient prior knowledge before they are exposed to more complex content. Prerequisites can also be set for lessons in the same way, however the prerequisites attribute for this is within the 'lesson' tag of the lesson XML file itself, as can be seen in Figure 8. If the prerequisites attribute is missing or contains an empty string it is assumed that the given topic or lesson has no prerequisites and can be included in the options immediately available to the learner.

Figure 11 also highlights how a minimum required 'correctness' and 'accuracy' can be set for each lesson and each objective. The value must be between 1.0 and 0.0, with 1.0 representing 100% accuracy or correctness and 0.0 representing 0%. Correctness here refers to the final state of the activity only, and whether the student has met all of the activity's requirements. Accuracy gives an indication of how many mistakes the learner has made while completing the task. To calculate accuracy, a tally is kept of both the total number of moves and the number of incorrect moves that the student makes during their interaction with the system. Accuracy is then calculated simply as shown in (1).

```
(total_moves - incorrect_moves)/total_moves        (1)
```

Consider a 'drag and drop' activity such as that illustrated in Figure 8 and Figure 9. 100% correctness would be awarded if the learner places all 'draggable' objects onto their corresponding 'drop boxes', with no 'draggable' objects in the wrong boxes and any 'draggable' objects without a specified 'drop box' being left in the blank space. 100% accuracy would be awarded only if the learner completed this task without misplacing any 'draggable' objects. If the learner placed a 'draggable' object into the wrong box then recognised their mistake and moved that object into the correct box, they would still be awarded 100% correctness but their accuracy would be lower.

Setting minimum requirements for lessons and objectives allows educators to specify the level of mastery they require from their learners in order to consider the section complete to a sufficient level and unlock new lessons. An example of this is to have correctness for an activity set to '1.0' and accuracy set to '0.8', thus the learner must have the task 100% correct when they submit it, but can have an accuracy of only 80%, meaning that they can make a few mistakes while they are doing the activity provided they fix them before submitting it. If the learner does not reach the minimum required correctness and accuracy they will be able to repeat the task and try again.

The automated assessment function of the Thinking Head Whiteboard endeavours to present all 'core' lessons to the learner as quickly as possible while adhering to the set curriculum requirements. To achieve this, the

Assessment Manager generates a list of 'accessible lessons' consisting of lessons with no prerequisites or with all prerequisite tasks already complete, and selects up to three activities for the learner to choose from. The Assessment Manager begins by extracting a list of all core lesson IDs from the curriculum file. Core lessons that are already accessible are added directly to the list, while lessons with incomplete prerequisites are not. For these, the Assessment Manager performs a depth-first search through the lesson's prerequisite sequence until it finds a lesson that is already unlocked. For example, if my core lesson "TargetLesson" had a prerequisite of "MiddleLesson" and "MiddleLesson" had a prerequisite of "FirstLesson" but "FirstLesson" had no prerequisites, "FirstLesson" would be the ID added to the 'accessible lessons' list as it is the first available lesson in the sequence required to unlock the required 'core' lesson. For topics this same prerequisite checking process occurs, but instead of limiting the learner to three choices, all unlocked topics are displayed for the user at all times.

A maximum of three lesson options are provided to the user at any one time in an attempt to avoid overwhelming the learner with too many choices and to guide them towards the most direct path through the lesson content, but at the same time giving them a sense of control over their learning. The 'accessible lessons' list is typically much longer than this. When lessons are added to the 'accessible lessons' list, they are assigned a status – core new, extra new, core previously attempted or extra previously attempted. Priority is given first to new core lessons, followed by previously attempted core lessons, then new extra lessons and finally previously attempted extra lessons. Lessons that have been completed to their minimum required correctness and accuracy are only added to this list if all necessary lessons are complete, but the average correctness or accuracy of all lessons within the objective falls below the minimum requirements to pass. In this case, lessons will require repetition until the objective's requirements are met. This case depends highly on how the educator structures their curriculum within the XML file. In most cases, once a lesson is complete it will no longer be eligible for inclusion in the list of activity choices presented to the learner, however all previously attempted activities can be accessed via the 'Repeat Previous Lesson' option on the main menu and repeated as often as desired.

In addition to this process, if a student struggles with a lesson there exists a mechanism to take a step back in the lesson sequence. After a student has attempted a lesson and obtained a correctness or accuracy score below a set threshold, the system resets that lesson and all of its prerequisites, making them all available for the student to repeat. The goal of this process is to combat instances where the learner has completed a lesson without properly understanding it or needs more repetition to consolidate their knowledge, and ensures they do not simply become stuck repeating the same lesson with no way to obtain the extra and varied practice they need.

The simple automated assessment and dynamic lesson sequencing approach described here aims to allow learners with more advanced knowledge of the topic being presented to fast-track through and complete only the minimum required lesson sequence, while enabling learners who need more time to access the extra lessons to supplement their learning as well as repeat previous activities until their understanding is strong

enough to complete the core lessons to the required level. While this implementation is very simple and strongly driven by the XML curriculum definition file written by educators, it is nonetheless expected to provide a more relevant and engaging educational experience for the learner than can be provided using a static lesson sequence.

4.6.4 Rewards and Reinforcement

A number of mechanisms have been put in place with the aim of encouraging and motivating students to continue to use the software and to support generalisation of skills to novel contexts. The simplest feature is that the virtual tutors are themselves encouraging, providing immediate feedback in the form of simple praise or prompts as learners interact with their activities. Hattie and Timperley (2007) found that providing praise, reward or punishment alone only has a small influence; however, feedback with helpful suggestions about the task or skill can be very powerful and beneficial for learners. In the Social Tutor software, if students make a mistake while working through a task, the teacher Kate provides them with constructive feedback. While the exact nature of the comment and when it is delivered varies across activity types, it can include a hint or feedback on what the result of the student's selection would be, e.g. "Oops! That would make Jack feel sad. Try again!". In several activity types, such as the concept map, if the student continues to make mistakes the hints become more detailed, for example ranging from "a connection is missing" all the way to the explicit "node A should be connected to node B". Finally, the virtual tutors typically provide a short 'recap' at the completion of each activity so that the key steps of the skill being targeted are continually highlighted and reinforced.

As discussed in Subsection 4.6.3 a three tiered rewards system has been implemented to provide extrinsic reinforcement. In the first tier, students automatically gain a gold star for each lesson they complete. In the second tier, students can choose to trade five gold stars in for a virtual 'sticker' and work towards completing their sticker collection. This is anticipated to help maintain motivation and provide students with a sense of progress, but also not detract too much time from their core learning activities. The third stage involves unlocking reward activities, where content reinforcing games are made available to the student at 50% and 100% completion of each topic. Students can play one of their available games per session, and can only play it after they have attempted at least one of their normal lesson activities first. This is to ensure they continue to progress through their regular activities at a suitable rate. These activities include turn taking games like "Go Fish" and "Guess My Number" as well as "Social Explorer" games where the student has to work out what each character's favourite toy is and take it to them in order to complete a maze, or interact with the characters using particular pro-social behaviours to complete the activity.

In addition to the rewards system, a homework system has also been implemented to encourage students to practice their skills outside of the software. Homework activities become unlocked after students complete prerequisite activities within the software that introduce the steps of the target skill. Once unlocked, the virtual teacher Kate lets students know that there is homework available. As mentioned in Section 4.1, the format of homework lessons is based on those in the Skillstreaming curriculum (McGinnis and Goldstein

2012). Modelling the method used by McGinnis and Goldstein (2012) and taking a similar approach to the Say-Do Correspondence Training advocated by Rosenberg et al. (2015), homework activities are a two-step process, a screenshot of which can be seen in Figure 12.

Rosenberg et al. (2015) asked students to identify who they were going to talk to at recess and provided reinforcers for students who followed through. It was shown that this resulted in an increased number of social exchanges for all participants. It was hoped that taking a similar approach in the software would likewise facilitate generalisation of skills for learners in the current study.

When a homework activity is first accessed, the learner is given a refresher about the steps involved in the skill, and is then asked to write down who they will practice with and when they will practice. The next time they log into the software and access the same homework activity, they will again be given a refresher on the steps of the skill, and this time will be asked to write down what happened, rate how they went, and explain their rating. It is acknowledged that the homework activity requires extensive typing which is not ideal for the target learner group. Preferably this would be completed using speech recognition technology instead, however as discussed in Subsection 2.6.4 there are some difficulties with speech recognition for this user group, as well as simply not being able to guarantee that all participants would have a suitable microphone. In the meantime, parents were instructed that they could do the typing for their child if they wished.

Homework is optional but encouraged, so if an incomplete, unlocked homework activity exists, Kate will ask the student if they completed it each time they log into the software and remind them that it exists each time they exit. It is hoped that this will encourage generalisation of skills learned in the software to novel contexts such as home and school.

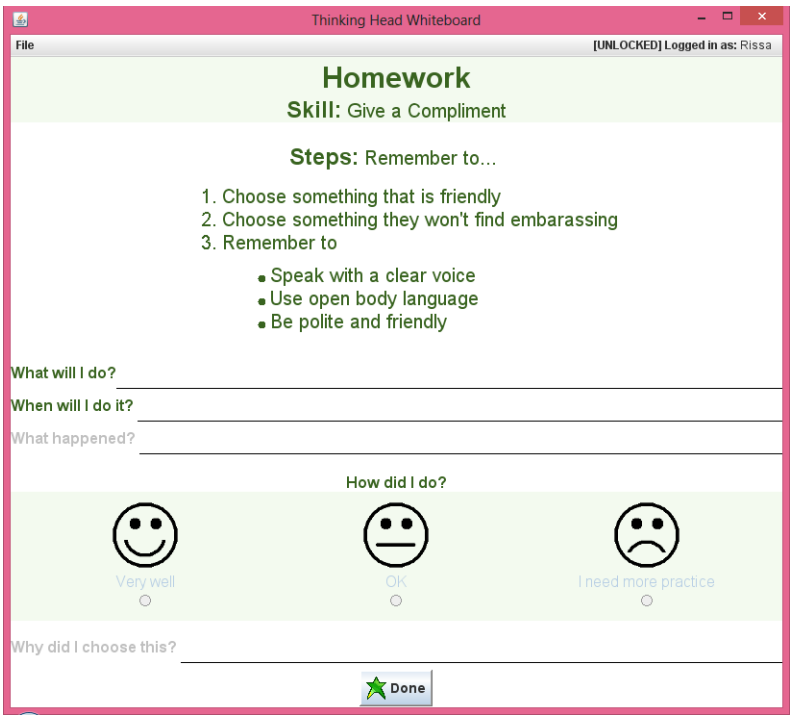


Figure 12: Screenshot of a homework activity

CHAPTER 5. EVALUATION METHOD

A review of the efficacy of existing social skills interventions for children with autism identified many recommendations for improving the experimental methods used to assess the interventions (Rao et al. 2008). The more recent follow-up work by Neely et al. (2016) suggests that, while some progress has been made, there is still room for improvement. Wherever possible, this study aimed to meet these recommendations. This chapter outlines the methodology used to evaluate the Social Tutor software developed here including participant recruitment procedures, selection and design of measurement tools, and an explanation of the data collection process as a whole.

5.1 Overview and Justification

One of the major issues identified by Rao et al. (2008) was a lack of studies using a control group. To address this two sets of content were developed for this software, one which explicitly teaches social skills and one that does not (for more detail on the differences see 5.3). In the review it was also found that only three of the ten studies investigated involved more than ten participants (Rao et al. 2008). Having a sample size with sufficient power for meaningful data analysis is important, not only so that useful conclusions can be drawn that help inform intervention use and future research, but also to ensure that burden is not being placed on participants and their caregivers unnecessarily. Another limitation of existing studies was the lack of blinded observer ratings. In response to these identified issues and the associated recommendations, this study aimed to recruit sixteen participants in each group for a total of thirty two participants overall, and allocated participants to the control and experimental groups using a matched pairs approach as described in Subsection 5.2.2. Caregivers were not informed which group their child had been assigned to until after all data completion was complete, and were simply informed that there were two versions of the software being evaluated and their child would be allocated to one of these groups. Combined with the participant inclusion criteria, taking this allocation approach helped ensure that both the control and experimental groups were as similar in profile as possible so that results from the two groups could be compared with higher confidence, providing sufficient power for meaningful data analysis.

Limitations were also identified regarding generalisation and maintenance, two widely acknowledged issues for individuals with autism in relation to interventions (Rao et al. 2008, Neely et al. 2016). Some studies failed to explicitly promote or measure generalisation at all, while several others who did attempt it unfortunately had poor results (Rao et al. 2008). As discussed in Subsection 2.6.2, the more recent work by Neely et al. (2016) found that researchers are often still failing to explicitly program for or measure generalisation and maintenance, and those that do are employing a "train and hope" approach. Rao et al. (2008) found that the single study which experienced some success with generalisation incorporated a component allowing participants to interact with their neurotypical peers outside of the treatment setting, which they note is also a successful strategy for children with social anxiety disorder. Learning from this, and as described in Chapter 4, interaction with neurotypical peers is encouraged via the homework exercises

in the software, as well as exposing the learner to a variety of cartoon and photo depictions of facial expressions and scenarios, and learners are required to apply the same skills in multiple different activity types within the software. In addition to the in-software quiz to measure retention of the presented educational content, this study also explicitly measures generalisation of skills and changes in real-world behaviours by using the Vineland Adaptive Behaviour Scales. Rao et al. (2008) noted that most interventions at the time failed to measure for maintenance effects post-intervention, however the more recent work by Neely et al. (2016) indicates that the situation has improved, with most studies reviewed including some data collection following the intervention period, however typically this only happens once and does not extend past three months after the end of the intervention. To address this identified lack of follow up assessment, data in this study is collected not only immediately before and after the three weeks of software use, but also at two and four months post-intervention, with the goal of analysing the stability of any effects over time, as per research objective four described in Chapter 3: Research Aims.

5.2 Methodology

When designing the evaluation methodology, particular care was given to minimising the burden on families participating in the study, while also aiming to collect sufficient data to draw meaningful conclusions from its analysis. It is appreciated that families of children with additional needs such as those with autism are already under enormous time and resource pressures, and the software being developed was intended to fit in to their daily routines. Wherever possible, data collection has been automated or provided in a way that enables families to have flexibility while also maintaining validity and consistency across the data collection period.

In order to conduct this software evaluation, ethics approval was first obtained from the Flinders University Social and Behavioural Research Ethics Committee (Project 5703). Following this, ethics approval was obtained from Autism SA (Project PP201611) allowing advertising to be conducted through their channels. The approval documentation for both of these committees can be seen in Appendix L. Ethics approval was also obtained from the Department for Education and Child Development (project CS/16/00068-1.8) and Catholic Education South Australia (reference 201618) to enable the evaluation to be run in the school environment rather than being restricted to only the home environment, however no participating families requested this and all data included here is from participants who used the software in their own homes.

5.2.1 Recruitment and Inclusion Criteria

Participants were recruited via advertising with the local autism support provider, Autism SA. Following ethics approval by their professional practice committee, Autism SA placed an advertisement on their website and included advertisements in several of their electronic newsletters that are sent to families and educators who have signed up to receive it, a copy of which can be seen in Appendix M. Interested parties then contacted the researcher via the details provided. Following this initial contact, the researcher confirmed if the children were eligible for participation in the study and provided the family with the relevant

information sheets and consent forms via email, a copy of which can be seen in Appendix N, along with instructions for the next steps if they wished to take part. The advertisement used with Autism SA was likewise added to the Flinders University website, and again interested parties contacted the researcher and the same process was followed.

In order to be eligible for the current study, participants had to be aged six to twelve years old, have an existing diagnosis of Asperger Syndrome or high functioning autism, and currently be attending a mainstream school. These inclusion criteria provided several functions, the first being to ensure that the participant group was sufficiently homogenous for the data to be informative, and the second to ensure that the content presented was both age and context appropriate for all participants. Finally, including only individuals high functioning enough to attend mainstream school allowed a minimum level of communication skill and general functioning to be assumed when selecting the target topics and designing lesson activities.

Finally, families had to have access to a suitable computer for the duration of the study, which could either be at home or, with permission, a school computer. The computer had to have the Windows 7, 8 or 10 operating system installed and an internet connection. All families opted to use their own home computers.

5.2.2 Group Allocation

A matched pairs approach was taken when allocating participants into the experimental or control group. Once the recruitment process has progressed to the point where the family and researcher had agreed on a date for the first visit, the participant was allocated into a group. Participants were matched on age according to three 'buckets' - six to eight year olds, nine and ten year olds, and eleven and twelve year olds. The first participant was placed into the experimental group. The next participant, if their age fell into the same 'bucket' as the first participant, was matched with them and placed in the control group. If their age did not fall into the same 'bucket', they were instead allocated into the experimental group. This process continued for each participant. Participant gender was not explicitly controlled for as it was expected to be difficult to conduct in practice and have negligible effect on the outcomes given the relatively small total sample size being recruited and the typical gender ratio for autism diagnoses of approximately 1 female for every 4 males, although more recent research suggests that this ratio may be closer to approximately 1:3 (Loomes et al. 2017) and thus in future research consideration should be given to controlling for gender explicitly. A summary of participant demographics can be found in Section 6.1. Finally, it should be noted that siblings were kept in the same group as each other to avoid issues where the assessor (caregiver) guesses which group the children had been assigned to, and to avoid issues relating to fairness and jealousy between siblings. Three pairs of siblings were involved in the study, with two sets allocated to the experimental group and one to the control group.

5.2.3 Selection and Design of Tools

Given the intention of utilising a research methodology that minimises the burden on families, the selection of measurement tools favoured those that could be completed electronically and independently by the participant and caregiver as appropriate. For this reason, more time-consuming or invasive measurement tools that required play-based sessions or observations of natural behaviours were deemed unsuitable for the current study. However, in order to ensure an accurate picture of the impact of the Social Tutor software can be gleaned, a multifaceted data collection approach has been taken.

As described in Chapter 3: Research Aims, the first core objective was to design and implement software to teach social skills to children with autism. The second and third objectives focussed on determining if changes in social knowledge and behaviour occurred following interaction with this system, the fourth core objective was to determine if any changes that did occur were maintained after software use ended, and the fifth objective was to determine participant and caregiver perceptions of the software. As discussed in Subsection 2.6.2, generalisation to novel contexts is a known difficulty for individuals with autism, so it is important to assess both theoretical level understanding and real-world skill performance in order to gain a full picture of the efficacy of the Social Tutor software. Given this, 'near transfer' of skills was measured using the in-software content quiz, which presented similar activities but with content modified from that used during teaching, and 'far transfer' of skills to real-world situations was measured using the Vineland-II behavioural assessment. Additionally, to provide better insight into the way that participants interacted with the software, how much content they were able to cover in their active software use period, and the profile of this interaction, ongoing data collection was automatically conducted within the software itself, as described in Subsection 4.6.2. A known issue with children on the autism spectrum is that they can learn how to 'do the intervention' without applying what they are learning to situations outside of the intervention context. Comparison of results from the Vineland-II and content quiz provides insight into whether knowledge gained from using the software is likely to have transferred to behaviour changes outside of the intervention setting, and investigation of log data provides an indication of which components in the software had the most impact or require change.

To address the fifth core research aim, participants completed a questionnaire prior to software use to provide an indication of their expectations and computing expertise, and both participants and caregivers were asked to complete post-test questionnaires reflecting on their experience with the software. This is intended to direct future development of the software and provide insight into what aspects were best received and where difficulties arose.

Vineland-II

The Vineland-II (Sparrow et al. 2005b, Sparrow et al. 2005a) was selected as the behavioural assessment most appropriate for the current study, as it provides a fine grained level of detail sufficient to allow for detection of subtle behaviour changes in the target population. Thus, it was used to assess 'far transfer'

generalisation of skills to real-world scenarios. A detailed discussion of the justification for selection of the Vineland-II can be found in Section 2.8 Assessment Tools and Techniques.

As indicated in Table 9, caregivers were asked to complete the Vineland-II at pre-test and all three post-test data collection points. Only the areas relevant to the current study were included in the electronic version of the Vineland-II used for this study, specifically these were the Receptive and Expressive subdomains within the broader Communication domain, all of the Socialization domain, and all of the Maladaptive Behaviours domain. The Written subdomain of the Communication domain and the entire Motor Skills domain have been omitted as these areas were not addressed in any way by the content of the Social Tutor. The Vineland-II was presented as a password protected Google Form, with each statement in the assessment displayed with a set of multiple choice radio button answers, and typically half to all of the questions in a single subdomain displayed per page depending on length. At the end of each page a text box was provided so caregivers could clarify or provide detail on their answer for a particular question if they wished. The Vineland-II was displayed on a provided iPad at pre-test, and then accessed on the caregiver's own computer via a supplied URL at the three post-test data points. The behavioural assessment typically took 20 to 30 minutes to complete. Following data collection, raw scores for the included domains were converted to v-scale scores according to standard Vineland procedure.

Content Quiz

The content quiz was designed to directly measure basic and procedural knowledge (Shute and Towle 2003), and consists of four short activities drawn from each topic within the experimental group curriculum, the three topics being greeting, listening and turn taking, and starting and ending conversations, totalling twelve activities all together. The twelve activities were presented within the software itself, and displayed in random order each time the content quiz was run. The content quiz activities were assessed automatically by the software, which then provided a percentage correctness and percentage accuracy score for each question, along with timestamps that could be used to determine the time taken to complete each of the twelve activities. Content quiz scores across the data collection time points were compared to determine what information participants had learned and retained over the period of the study. The content quiz questions and expected responses can be seen in Appendix A.

Pre-test Questionnaire

Participants were asked to complete a very short questionnaire prior to software use. This was presented as a Google Form, and caregivers were instructed that they could assist their child to enter their answers if desired. The pre-test questionnaire aimed to provide a sense of the learner's experience and confidence with software and learning new things in general, and to gain insight into the learner's expectations of the software. This data provides insights into commonalities between users relating to both educational outcomes and perceptions of the software after use. The pre-test questionnaire can be seen in Appendix B.

Post-test Questionnaire

Following the three weeks of software use, participants and their caregivers were asked to each complete a questionnaire assessing their experiences with the software and their recommendations for future development. Some elements of the questionnaire were presented using a rating scale so that responses could be numerically evaluated, while others were open-ended to provide participants and caregivers with an opportunity to include any feedback they deemed beneficial or noteworthy. Families were encouraged to use the post-test questionnaire as an opportunity to recommend directions they would like to see software such as this taken in the future. The post-test questionnaire can be seen in Appendix C.

Software Log Data

In addition to the data collection tools already discussed, the Social Tutor software itself continually logged user interactions with the system and saved user progress across learning tasks. From these logs and summary files, information about the number of lessons that participants completed, the average amount of time spent on these lessons, and which topic areas and lessons they attempted could all be extracted. This has the potential to provide a clear picture of user interaction with the system, and in conjunction with the other measures previously discussed, assists in identifying areas of strength and weakness within the software.

5.2.4 Data Collection Schedule

As discussed previously, generalisation and maintenance of skills from the intervention to the real world must be explicitly addressed and measured in any intervention aimed at individuals with autism. To allow maintenance to be measured across time, data was collected at four points - immediately prior to software use, immediately following software use, and both at two and four months following software use. Data collection took place over 8 months, with the first participants recruited and beginning the evaluation in mid-August 2016, the last participants beginning the evaluation mid-November 2016, and the final data being collected in early April 2017. A summary of the data collection schedule can be seen in Table 9.

The evaluation methodology was designed so that only a single visit by the researcher to the participants' home or school was compulsory, with all data collection performed electronically. During the researcher's visit, a discussion of the study process took place and informed consent was obtained, then the researcher installed and tested the software on the family computer. While the researcher was installing the software, the caregiver was asked to complete the Vineland-II behavioural scale electronically via a Google Form displayed on the researcher's iPad. This was done with the researcher present so that any queries the caregiver had while answering questions could be quickly answered. Once the Vineland-II was completed by the caregiver, the participant was asked to complete a pre-test questionnaire electronically, again in the format of a Google Form and displayed on the iPad. Caregivers were instructed that they could assist their child to input their answers. Once the software was installed, the participant was shown how to set up their account and log into the software, and was asked to complete the pre-test content quiz. Again this was done with the researcher present to ensure any questions about how to input answers or use the software could be

answered or demonstrated immediately if required. This visit from the researcher typically took under an hour and a half to complete, although where technical difficulties were encountered could occasionally take longer. Following this visit, no more visits from the researcher were compulsory, however families were informed that if they encountered any difficulties with the software or wanted support with data collection, the researcher would be happy to visit and assist them.

Table 9: Data collection schedule

	Pre-Test	**Software Use**	**Immediate Post-Test**	**2 Month Follow Up Post-Test**	**4 Month Follow Up Post-Test**
Timeline	**Day 1**	**Day 2 - Day 22** Ongoing for 3 weeks	**Day 23** The day after software use ends	**Day 83** 2 months after post-test	**Day 143** 2 months after second post-test
Researcher Visit	**Yes**	**No**	**On request**	**On request**	**On request**
Caregiver actions	Complete Vineland-II online *	Support participant to use software if required	Complete Vineland-II & questionnaire online	Complete Vineland-II online	Complete Vineland-II online
Participant actions	Complete content quiz via software & questionnaire online *	Software use 10-15 min a day, 3-5 times a week	Complete content quiz via software & questionnaire online	Complete content quiz via software †	Complete content quiz via software †
Researcher actions	Install software * Support caregiver to complete Vineland-II and participant to complete quiz and questionnaire		Contact caregiver to remind them to complete assessments	Contact caregiver to remind them to complete assessments	Contact caregiver to remind them to complete assessments. When all data received, provide unlock instructions and reimbursement

† As described in Section 5.4 Challenges Encountered During Evaluation, technical difficulties necessitated some participants completing the content quiz at two and four month follow up via an electronic Word document.

Families were instructed that, starting the day after the researcher visits, participants should use the software for one 10-15 minute session per day, 3-5 days per week, for three weeks. As previously discussed, a timer in the software allows users to self-manage their session times, reducing the burden on caregivers.

Three weeks and one day from set up of a new account, the software automatically presented the user with their first post-test content quiz. At this point, users could no longer access their lesson activities until the final four month post-test was complete, at which point they could login as usual and resume their lessons if they wished. In addition to the content quiz that participants completed using the software, caregivers were again asked to complete the Vineland-II, this time via the browser on their home computer, and both caregivers and participants were asked to complete the post-test questionnaire addressing their experiences and recommendations. Following this, no action was required from participants or their caregivers for the next two months.

Two months after the previous post-test, participants were asked to log into the software and complete the content quiz again, and caregivers were asked to complete the Vineland-II. Following this there was another two month break, after which participants were once more asked to log into the software to complete a fourth

and final content quiz, and caregivers were asked to complete their final Vineland-II. On completion of the content quiz, the software automatically unlocked the user's account and their lesson activities became accessible once again. Once all data was complete and received by the researcher, the family was also reimbursed $30 for their participation and were provided with the necessary instructions for them to access the version of the software that they were not initially assigned to.

At each 'post-test' data collection point, the family received an email from the researcher a few days before their data collection was due which provided links and instructions for the required tasks, as well a small window of time in which it was requested they complete them. A small number of families required additional reminders before completing data collection tasks, however most did so within the requested time frame. Additionally, it should be noted that completeness of data collection tasks varied across participants, a summary of which can be seen in Appendix E.

5.3 Experimental Software

As discussed in Chapter 4, the backend software used for both the experimental and control groups in the current study was identical and consisted of three instances of the virtual character software Head X and one instance of the Thinking Head Whiteboard software run simultaneously. Only the lesson content used by the groups was different. This was to ensure that the experience of both groups was as identical as possible. The workflow for all students and the differences between the content for both groups is described below.

5.3.1 Typical Student Workflow

For all study participants, the overall workflow through the software was identical. An overview of a typical session during the three weeks of software use can be seen in Figure 13. A typical session started with the user logging into the software with their provided credentials, which consisted of their name and a password indicating to the software which version of content to present. They were then welcomed by the virtual teacher while the software read in the details of their student file to determine what lessons have been completed previously, allowing the assessment manager to determine which topics to present to the learner. Once loaded, the virtual teacher prompted the learner to choose one of these topics, and the student clicked on their selection. If the 'read lesson buttons' option was active, clicking the topic's button once resulted in the virtual teacher reading the title of the topic, and the student then clicked a second time to select it. If 'read lesson buttons' was not active, the student only clicked once to select a topic. A maximum of three lessons from the chosen topic were then presented to the learner to choose from. Again, the virtual teacher prompted the learner to pick one and they clicked their choice. If 'read lesson buttons' was active, the teacher read the title of the lesson and a second click selected the lesson, otherwise a single click loaded the lesson.

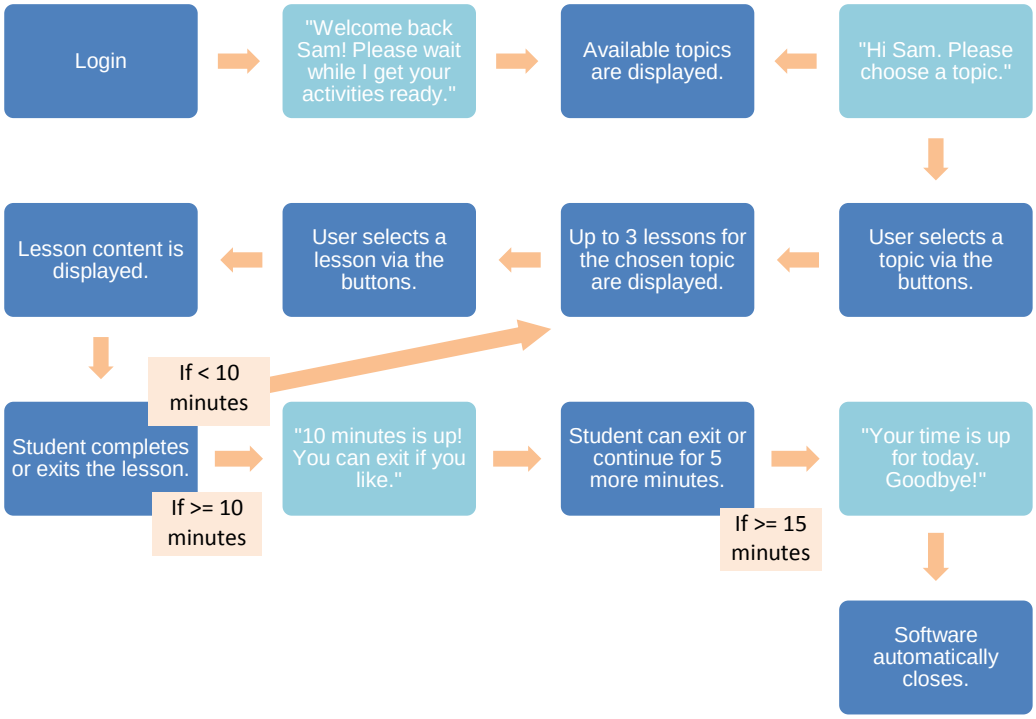


Figure 13: Student workflow

A typical lesson involved a short introduction and explanation by the virtual tutor Kate, and then the user interacted with the Thinking Head Whiteboard to complete the presented activity. The activity content varies depending on the lesson selected and the group the participant is allocated to. Typically, correct actions within the activity elicit a simple praise utterance from the virtual teacher, and incorrect actions typically elicit a simple prompt utterance. More detail on lesson content is provided in Subsections 5.3.3 and 5.3.4.

Once the student completed the activity they returned to the lesson selection screen. Alternatively, they could use the menu to choose to exit without completing the activity. If they chose "select a new lesson" they were returned to the lesson selection screen or if they chose "select a new topic" they were returned to the topic selection screen. In the case of the lesson selection screen, they were presented with a maximum of three lessons again, however due to the assessment manager process these were not always the same three options presented previously. The student continued to work on lessons for ten minutes, at which point the virtual teacher let them know both verbally and via a pop up that ten minutes is up and they could choose to exit or continue for another five minutes. At this point the user could simply close the software. If they chose to continue working on lessons, once another five minutes had passed the virtual teacher would wait for them to complete or exit the activity they were working on and return to the lesson selection screen, and then let them know that time is up, say goodbye, and automatically close the software. The exit process involved closing all three Head X instances fully, then making the Thinking Head Whiteboard transparent while the user's progress was saved and uploaded to the secure server in the background. Once this was done the Thinking Head Whiteboard was also closed fully.

5.3.2 Data Collection Workflow

The process above describes a typical session during the three weeks of software use; however, the process was different at the data collection time points. For the pre-evaluation time point, the student was required to create their account rather than login. As credentials were supplied, the student simply entered their name and password then selected "New Account" rather than "Login" on the welcome screen. For the post-evaluation time points the student simply logged in as usual. Following this, the two instances of Head X that display virtual children were closed and the Head X instance containing the virtual teacher expanded to fill more of the screen. The virtual teacher then greeted the user and automatically displayed the content quiz. The virtual teacher thanked the student for agreeing to help evaluate the social tutoring software, explained or recapped some key features of the interface that are useful during the quiz as appropriate, and then presented the quiz content in random order. The content quiz was the same for both groups and at all data collection time points, with identical questions always presented in random order, prefaced by a short welcome and some instructions by the virtual teacher. Students worked through their questions sequentially, and once all questions were complete the virtual tutor praised the student and instructed them to close the software.

5.3.3 Experimental 'Social Content' Group

Lesson content within the social tutoring curriculum varies greatly between activities, but there are commonalities relating to both lesson presentation order and individual lesson activity format. Regarding lesson presentation order, each set of activities grouped together on a target skill follows a standard pattern where initial lessons introduce, explain and demonstrate the target skills through Social Story style activities, virtual role-plays and skill modelling by the virtual characters, and videos where available. Following this, interactive lessons that encourage learners to identify, explore and apply the skill steps and common features of the target skill in a variety of activities are provided. These typically include highly visual sorting style activities such as drag and drop tasks, Venn Diagrams, and word grids. Depending on the target skill, other performance focussed tasks are also included such as basic speech recognition activities and interactive role-plays where the student controls what the virtual peers do in response to certain social stimuli, and reflective tasks such as virtual drawing are likewise included where appropriate.

Within these individual lessons, each activity begins with the virtual teacher Kate introducing the task, sometimes with assistance from the virtual peers Jack and Anna, and then the student completes it at their own pace. For most activity types, each action the learner takes is met with either praise for a correct action or a prompt or hint for an incorrect action. Completion of the lesson automatically returns the student to the lesson selection screen so they can choose their next activity.

In addition to the standard lessons, homework lessons cover content that directly relates to the topics and lessons the student has previously accessed and encourage the student to practice their skills with their families or peers at school. The reward activities for the social skills curriculum likewise aim to reinforce

lesson content, including the turn-taking games 'Go Fish' and 'Guess My Number' and games that require the user to interact with cartoon characters in pro-social ways.

5.3.4 Control 'Maze Games' Group

In contrast to the social skills curriculum, the lesson content for the control group consists of a series of mazes. The mazes are still grouped into 'topics' as can be seen in Figure 14, with the mazes within each topic fitting in with the theme. An example maze from the 'Spooky and Kooky' topic can be seen in Figure 15. Mazes increase in difficulty and length as the user progresses, from easy single page mazes such as that shown in Figure 15, through to medium, hard and extra hard mazes where the mazes have more obstacles, multiple pages or require more than one item per obstacle.

The format for every lesson is identical, with one of the virtual characters randomly selected to introduce the maze, and the student completing the maze at their own pace and being returned to the lesson selection screen once the maze is successfully navigated.

Figure 14: Control group topic options

In maze activities, if the user is stuck they can type in keywords, full sentences or partial sentences to request a hint from the virtual characters, e.g. "how do I get past the bat ghost?" The images and correct spellings of each obstacle and item are displayed to assist users to do this successfully.

To maintain consistency between the experimental and control groups, homework and reward lessons are also included in the mazes curriculum. Homework activities contain no social content and instead ask the student to do tasks like draw a picture, do some reading, or design their own maze. Reward activities likewise contain no social content, instead consisting of retro style games such as Tetris, Snake and Memory.

5.4 Challenges Encountered During Evaluation

There were a number of technical challenges that arose during the software evaluation period. The first was that installing software of this nature, with multiple independent but critical components, meant that the installation process ended up being slightly different for every individual computer encountered. In only a small number of cases did installation proceed exactly as anticipated. In most cases some additional libraries had to be installed or small settings tweaked on firewall or anti-virus software to allow the Social Tutor software to be installed and to function correctly with all components able to communicate with one another. At one point when over three quarters of the participants had already had the software installed and

commenced using it, a virus signature update was rolled out that caused a core component of the virtual human software to be flagged as suspicious and automatically deleted when trying to install the Social Tutor on new computers. A workaround was found for the first participant affected, and following that the software itself was modified so that the component of the virtual human software in question was no longer needed but the Social Tutor itself still performed exactly as it had before the change, ensuring a consistent user experience for all participants.

After all participants had completed the three weeks of software use and post-test, one final technical hurdle was faced when a Java update was released that caused an older library used by the Social Tutor to display media files to fail. This in turn caused the Social Tutor to report an error and fail on loading, which meant that for some participants they were not able to complete the two and four month follow up post-test quizzes via the software. In this case many families opted to complete this component of the data collection manually using a supplied Microsoft Word file that contained screenshots of the content quiz questions and instructions on how to complete each of the activities.

The final challenge encountered was simply that of 'disruptive life events' where over the course of the study several participants moved house or school, or left mainstream schooling and began homeschooling, while one moved overseas. These are significant life events for anyone, and for a child on the autism spectrum can be particularly confronting and disruptive. Additionally, for many participants one or more of their follow up post-tests fell during the Christmas holidays or start of the new school year, again quite unsettled and disruptive times of year particularly for children on the autism spectrum given their need for consistency and predictability. Several parents commented on this and noted that they were curious to see if their children's results would be reflective of these disruptive times. A discussion of the data in relation to these challenges and issues is provided where appropriate in the following chapter.

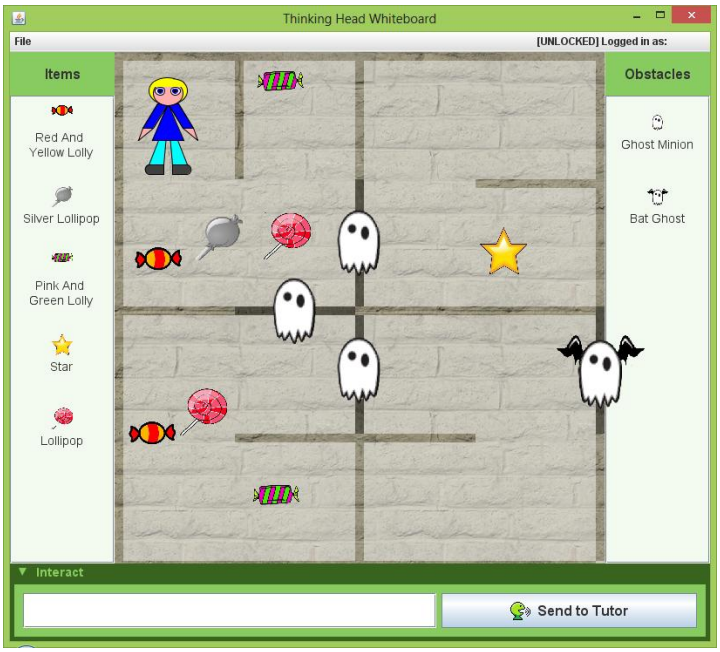


Figure 15: Example of a control group maze activity

CHAPTER 6. RESULTS AND ANALYSIS

Several measurement tools were used throughout the software evaluation in order to address the stated research objectives, including log data from the software itself, a content quiz directly assessing student knowledge, the Vineland-II behavioural scales for assessing participants' real-world social behaviours, and both pre-test and post-test questionnaires for assessing participant and caregiver expectations and perceptions of the software. Unless otherwise stated, all statistical analysis was carried out using the statistical software package SPSS version 17.0.0, with a p-value less than 0.05 being deemed significant and a p-value of less than 0.1 being marginally significant. Marginally significant results are included in line with current statistical trends suggesting that results that fall just below the traditional $p < 0.05$ cut off may still provide valuable insight for future investigation, particularly given the small sample size of the current participant cohort (Pritschet et al. 2016). For further technical details on the conventions and measures presented in the following analysis see Appendix D: Statistical Formulae and Conventions.

6.1 Participant Demographics

To provide a framework for the results in the remainder of this chapter, here an overview of the participants who were successfully recruited and retained throughout a sufficient portion of the evaluation process is provided. As discussed in Subsection 5.2.1, all participants were aged between six and twelve years old, were attending mainstream school when they commenced participation in the study, and had an existing diagnosis of autism spectrum disorder.

Participants were assigned to either the experimental or control condition according to the procedure outlined in Subsection 5.2.2. Seven potential participants completed the pre-test data collection but did not proceed with participation in the study. While only one of these participants explicitly provided a reason for this, which was unforeseen ill health, anecdotal observation indicates that in three other cases the individuals were in the lowest age range and found maintaining the concentration needed to complete the pre-test content quiz challenging, and thus may have had similar difficulties in using the Social Tutor independently. In one of the remaining cases installation of the Social Tutor was difficult, suggesting possible ongoing technical issues may account for their lack of participation, and the other two participants were in the oldest age range and possibly felt the Social Tutor content was too basic for their needs or not sufficiently engaging. These participants were approximately balanced across the intervention groups, with four having been assigned to the control group and three to the experimental group.

Only participants who at a minimum completed both the pre-test and post-test content quiz or both the pre-test and post-test Vineland-II are included in the final cohort of participants. A total cohort of thirty one children participated, with sixteen individuals in the experimental group ($M = 8.81$ years, $SD = 1.83$) and fifteen in the control group ($M = 9.20$ years, $SD = 2.08$). An independent samples t-test was carried out and

found no significant difference in mean age between the two groups ($p = 0.497$, $r = 0.10$). Completeness of data collection tasks varied across participants, a summary of which can be seen in Appendix E.

Participant gender was not specifically controlled for, however in total five of the children in the experimental group were female (31.3%) and three in the control group were female (20.0%). A 2-sample t-test for equality of proportions was carried out and found no significant difference in ratio between the two groups, $\chi^2(1, N = 31) = 0.51$, $p = 0.47$. Similarly, participant socio-economic status was not controlled for but analysis found no unexpected differences between the intervention groups (see Appendix F for details), meaning that the groups were sufficiently balanced according to age, gender and socio-economic status.

Baseline social and communication skills were not directly controlled for beyond specifying in the inclusion criteria that participants must be attending mainstream school and have an existing diagnosis of autism spectrum disorder. As previously discussed, this was done to ensure that the content of the Social Tutor would be an appropriate match for participants' existing skills and needs. However, Vineland-II pre-test results were used to confirm that there were no significant differences initially between the control and experimental groups. Participants' adaptive behaviour scores were compared using unpaired t-test and no significant differences were found between the groups for either the Socialization domain ($p = 0.534$, $r = 0.08$) or the Communication domain ($p = 0.715$, $r = 0.05$). Vineland-II pre-test results also indicated that participants' adaptive levels fit appropriately with the intended recruitment aims for the study, with no participant scoring lower than a "mild deficit". Further demographics are discussed individually for each data collection task in their relevant sections.

6.2 Research Objective 1 - Software Implementation

To address Research Objective 1, namely to 'design and implement an evidence-based Social Tutor software program that can be used by children with autism', first the Social Tutor software was created as described in Chapter 4. To confirm that the software was used as intended and to provide insight into participant interaction patterns, the software automatically collected log data throughout the evaluation period, continuously recording participant actions as they engaged with the Social Tutor. The data and basic analysis is presented here, with discussion of implications to follow in Section 7.1.

From the software log data an array of insightful information about how users interact with the software can be extracted. The key features extracted include the number of days each participant opened the software, the total time they spent engaging with lesson activities, the number of lessons completed across the course of the experimental period, and what lessons each individual attempted. This data is taken from the three weeks of software use only, and does not include time spent during pre-test and post-test data collection activities.

A summary of user log data can be seen in Table 10. It should be noted that users were instructed to use the software once a day, however there was no mechanism preventing them from using it more often. Each time the software was opened a new backup folder was created, indicating total number of sessions with the

software rather than just total number of days, thus if a student opened the software twice on a single day, that day would have two backup folders. However, the technical difficulties some participants experienced necessitated them opening the software multiple times some days even when they only completed a single session of work. Thus 'days' of software use rather than 'sessions' of use was determined to be more reliable.

From the extracted data, several calculated values were also obtained and can be seen in Table 11. Calculated values include the number of lessons students completed per day, the amount of time they spent doing lesson activities per day, and the time spent per lesson. As previously described, students were prompted ten minutes after logging into the software that time was up, and the software forcibly closed once they exited their current activity after fifteen minutes had passed since logging in. This ten to fifteen minute period of software interaction included activities such as choosing which topics and lessons to do and interacting with the rewards system. The times shown in Table 10 and Table 11 indicate time spent actively engaged with the lesson activities specifically.

Table 10: Extracted software log file data summary

Participant group	**No. users**	**Number of days software was used**				**Total time engaging with lessons (minutes)**				**Total number of lessons completed** *(Total lessons available E = 181, C = 115)*			
		Mean	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Experimental	16	11.44	3.27	6	17	72.99	40.28	24.08	183.82	39.38	15.63	18	65
Control	15	10.33	2.89	4	15	89.25	40.84	42.88	192.48	24.60	7.19	15	38

From the software log data it was also possible to determine which lessons each participant attempted. A summary of all lessons attempted by participants in the experimental group and how many participants attempted each one can be seen in Table 12. Any lesson that was not tied to particular educational content, such as the 'welcome' lesson, lessons explaining how to use the interface and lessons covering how to do mind map style activities, were classified under a separate 'admin' topic, while lessons covering content from the implemented curricula were separated into their own topics.

Table 11: Calculated software log file data summary

Participant group	**No. users**	**Number of lessons per day**[1]				**Time spent engaging with lessons per day (minutes)**[2]				**Time spent per lesson (minutes)**[3]			
		Mean	**Min**	**Max**	**SD**	**Mean**	**Min**	**Max**	**SD**	**Mean**	**Min**	**Max**	**SD**
Experimental	16	3.50	1.85	6.50	1.21	6.51	2.29	18.38	3.73	1.80	1.24	2.83	0.48
Control	15	2.50	1.73	5.75	0.95	9.02	3.90	17.40	3.86	3.70	1.96	7.18	1.46

1. Calculated as total number of lessons divided by total number of days the software was used
2. Calculated as total time spent using the software divided by total number of days the software was used.
3. Calculated as total time spent using the software divided by total number of lessons completed.

Table 12: Summary of lesson interaction data

Topic	Objective	*Total lessons available*	Lessons attempted		
			Unique lessons	*Total attempts*	*Percent of unique*
Admin	Admin	5	4	29	5%
Greeting	Basic greeting	17	13	200	17%
	Advanced greeting	40	28	138	36%
Listening and Turn Taking	Basic listening	17	13	136	17%
	Advanced listening	20	13	71	17%
	Turn taking	32	7	56	9%
Beginning, Ending and Maintaining Conversations	Starting conversations	18	0	0	0%
	Ending conversations	5	0	0	0%
	Continuing conversations	27	0	0	0%
	Total	*181*	*78*	*630*	*100%*

From Table 12 it can be seen that the third core topic of 'Beginning, Ending and Maintaining Conversations' was not attempted by any of the participants. This topic required completion of the 'Listening and Turn Taking' topic before it was unlocked and made available, and from inspection of the data it can be seen that this did not occur for any participants. While a longer study period may have resulted in Topic 3 being accessed and is recommended for future evaluations, it was a deliberate design choice to develop more content than participants were expected to get through in three weeks. This was to ensure that there would be enough content for participants to have a choice of activities at all times, and to ensure that participants who worked quickly through the tasks would not run out of activities to access. From Table 12 it can also be seen that more student interaction happened in the first core topic, with 53% of interaction focussed on 'Greeting' and 42% of interaction focussed on 'Listening and Turn Taking'. Student interaction patterns varied across individuals, with a few participants choosing the same topic each time they logged in, but most oscillating between topics. The sequence each participant in the experimental group took through the software can be seen in Table 13, along with a count of objective changes for each participant. For clarity, homework and admin lessons were omitted. As can be seen in Table 13, some participants attempted lessons from as few as two objectives, both within the 'Greeting' topic, while others continually swapped back and forth between topics and objectives.

From this analysis of the software log data it appears that Research Objective 1 was successfully met, with participants being able to use the software successfully, generally spending the intended amount of time on each individual lesson, and completing a satisfactory number of sessions across the intervention period.

Table 13: Sequence of objectives as taken by experimental group participants

Objective changes	Objective Sequence										
2	basic greeting	advanced greeting									
2	basic greeting	advanced greeting									
3	basic greeting	turn taking	listening								
3	basic greeting	turn taking	listening								
3	turn taking	basic greeting	listening								
3	turn taking	basic greeting	advanced greeting								
4	basic greeting	turn taking	listening	basic greeting							
5	turn taking	listening	basic greeting	listening	advanced greeting						
5	basic greeting	turn taking	listening	basic greeting	advanced greeting						
6	basic greeting	turn taking	listening	basic greeting	listening	advanced greeting					
6	turn taking	listening	basic greeting	advanced greeting	turn taking	advanced greeting					
7	basic greeting	turn taking	listening	basic greeting	listening	basic greeting	listening				
7	basic greeting	turn taking	basic greeting	listening	basic greeting	advanced greeting	listening				
7	turn taking	listening	basic greeting	listening	basic greeting	advanced greeting	listening				
8	turn taking	basic greeting	listening	basic greeting	advanced greeting	listening	advanced greeting	listening			
11	basic greeting	turn taking	listening	basic greeting	listening	basic greeting	listening	advanced greeting	listening	advanced greeting	listening

6.3 Research Objective 2 - Changes in Knowledge

To address Research Objective 2 of determining if knowledge of the targeted social skills changed due to interaction with the Social Tutor, participants were presented with an in-software content quiz at all data collection points as described in Subsection 5.2.4: Data Collection Schedule. Specifically, participants were presented with the quiz at pre-test when they logged into their user accounts for the first time, then at immediate post-test which was conducted at the end of the three weeks of software use, and at both two and four months following the immediate post-test. The data collected from pre-test to immediate post-test is considered 'intervention data' that directly addresses Research Objective 2 and this is presented here, while the two and four month follow up results are considered 'longitudinal data' and instead discussed in relation to Research Objective 4 and presented in Section 6.5.

From the automatically recorded data, measures of correctness, accuracy and duration were available both for the content quiz as a whole, as well as for each individual question within the quiz. As each quiz question is a short activity in itself (see Appendix A), correctness and accuracy data is available as a percentage for each question. Duration data is calculated as the difference in total running time between the entry of the question at hand and the previous entry, and is presented as the time taken to complete that question in seconds. Analysis of correctness data at the whole-quiz level is intended to directly address Research Objective 2, with exploratory analysis of correctness, accuracy, and duration data also presented here, including analysis conducted at the whole quiz, topic and question levels as appropriate.

6.3.1 Demographics and Assumption Testing

All participants completed the in-software content quiz at pre-test, with most (90.3%) also completing the immediate post-test. Three participants were either unable to complete the post-test quiz or unable to submit their completed results due to technical difficulties, such as failure of the home computer.

In total twenty eight participants completed the content quiz at both the pre-test and immediate post-test data points, including all participants in the experimental group (N = 16, M = 8.81 years, SD = 1.83, 95% CI [7.84, 9.79]) and 80% of participants in the control group (N = 12, M = 9.42 years, SD = 2.02, 95% CI [8.13, 10.70]), a box plot of the distribution of ages for each group can be seen in Figure 16. In order to conduct t-test analysis of content quiz results, both homogeneity of variance and normality of distribution are assumed, and to test these assumptions Levene's Test for Homogeneity of Variance and Shapiro-Wilk normality tests are used.

First Levene's Test for Homogeneity of Variance was conducted comparing the experimental and control groups and the deviations were found to be non-significant for correctness at pre-test and post-test ($p = 0.226$ and $p = 0.432$) and change in correctness ($p = 0.581$), pre-test and post-test accuracy ($p = 0.964$ and $p = 0.395$) and change in accuracy ($p = 0.313$), pre-test duration ($p = 0.661$) and change in duration ($p = 0.207$), but not post-test duration ($p = 0.017$). Thus the assumption of equal variances can be accepted for all of these datasets with the exception of post-test duration.

Next, Shapiro-Wilk normality tests were conducted and for the control group resulted in non-significance for pre-test correctness ($p = 0.307$), post-test correctness ($p = 0.833$), and for change in correctness ($p = 0.683$), along with pre-test accuracy ($p = 0.582$), post-test accuracy ($p = 0.532$), and change in accuracy ($p = 0.826$) indicating that these datasets are normally distributed. For duration data, change in duration was non-significant ($p = 0.227$) however both pre-test duration and post-test duration did reach significance ($p = 0.015$ and $p = 0.031$ respectively) indicating that these datasets are not normally distributed.

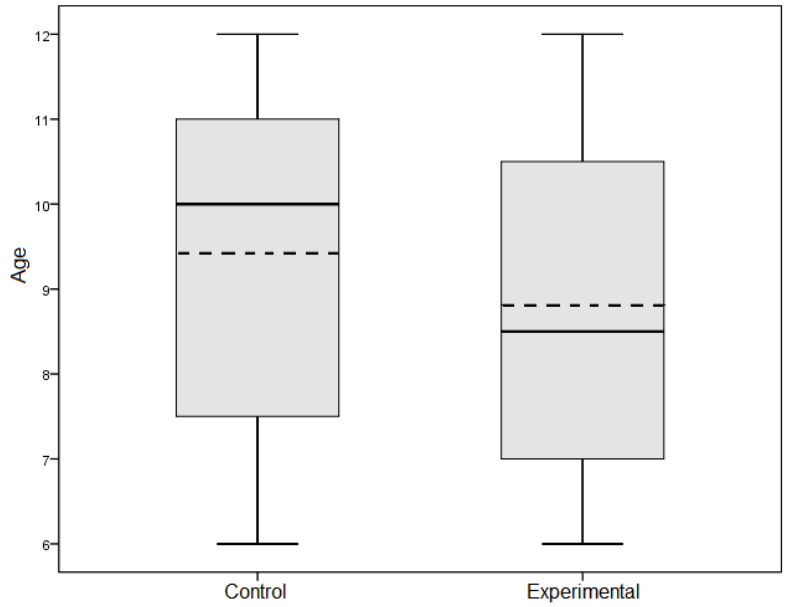


Figure 16: Box plots showing distribution of participant ages by group
Solid line indicates median, dotted line indicates mean.

For the experimental group, Shapiro-Wilk normality tests indicated non-significance for pre-test accuracy ($p = 0.295$), post-test accuracy ($p = 0.195$) and change in accuracy ($p = 0.538$) as well as pre-test duration ($p = 0.180$), post-test duration ($p = 0.346$) and change in duration ($p = 0.373$). For correctness scores, Shapiro-Wilk normality tests indicated non-significance for pre-test and post-test scores ($p = 0.201$ and $p = 0.278$ respectively), however for change in correctness statistical significance was reached ($p = 0.001$) indicating that the assumption of normality must be rejected in this case. Moreover, as can be seen in the quantile-comparison plots in Figure 17, the data appear skewed. This is particularly prominent for the experimental group. Investigating further, box plots of the change in correctness scores for the control group and experimental group shown in Figure 18 suggest two possible outliers in the experimental group data. Removing these results in improved Shapiro-Wilk normality test results where significance is not reached ($p = 0.643$), and thus normality can be assumed under these conditions.

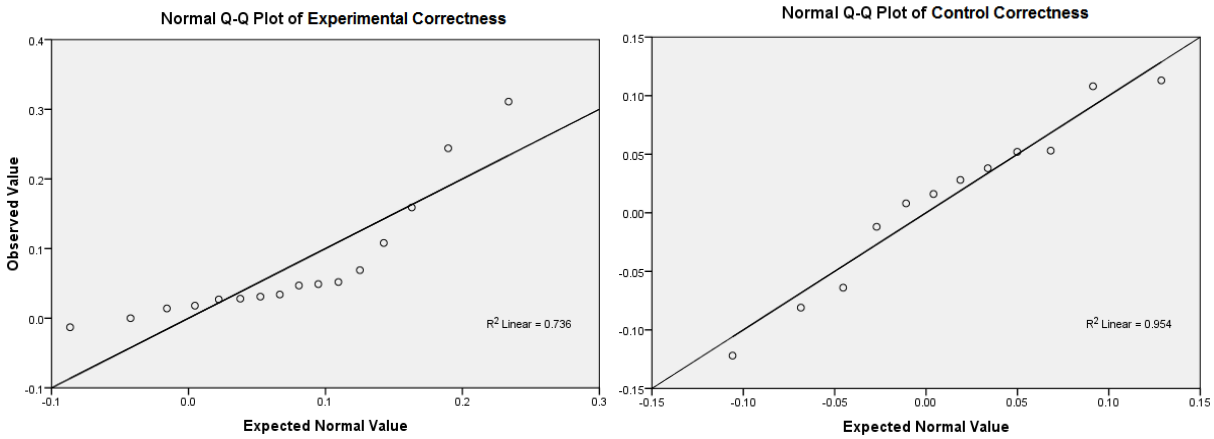


Figure 17: Quantile-comparison plots showing pre-test to post-test change in content quiz correctness. Experimental group (left) and control group (right)

Table 14: Experimental participants grouped into response levels

	Low responding				Average responding								High responding			
Change in score (%)	-1.25	0.00	1.42	1.75	2.67	2.83	3.08	3.42	4.67	4.92	5.17	6.92	10.83	15.92	24.42	31.08
Lessons completed	27	51	65	48	32	53	41	24	52	18	64	49	37	27	18	24
Total time on lessons (s)	2306	4625	11029	3919	4037	6056	3752	3929	5230	1445	6425	6805	3777	3353	1597	1783

One possible post hoc explanation for the identified outliers is that they are part of a high response subgroup, as from observation of the experimental group plot in Figure 17 and histogram in Figure 19, the data for the experimental group appears skewed, with a number of higher performing individuals becoming apparent. The hypothesis here is that these individuals may have needed less time with the software than their peers in order to benefit from it. Investigating this further, participants were classified as high, average and low responders according to their content quiz correctness scores. To achieve this, participant data was ordered according to change in correctness from pre-test to post-test, and assessed for the presence of any patterns. A distinct 'jump' in scores becomes apparent: where most participants had improved their score by 5% or less and were quite continuous in their scores, there was a large gap and then four 'high responding' participants who obtained an improvement of 10% or more. Two participants did fall within the gap, however observation of their log data indicated that they had both used the software an above-average amount, whereas the four high responding participants had used it a below-average amount as measured by both number of lessons completed and total time spent actively doing lessons with the software. A summary of the experimental group participants' change in scores, total number of lessons completed, and total time doing lessons in the software can be seen in Table 14.

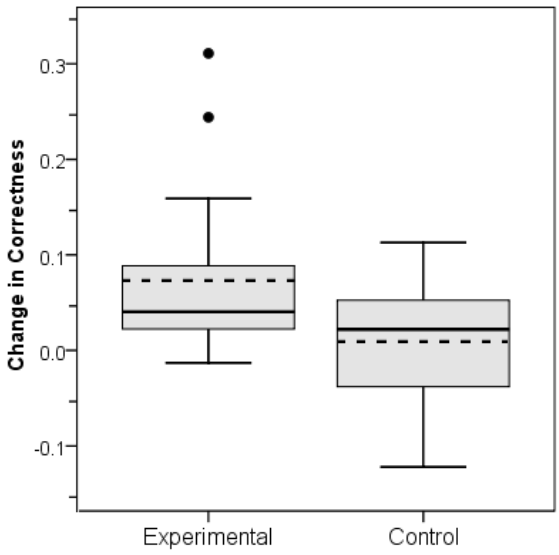


Figure 18: Change in content quiz correctness score from pre-test to post-test by group with median denoted by a solid line and mean by a dotted line

The Shapiro-Wilk test of normality was non-significant for change in correctness for the high responding group (p = 0.854), the low responding group (p = 0.558), and the average responding group (p = 0.325), however Levene's Test for Homogeneity of Variance found deviations to be significant between the three groups (p = 0.006). Given the small number of participants in each group this is not unexpected, however it does mean that the assumptions are not met for ANOVA analysis of the change in correctness datasets.

Shapiro-Wilk tests were non-significant for distribution of age in both the control (p = 0.355) and experimental groups (p = 0.241), and Levene's Test for Homogeneity of Variance was also found to be non-significant between the two groups (p = 0.742). With all values included, one way ANOVA found no significant difference in mean age between the experimental and control groups (N = 16, M = 8.81 years, SD = 1.83, 95% CI [7.84, 9.79] and N = 12, M = 9.42 years, SD = 2.02, 95% CI [8.13, 10.70] respectively, $F(1, 28) = 0.68$, $p = 0.42$, $\eta_p^2 = 0.03$), and likewise once outliers were removed ($F(1, 26) = 1.008$, $p = 0.325$, $\eta_p^2 = 0.04$), however the mean age for the experimental group did drop slightly (N = 14, M = 8.64 years, SD = 1.91, 95% CI [7.54, 9.74]).

The matched pairs process balanced the groups by age, but did not explicitly balance by participant gender. Equality tests were therefore conducted to assess whether balance had been obtained. With all values included, a 2-sample test for equality of proportions found no significant difference in the ratio of the groups, $\chi^2(1, N = 28) = 0.13$, $p = 0.72$, V = 0.07, with the control group having 25.0% female participants and 75.0% male, and the experimental group having 31.3% female participants and 68.7% male. With the same two outliers identified in Figure 18 removed, the experimental group had 29.0% female participants and 71.0% male participants, and again a 2-sample test for equality of proportions found no significant difference in gender ratio between the groups $\chi^2(1, N = 26) = 0.04$, $p = 0.84$, V = 0.04.

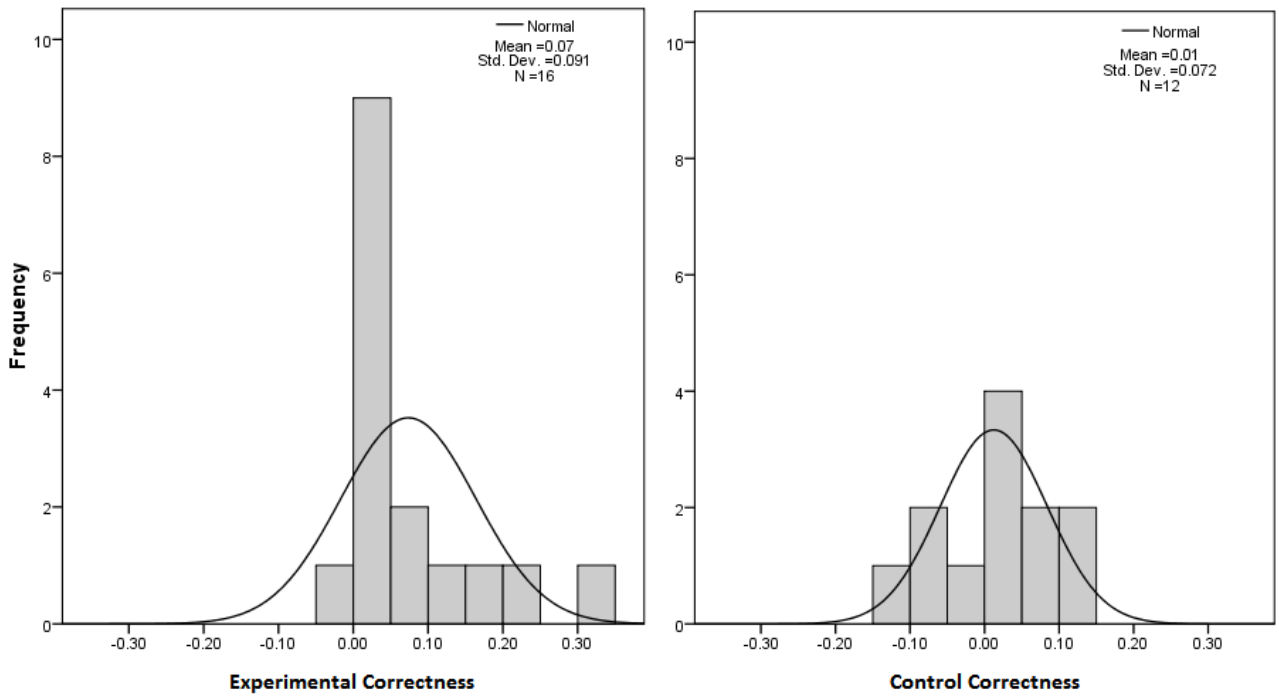


Figure 19: Distribution of correctness scores for the experimental and control groups

Finally, a summary of the correctness, accuracy and duration data can be seen in Table 15. Interestingly, it can be seen that pre-test correctness results for the control group are approximately 10% higher on average than pre-test correctness results for the experimental group, however independent samples t-test indicated that this was not significant ($p = 0.226$, $d = 0.33$). Further breakdown of the data in Table 15 follows.

Table 15: Summary of correctness, accuracy and duration data from pre-test and post-test content quiz

		Experimental (all values) M (SD)	Experimental (outliers removed) M (SD)	Control M (SD)
Correctness (%)	**Pre-test**	71.51 (16.10)	74.81 (15.42)	81.20 (12.15)
	Post-test	78.87 (13.76)	78.38 (14.38)	82.34 (11.15)
	Change	7.36 (9.05)	3.57 (3.10)	1.14 (7.16)
Accuracy (%)	**Pre-test**	69.46 (12.31)	70.85 (11.64)	74.87 (12.84)
	Post-test	75.85 (13.63)	75.75 (14.98)	78.39 (11.43)
	Change	6.4 (8.18)	4.90 (6.84)	3.52 (6.34)
Duration (s)	**Pre-test**	90.35 (33.51)	96.81 (32.43)	93.04 (40.41)
	Post-test	62.36 (20.36)	62.15 (20.76)	79.32 (38.24)
	Change	-27.99 (24.66)	-34.66 (22.37)	-13.72 (46.54)

6.3.2 Primary Analysis

To address Objective 2 and "determine if knowledge of targeted social skills changes due to interaction with the Social Tutor" and to address part 1 of Objective 4 and "determine if any changes in knowledge or behaviour are maintained after software use ends", participant correctness scores at the whole-quiz level were analysed. Paired t-tests have been chosen as the most appropriate statistical test to apply given the small sample sizes and underlying assumptions discussed in subsection 6.3.1, with complete results for each comparison after applying correction for multiple comparisons provided in Appendix G. The longitudinal data related to Objective 4 is presented in Chapter 6.5 below.

The total percentage correctness was taken from the pre-test and post-test content quiz data for all twenty eight included participants and the difference between them calculated. As anticipated, the results for the control group showed only a very small mean change from pre-test to post-test ($M = 1.14\%$, $SD = 7.16$, 95% CI [-3.41, 5.69]), and paired t-test showed this to be statistically insignificant ($p = 0.799$, $d = 0.08$). In contrast, analysis of data from the experimental group resulted in a slightly larger mean change from pre-test to post-test ($M = 7.36\%$, $SD = 9.05$, 95% CI [2.54, 12.19]), and paired t-test with False Discovery Rate (FDR) correction indicated statistical significance ($p = 0.01$, $d = -0.73$). Raw and FDR-corrected p-values can be seen in Appendix G.

For completeness, a repeated measures two-way ANOVA was also conducted. This indicated a significant interaction in correctness scores between group and test period (i.e. pre-test, post-test, 2 month post-test or 4 month post-test) $F(3,30) = 3.421$, $p = 0.03$, however simple main effects analysis found no significant differences between the control and experimental groups or between test periods. As previously discussed, this is likely due to the small sample sizes and supports the use of t-tests for further analysis of this data.

6.3.3 Exploratory Analysis

Following analysis of the whole-quiz data required for primary Objectives 2 and 4, further exploratory analysis of the available content quiz data was conducted to investigate the possible presence of any patterns or characteristics that could be beneficial to future development of the software. Of particular interest was identifying characteristics of users who benefitted most from the software, identification of any problem areas within the software itself or research methodology, and identifying areas where more data should be gathered in future research. As the following analyses are exploratory in nature only, no correction for multiple comparisons has been applied.

Correctness Overall

As previously discussed and as can be seen in Figure 18, two outliers were identified in the experimental group results. Removing these outliers from the data analysis results in a smaller mean change for the experimental group ($M = 4.45\%$, $SD = 4.44$, 95% CI [1.89, 7.02]) with the results remaining statistically significant by paired t-test ($p = 0.002$, $d = -1.00$). The assumption of normal distribution was not met for the experimental group change in correctness dataset, so the non-parametric Wilcoxon rank sum test was employed to compare change in the control group to change in the experimental group, however this did not reach significance ($p = 0.173$, $r = 0.26$). When the two identified outliers are removed from the experimental dataset the normality assumption is met and the more sensitive independent samples t-test can be run, however even under these conditions there is no significant difference detected in change in correctness scores between the control and experimental groups ($p = 0.164$, $d = -0.28$). Again reiterating previous analysis, one possible explanation for these 'outliers' is that they are part of a high responding group, and as such experimental group participants can be organised into three response levels as indicated in Table 14: high, average and low responding. When organised in this manner the underlying assumption regarding homogeneity of variance required to conduct ANOVA analysis is not met, therefore Kruskal-Wallis rank sum tests are used. As expected and confirming the appropriateness of the groupings, a significant difference was found between the three response subgroups and the control group for change in overall content quiz scores from pre-test to immediate post-test ($\chi^2 = 12.98$, $p = 0.005$) with mean change of 20.56% for high responders, 4.21% for average responders, 0.48% for low responders, and 1.14% for the control group (for more detail see Appendix H: Table 44).

As the assumption of equal variances is not met for the response subgroups, post hoc analysis was conducted using pairwise Wilcoxon rank sum tests. These indicated significant differences between high responders

and average responders, average responders and low responders, high responders and low responders, and high responders and the control group ($p = 0.004$, $r = 0.55$; $p = 0.004$, $r = 0.55$; $p = 0.029$, $r = 0.58$; and $p = 0.002$, $r = 0.49$ respectively), but not between the control group and average responders or the control group and low responders ($p = 0.315$, $r = -0.16$ and $p = 0.521$, $r = 0.13$ respectively).

Correctness by Topic

The activities within the content quiz were designed to directly address the topics available to students in the Social Tutor, thus questions one to four of the quiz were designed to address 'Topic 1: Greeting' within the software, questions five to eight were designed to address 'Topic 2: Listening and Turn-Taking,' and questions nine to twelve were designed to address 'Topic 3: Beginning, Ending and Maintaining Conversations'. The content quiz data was therefore broken down into these datasets for finer grained analysis. Given that no student unlocked Topic 3, a comparison between Topic 1 and Topic 2 combined and Topic 3 alone was also conducted.

Each dataset was assessed for the assumptions of normality and equal variance, however several failed one or both of these tests making non-parametric statistical techniques most appropriate, therefore Wilcoxon signed rank tests have been used to compare pre-test results to post-test results for statistical significance. A summary can be seen in Table 16, with a detailed table including p-values, effect sizes and confidence intervals provided in Appendix F: Table 44.

As anticipated, control group results for all subsections indicated a very small mean improvement that was relatively consistent regardless of question breakdown and did not reach statistical significance. Also as expected, when questions nine to twelve were analysed separately the results did not reach significance for any group or subgroup. When only questions one to eight are considered it can be seen that experimental participants on average did make a small but significant improvement from pre-test to post-test ($p = 0.013$, $r = -0.44$) and that this improvement is slightly higher than when the results of all questions are included. This trend holds true for average responding students ($p = 0.017$, $r = -0.60$). Interestingly, low responding students performed more poorly at post-test than pre-test for Topics 1 and 2 combined, although this did not reach significance and appears to be influenced by the results of Topic 2 alone.

In Table 16 it can also be seen that when further broken down into topics, the mean improvement from pre-test to post-test for questions one to four, the questions which align with Topic 1, is significant for the experimental group ($p = 0.041$, $r = -0.36$) and in this case the mean improvement is also larger than any other subsection of questions ($M = 9.67\%$, $SD = 16.01$, 95% CI [1.14, 18.2]). This same trend holds true for the high responding subgroup, although no results reached significance in this case. Interestingly, of all the topics, the experimental group performed most poorly on questions five to eight, those aligning with Topic 2. While all other groups and subgroups exhibited a small to moderate mean improvement for Topic 2, the low responding subgroup actually displayed a decrease in mean score from pre-test to post-test ($M = -6.81\%$, $SD = 1.68$, 95% CI [-9.48, -4.15]).

Table 16: Mean change in content quiz correctness scores from pre-test to post-test

	Experimental group N = 16	High responders N = 4	Average responders N = 8	Low responders N = 4	Control group N = 12
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
All questions	7.36% (9.05) *	20.56% (8.98) .	4.21% (1.47) *	0.48% (1.38)	1.14% (7.16)
Topic 1 & 2 (questions 1-8)	7.65% (10.77) *	21.97% (9.68) .	5.80% (4.27) *	-2.97% (3.27) .	1.05% (7.62)
Topic 1 (questions 1-4)	9.67% (16.01) *	32.38% (7.15) .	2.72% (10.77)	0.88% (6.42)	1.50% (14.10)
Topic 2 (questions 5-8)	5.63% (14.99)	11.56% (16.20)	8.88% (15.73)	-6.81% (1.68) .	0.60% (8.64)
Topic 3 (questions 9-12)	6.80% (14.64)	17.75% (21.59)	1.03% (10.66)	7.38% (9.52)	1.31% (12.76)

*Note: * denotes statistical significance ($p < 0.05$), . denotes marginal significance ($0.05 < p < 0.1$)*

Correctness by Question

Each question in the content quiz consists of a short activity, and while the assessment algorithm differs for each question depending on the activity type, every question has a possible correctness score of up to 100%. Given this, Wilcoxon signed rank test analysis was conducted comparing the pre-test and post-test data for each individual question within the content quiz for the control group, the experimental group, and all response level subgroups, however no individual question reached significance.

Accuracy Overall and By Topic

While the method used to measure accuracy differs according to the particular task type and is not applicable in some contexts such as watching a video or simply listening to instructions, it provides an indication of how many incorrect moves were made before the task reached completion. For example in a drag and drop style activity accuracy measures how many times responses were placed into incorrect categories prior to exiting the task, either due to completion or manually moving to another activity. The control and experimental groups and all response level subgroups displayed a mean increase in accuracy from pre-test to post-test. Paired t-tests comparing pre-test to post-test reached significance for the experimental group ($p = 0.007$, $d = -0.78$) but not the control group ($p = 0.081$, $d = -0.56$), and independent samples t-test found no significant difference in change scores between the experimental and control groups ($p = 0.322$, $d = -0.19$). As the response subgroups failed to meet the normality assumption for a t-test, Wilcoxon rank sum tests were used to compare change in accuracy between the subgroups, however statistical significance was not reached for any pairwise comparison. Similarly, pre-test to post-test results were compared for the response subgroups, with no subgroup reaching significance.

Aligning with the trends found in correctness data, the control group displayed the lowest mean improvement in accuracy overall ($M = 3.52\%$, $SD = 6.34$, 95% CI [-0.51, 7.55]) and the experimental group

as a whole outperformed them noticeably (M = 6.40%, SD = 8.18, 95% CI [2.04, 10.75]). The high responding subgroup displayed the highest mean improvement overall (M = 9.90%, SD = 11.43, 95% CI [-8.29, 28.08]), with the average and low responding subgroups performing similarly to each other (M = 4.69%, SD = 7.16, 95% CI [-1.30, 10.67] and M = 6.31%, SD = 7.75, 95% CI [-6.01, 18.64] respectively). For completeness, the data was broken down by topic and Wilcoxon signed rank tests were conducted comparing pre-test to post-test for each group and subgroup, and Wilcoxon rank sum tests used to compare change in score between groups, however again no significant results were found.

Duration Overall

The time spent on content quiz questions at pre-test and immediate post-test was analysed for all twenty eight eligible participants, with the difference in total duration from pre-test to post-test compared. Again, the assumptions of normality and equal variance were tested, with change in duration datasets meeting the assumptions for the control and experimental groups, but both the pre-test and post-test duration datasets failing these tests, along with the smaller response subgroup datasets. Parametric t-tests and non-parametric Wilcoxon tests have been used to test statistical significance as appropriate.

Using Wilcoxon signed rank tests comparison of pre-test duration to post-test duration found no statistically significant difference for the control group separately ($p = 0.182$, $r = -0.27$), however significant differences were revealed when all data was combined ($p = 0.016$, $r = -0.32$) and for the experimental group alone ($p = 0.020$, $r = -0.41$). This indicates that on average experimental group participants were faster at answering the content quiz questions at post-test than they had been at pre-test (N = 16, M = -28.00 sec, SD = 24.66, 95% CI [-41.13, -14.85]), and this was also true of the combined cohort (N = 28, M = -21.88 sec, SD = 35.66, 95% CI [-35.70, -8.05]). Although the control group did not reach significance, a trend towards shorter durations was observed (N = 12, M = -13.72 sec, SD = 46.54, 95% CI [-43.29, 15.85]). These observations can be seen in Figure 20. Independent samples t-test found no significant difference between the change in scores for the control group and experimental group ($p = 0.304$, $d = -0.20$).

The data was further analysed to assess whether there was any difference in duration according to response subgroups within the experimental group, with a summary of performance by group shown in Table 17. A Kruskal-Wallis rank sum test comparing each response subgroup and the control group found a significant difference between the groups ($p = 0.032$). Following this, pairwise Wilcoxon rank sum tests revealed that significant differences exist between the high responding and average responding subgroups ($p = 0.007$, $r = 0.55$) and between the high responding and low responding subgroups ($p = 0.021$, $r = -0.58$) however no significant interactions were identified between the control group and the high, average or low responding groups ($p = 0.115$, $r = -0.28$; $p = 0.217$, $r = -0.20$; and $p = 0.182$, $r = -0.24$ respectively) or between low responding and average responding subgroups ($p = 0.396$, $r = -0.17$).

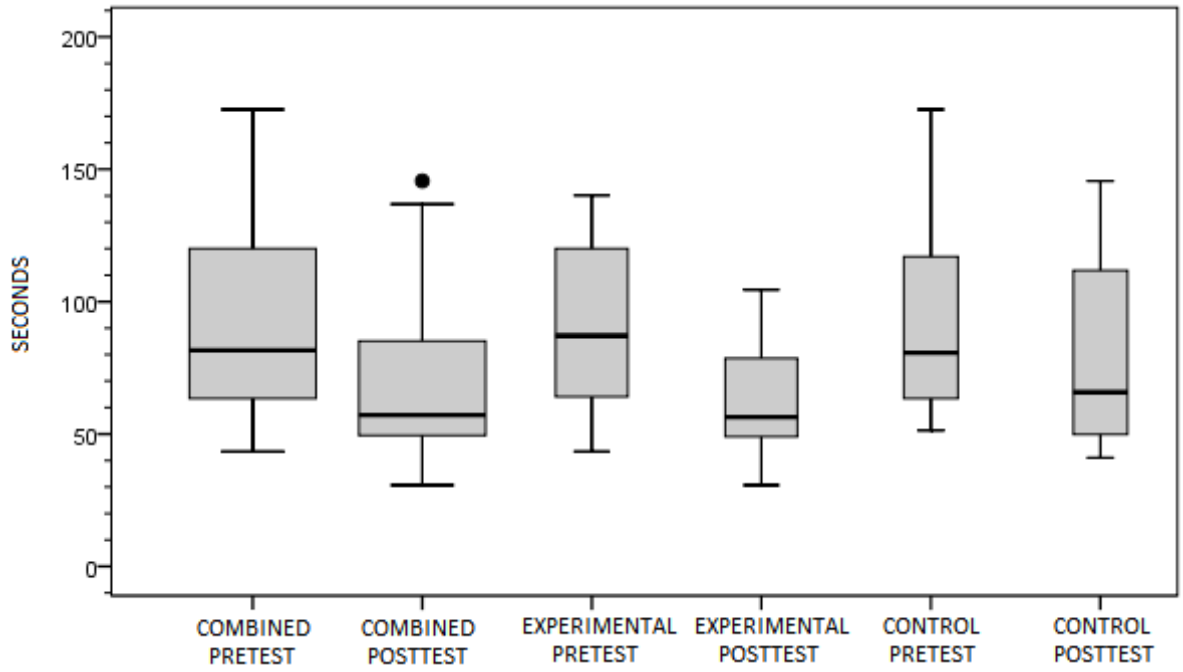


Figure 20: Time spent on content quiz questions by data collection point and group

As can be seen from Table 17, the high responding group showed a much smaller improvement in question answering duration than any other group or subgroup. It can also be seen that half of this subgroup answered more slowly at post-test than pre-test, while all participants in the average and low responding subgroups answered more quickly at post-test than pre-test.

Table 17: Summary of mean change in duration in seconds by response subgroup

		Experimental group				Control group
		High	Average	Low	Total	
Count	N	4	8	4	16	12
	Faster	2	8	4	14	8
	Slower	2	0	0	2	4
M (SD)		-1.29 (5.55)	-31.25 (15.11)	-48.17 (30.92)	-28.00 (24.66)	-13.72 (46.54)
95% CI		-10.13, 7.55	-43.88, -18.62	-97.36, 1.03	-41.13, -14.85	-43.29, 15.85

Duration by Topic

To determine if the observed improvements in response speed from pre-test to post-test were linked to a particular topic, mean change in duration was calculated for each topic and Wilcoxon signed rank tests were used to identify if statistically significant differences were present between pre-test and post-test durations. These tests were performed on the control group, the experimental group, and all response level subgroups separately, with a summary shown in Table 18 and a detailed table including p-values, effect sizes and confidence intervals available in Appendix I: Table 45. As with the earlier Table 16, it should be noted that the data fails to meet the assumptions for t-test use, making the Wilcoxon signed rank test the most appropriate analysis tool here.

Table 18: Comparison of change in content quiz question answering duration by topic

	Experimental group N = 16	High responders N = 4	Average responders N = 8	Low responders N = 4	Control group N = 12
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
Topic 1 & 2 (questions 1-8)	-26.03 (27.96) *	2.72 (6.22)	-26.16 (17.14) *	-54.53 (31.62) .	-7.30 (55.40)
Topic 1 'Greeting' (questions 1-4)	-18.68 (26.93) .	4.38 (23.43)	-20.44 (20.49) *	-37.75 (29.91) .	-14.56 (28.06) .
Topic 2 'Listening and Turn Taking' (questions 5-8)	-33.50 (36.39) *	1.06 (15.58)	-31.88 (19.37) *	-71.32 (44.70) .	-0.04 (100.87)
Topic 3 'Good Conversations' (questions 9-12)	-31.90 (27.56) *	-9.31 (19.86)	-41.43 (25.84) *	-35.44 (30.08)	-26.56 (36.83) *

** denotes statistical significance (p < 0.05), . denotes marginal significance (0.05 < p < 0.1)*

From Table 18 it can be seen that for the experimental group change in duration reached significance for Topic 2 and Topic 3, as well as the combined Topic 1 and 2 condition ($p = 0.02$, $r = -0.41$, $p = 0.023$, $r = -0.40$ and $p = 0.023$, $r = -0.40$ respectively). This trend was also found for the average responding subgroup ($p = 0.017$, $r = -0.60$; $p = 0.012$, $r = -0.63$ and $p = 0.012$, $r = -0.89$ respectively). In addition, the average responding subgroup reached significance for Topic 1 ($p = 0.017$, $r = -0.60$) however for the experimental group as a whole Topic 1 was only marginally significant ($p = 0.056$, $r = -0.34$). No topic reached significance for either the high or low responding subgroups, likely due to the very small sample sizes ($N = 4$ for both subgroups). For the control group, only Topic 3 reached significance ($p = 0.023$, $r = -0.46$).

In all of these cases the mean change was negative, indicating that participants were faster at answering questions at post-test than at pre-test. The experimental group as a whole and average responding subgroup performed roughly on par with each other, with the experimental group displaying an improvement of approximately 19 to 34 seconds per question and the average responding subgroup between 20 and 41 seconds per question. The low responding subgroup displayed a somewhat larger improvement, ranging from approximately 35 to 71 seconds. In contrast, the high responding subgroup displayed an improvement only in Topic 3, being just 9 seconds faster per question on average. For Topics 1 and 2 the high responding subgroup actually reduced their speed by 4.38 and 1.06 seconds respectively, although no topic reached significance for this subgroup. In contrast, the control group remained approximately the same speed for Topic 2 (0.04 seconds improvement) and displayed an improvement in speed for Topic 1 of approximately 15 seconds and for Topic 3 of approximately 27 seconds, performing similarly to the experimental group as a whole for these two topics.

Significant differences were identified from pre-test to post-test within several groups, and a Kruskal-Wallis test between the control group and all three response level subgroups also reached significance, suggesting differences exist between groups as well. Pairwise Wilcoxon rank sum tests revealed significant differences between high and average responders ($p = 0.007$, $r = -0.55$) and high and low responders ($p = 0.021$, $r = -0.58$) but not between average and low responders ($p = 0.396$, $r = -0.25$) or between the control group and high, average or low responders ($p = 0.115$, $r = -0.28$; $p = 0.217$, $r = -0.20$ and $p = 0.182$, $r = -0.24$ respectively). This aligns with the previous findings given that high responders were the only group not to display an improvement in question answering speed overall. To explore all of these findings further, the data was next analysed by individual question.

Duration by Question

To investigate the contribution of each individual question, Wilcoxon signed rank tests were used to compare the pre-test and post-test durations for both the experimental group and the control group. As can be seen in Table 19, for the experimental group statistical significance was reached for all questions except questions two, four, and five, and for the control group it was reached for questions four, eight, eleven, and twelve. A more detailed table showing p-values, effect sizes and associated pre- and post-test values can be seen in Appendix I: Table 46.

Table 19: Change in content quiz duration in seconds from pre- to post-test by questions and group

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
Experimental	-19.75 *	-19.31 .	-25.69 *	-9.50	-8.81	-30.06 *	-38.50 *	-56.63 *	-34.06 *	-37.38 *	-21.94 *	-34.25 *
Control	-13.83	-14.67	-5.67	-24.08*	66.92	-12.92	24.33	-78.50 *	-23.83	-19.58	-41.75 *	-21.08 *

** Denotes statistical significance ($p < 0.05$), . denotes marginal significance ($0.05 < p < 0.1$)*

This process was repeated for the response level subgroups of the experimental group, and from Table 20 it can be seen that no question reached statistical significance for the high or low responding subgroups, most likely strongly influenced by the small sample sizes ($N = 4$ in both cases), however for the average responding group questions two, three, six, eight, nine, ten, eleven, and twelve all reached significance. Again, a more detailed table is available in Appendix I: Table 47.

Table 20: Change in content quiz duration from pre- to post-test by question and response subgroups

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
High responders	-23.25	20.00	0.75	20.00	-13.50	14.50	2.50	0.75	-5.25	-19.75	7.50	-19.75
Average responders	-16.50 .	-21.63 *	-29.00 *	-14.63	2.50	-33.00 *	-47.50 .	-49.50 *	-43.25 *	-49.25 *	-36.75 *	-36.50 *
Low responders	-22.75	-54.00	-45.50 .	-28.75 .	-26.75	-68.75 .	-61.50 .	-128.25 .	-44.50 .	-31.25	-21.75	-44.25 .

** Denotes statistical significance ($p < 0.05$), . denotes marginal significance ($0.05 < p < 0.1$)*
Note: values shown are indicated in seconds.

The content quiz questions themselves can be seen in Appendix A, and from observing these it becomes apparent that they can be categorised as either 'high' or 'low' complexity based on the number of variables in each question, with low complexity questions having three or four variables participants must respond to, and high complexity having five to ten variables. A summary of this categorisation can be seen in Table 21.

For the control group, of the four questions that reached significance for improvement in duration, two are high complexity and two are low complexity. For the experimental group, three are low complexity and the remaining six are high complexity, this being all of the possible high complexity questions. For the average responding subgroup four low complexity and three high complexity questions reached significance.

Analysis by Demographics

As described in Subsection 5.2.2, participants were allocated to either the experimental or control group using a matched pairs process. While not explicitly matched on gender, a reasonable balance was obtained across both the experimental group and control group as described in Section 6.1.

Wilcoxon rank sum tests were conducted between gender and change in content quiz scores for both the experimental group and the control group, however no statistically significant results were found ($p = 0.692$, $r = -0.07$ and $p = 0.405$, $r = -0.17$ respectively), and this held true when the tests were repeated between gender and change in content quiz answering duration, again with no significant results found for either the experimental or control group ($p = 0.955$, $r = -0.01$ and $p = 0.229$, $r = -0.25$ respectively). However, when the tests were repeated for accuracy data, a significant result was uncovered for the control group ($p = 0.033$, $r = -0.44$) but not for the experimental group ($p = 0.496$, $r = -0.12$).

Participants were explicitly matched by age according to three age 'buckets' with six to eight year olds in the youngest group, nine and ten year olds in the middle group, and eleven and twelve year olds in the oldest group. A summary of the mean correctness, accuracy and duration for each age bucket can be seen in Table 22. In order to compare the three age buckets, Kruskal-Wallis rank sum tests were employed.

As the Kruskal-Wallis rank sum test assumes equal variances, Levene's Test for Homogeneity of Variance was conducted and for the control group deviations were found to be non-significant for change in correctness ($p = 0.141$) and change in accuracy ($p = 0.308$) indicating approximately equal variances, but did reach significance for change in duration ($p = 0.018$). For the experimental group, non-significance was found for change in accuracy ($p = 0.227$) and change in duration ($p = 0.171$) but significance was reached for change in correctness ($p = 0.001$). Thus the assumption of equal variances across age buckets must be rejected for change in duration in the control group and change in correctness for the experimental group. This may be primarily due to the small sample sizes in the various buckets and may not indicate that with a larger sample that the assumption of equal variances would not be met, however care should be taken when interpreting results from the Kruskal-Wallis rank sum test in the identified cases.

Table 21: Summary of type and complexity of content quiz questions

Topic	Question	Type	Complexity	Statistically significant change in duration from pre-test to post-test by test type				
				Experimental	High responders	Average responders	Low responders	Control
1	1	Long question list	High	X				
	2	Simple word grid	Low			X		
	3	Simple word grid	Low	X		X		
	4	Short question list	Low					X
2	5	Short question list	Low					
	6	Long question list	High	X				
	7	Complex word grid	High	X				
	8	Complex word grid	High	X		X		X
3	9	Short question list	Low	X		X		
	10	Drag and drop	High	X		X		
	11	Simple word grid	Low	X		X		X
	12	Drag and drop	High	X		X		X

When both groups were combined and Kruskal-Wallis rank sum test was applied to the experimental group, significance was reached for change in test scores ($p = 0.045$) and change in duration ($p = 0.029$) but was not reached for change in accuracy ($p = 0.118$). For the control group significance was not reached for change in test scores ($p = 0.694$), change in duration ($p = 0.246$) or change in accuracy ($p = 0.309$).

Table 22: Summary of mean duration, correctness and accuracy data by age bucket

		Experimental			Control		
		Change in dur. (s)	Change in corr. (%)	Change in acc. (%)	Change in dur. (s)	Change in corr. (%)	Change in acc. (%)
Youngest (6-8 years)	**M (SD)**	-42.85 (24.43)	2.65 (2.62)	3.81 (6.67)	-17.67 (18.54)	-0.85 (10.48)	8.33 (7.84)
	95% CI	-63.28, -22.43	0.46, 4.83	-1.76, 9.39	-47.16, 11.83	-17.53, 15.82	-4.14, 20.80
Middle (9-10 years)	**M (SD)**	-6.27 (12.45)	17.75 (12.13)	5.44 (11.77)	21.52 (55.60)	4.10 (5.15)	0.92 (2.41)
	95% CI	-26.08, 13.54	-1.55, 37.05	-13.29, 24.16	-66.96, 110.00	-4.09, 12.30	-2.92, 4.76
Oldest (11-12 years)	**M (SD)**	-19.98 (13.79)	6.42 (6.49)	12.52 (4.78)	-45.02 (39.02)	0.17 (5.79)	1.31 (5.81)
	95% CI	-41.92, 1.96	-3.91, 16.75	4.91, 20.13	-107.11, 17.06	-9.04, 9.38	-7.93, 10.55

To further investigate the significant differences identified using the Kruskal-Wallis tests, pairwise comparisons were made between the age bucket groups. Wilcoxon rank sum tests were applied for both the experimental group and the control group, with results of all comparisons shown in Table 23.

Table 23: Comparison of correctness, accuracy and duration data by age bucket

			Change in correctness	Change in accuracy	Change in duration
Experimental	Youngest vs. Middle	12	p = 0.017* r = -0.49	p = 0.671 r = -0.08	p = 0.017* r = -0.49
	Middle vs. Oldest	8	p = 0.149 r = -0.36	p = 0.248 r = -0.29	p = 0.149 r = -0.36
	Youngest vs. Oldest	12	p = 0.308 r = -0.21	p = 0.027* r = 0.45	p = 0.126 r = -0.31
Control	Youngest vs. Middle	8	p = 0.564 r = -0.14	p = 0.248 r = -0.29	p = 0.386 r = -0.22
	Middle vs. Oldest	8	p = 0.386 r = -0.22	p = 1.000 r = 0.00	p = 0.149 r = -0.36
	Youngest vs. Oldest	8	p = 1.000 r = 0.00	p = 0.149 r = -0.36	p = 0.248 r = -0.29

** Denotes statistical significance ($p < 0.05$)*

Confirming earlier findings, no comparison reached significance for the control group. For the experimental group, change in correctness reached significance for comparison between the youngest and middle age buckets ($p = 0.017$, $r = -0.49$). Referring back to Table 22, it can be seen that for the experimental group, mean change in correctness in the youngest age bucket (range: 6-8 years, N = 8, M = 2.65%, SD = 2.62, 95% CI [0.46, 4.83]) and the oldest age bucket (range: 11-12 years, N = 4, M = 6.42%, SD = 6.49, 95% CI [-3.91, 16.75]) are similar, however the youngest age bucket has double the number of participants. In comparison, the middle age bucket has the largest mean improvement in correctness (range: 9-10 years, N = 4, M = 17.75%, SD = 12.13, 95% CI [-1.55, 37.05]) with the same number of participants as the oldest age bucket. Caution must be taken when interpreting these results not only due to the small sample sizes but also due to the presence of outliers, as two of the previously identified outliers were aged 10 years old and thus are both included in the best performing age bracket. Removing these from the sample results in a much smaller, although still top-performing, mean for the middle age bucket (range: 9-10 years, N = 2, M = 7.75%, SD = 4.36, 95% CI [-31.43, 46.93]).

Change in accuracy reached significance for comparison of the youngest and oldest age buckets ($p = 0.027$, $r = 0.45$), with results indicating that the oldest group improved their accuracy from pre-test to post-test significantly more than the youngest group (N = 4, M = 12.52%, SD = 4.78, 95% CI [4.91, 20.13] and N = 8, M = 3.81%, SD = 6.67, 95% CI [-1.76, 9.39] respectively). Change in duration also reached significance for comparison of the youngest and middle age buckets ($p = 0.017$, $r = -0.49$). The youngest group showed the most improvement in duration from pre-test to post-test (N = 8, M = -42.85 sec, SD = 24.43, 95% CI [-63.28, -22.43]), with the oldest age bucket display approximately half this improvement (N = 4, M = -19.98 sec, SD = 13.79, 95% CI [-41.92, 1.96]) and the middle age bucket showing only a small improvement (N = 4, M = -6.27 sec, SD = 12.45, 95% CI [-26.08, 13.54]).

6.3.4 Summary

Analysis of the content quiz data indicates that use of the Social Tutor has led to an improvement in social skills knowledge for participants in the experimental group, but not for those in the control group. Post-hoc analysis also revealed the presence of response subgroups, with participants being grouped as either high, average or low responding. Participants identified as 'high responding' appear to have made notably higher gains in social skills knowledge from use of the Social Tutor software.

6.4 Research Objective 3 - Changes in Behaviour

To address Research Objective 3, namely to 'determine if behaviour of targeted social skills changed due to interaction with the Social Tutor', participants' caregivers were asked to complete the Vineland Adaptive Behaviour Scale, 2nd Edition (Vineland-II) at all data collection points as described in Subsection 5.2.4: Data Collection Schedule. At the pre-test data point, caregivers completed the Vineland-II on a provided iPad while the software was being installed by the researcher, prior to their child completing the questionnaire or pre-test content quiz. This meant the researcher was present to clarify any queries relating to the Vineland-II items or process. At the immediate post-test data collection point and beyond, caregivers were emailed a hyperlink and instructions so they could complete the Vineland-II independently via their home computer. In the discussion below, pre-test to immediate post-test data is considered 'intervention data' and the two and four month follow up results are considered 'longitudinal data'.

As described in Subsection 5.2.3: Selection and Design of Tools, only the Vineland-II subdomains that directly aligned with the content being taught in the Social Tutor were administered. These were the Receptive and Expressive subdomains of the Communication domain, all subdomains of the Socialization domain, and all of the Maladaptive Behaviours domain. To ensure accuracy and consistency, a small helper script was written to calculate the raw score for each subdomain according to the procedure set out in the Vineland-II Survey Forms Manual (Sparrow et al. 2005b). In accordance to this procedure, the v-scale score for each subdomain was then manually obtained from the provided tables.

The next step in the procedure is to calculate a standard score for both the Communication and Socialization domains as a whole, however for the Communication domain not all subdomains were administered. To allow an approximate standard score for the Communication domain to be obtained for each participant, the known and validated mean score for the omitted Written subdomain for verbal individuals aged three to sixteen years old was substituted in, as per Table 8.15 of the Vineland-II Survey Forms Manual (Sparrow et al. 2005b). This was deemed an acceptable approach given that the purpose of administering the Vineland-II was comparison across time points rather than standalone diagnosis. From here, standard procedure was followed to calculate a standard score for both the Communication and Socialization domains, then percentile ranks and adaptive levels were obtained from the tables provided in the Vineland-II manual. Analysis of this Vineland-II data is presented here, with a discussion of implications to follow in Section 7.3.

6.4.1 Demographics and Assumption Testing

All caregivers completed the Vineland-II behavioural assessment at pre-test, with most (93.5%) also completing the immediate post-test. In total twenty nine caregivers completed the Vineland-II at both the pre-test and immediate post-test data points, including 93.8% of those in the experimental group (N = 15, M = 8.93 years, SD = 1.83, 95% CI [7.74, 9.83]) and 93.3% of those in the control group (N = 14, M = 9.21 years, SD = 2.15, 95% CI [7.97, 10.45]). Independent samples t-test found no significant difference in age between the groups (p = 0.59, d = 0.08), and a 2-sample test for equality of proportions found no significant difference in ratio, $\chi^2(1, N = 29) = 0.51, p = 0.47$, with the control group having 21.4% female participants and 78.6% male participants, and the experimental group having 33.3% female participants and 66.7% male participants. No outliers were identified in the Vineland-II data.

To enable parametric testing to be conducted, first the assumptions of equal variance and normal distribution of data must be checked. Shapiro-Wilk normality tests failed for five of the eighteen subdomains for the control group and ten of the subdomains for the experimental group, resulting in the assumption of normal distribution of data being rejected and necessitating the use of non-parametric statistical approaches in analysing the domain and subdomain data.

6.4.2 Primary Analysis

To address Objective 3 and "determine if behaviour of targeted social skills changes due to interaction with the Social Tutor" and to address part 2 of Objective 4 and "determine if any changes in knowledge or behaviour are maintained after software use ends", participant Vineland-II scores at the domain and subdomain level were analysed. As discussed above in subsection 6.4.1, the Vineland data failed to meet the assumptions of normality and equal variance required for parametric testing, and as such Wilcoxon rank sum and Wilcoxon signed rank tests were chosen as the most appropriate statistical test to apply here. The longitudinal data related to Objective 4 is presented in Chapter 6.5 below.

Change from pre-test to post-test was calculated for each participant and Wilcoxon rank sum tests used to compare the control and experimental groups, however no domain or subdomain reached significance. Following this, the pre-test and post-test scores were compared using Wilcoxon signed rank tests for both the experimental group and the control group. For the experimental group the Play and Leisure Time subdomain of the Socialization domain reached significance (p = 0.005, r = -0.52) and for the control group the Expressive subdomain of the Communication domain was marginally significant (p = 0.083, r = -0.33), however once Holm-Bonferroni correction for multiple comparisons was applied neither of these subdomains remained significant and no other domain or subdomain reached significance (Holm 1979).

A summary of the mean v-sum and standard scores for the Socialization domain, Communication domain and a combination of the two is provided in Table 24, and a summary of the v-scale scores for the subdomains is provided in Table 25. It can be seen that both the experimental group and control group

performed similarly overall, with both achieving a small positive change from pre-test to post-test in most domains and subdomains.

Table 24: Summary of mean change in v-sum and standard scores for Vineland-II domains and overall

		v-Sum Domain Scores			Domain Standard Scores		
		Communication	**Socialization**	**Combined**	**Communication**	**Socialization**	**Combined**
Experimental	**M (SD)**	0.20 (1.90)	1.87 (3.94)	2.07 (4.86)	0.27 (3.35)	3.33 (7.07)	3.60 (8.53)
	95% CI	-0.85, 1.25	-0.32, 4.05	-0.63, 4.76	-1.59, 2.12	-0.58, 7.25	-1.13, 8.33
Control	**M (SD)**	0.36 (1.15)	1.29 (3.41)	1.64 (3.71)	0.64 (2.10)	2.43 (5.96)	3.07 (6.57)
	95% CI	-0.31, 1.02	-0.68, 3.25	-0.50, 3.79	-0.57, 1.85	-1.01, 5.87	-0.72, 6.86

Note: Larger values indicate better performance.

The experimental group did display a higher improvement in the Socialization domain overall than the control group, however this did not reach significance. For the subdomains, the experimental group outperformed the control group on the Receptive subdomain of the Communication domain and the Play and Leisure Time subdomain of the Socialization domain, however as discussed only the Play and Leisure Time subdomain reached significance, and even then only before correction for multiple comparisons was applied.

Table 25: Summary of mean change in v-scale scores for Vineland-II subdomains

		Communication		Socialization			Maladaptive Behaviours †	
		Receptive	**Expressive**	**Interpersonal Relationships**	**Play and Leisure Time**	**Coping Skills**	**Internalizing**	**Externalizing**
Experimental	**M (SD)**	0.33 (0.90)	-0.13 (1.60)	0.33 (1.35)	1.47 (1.73)	0.07 (1.75)	0.27 (1.87)	0.13 (1.30)
	95% CI	-0.16, 0.83	-1.02, 0.75	-0.41, 1.08	0.51, 2.42	-0.90, 1.04	-0.77, 1.30	-0.59, 0.85
Control	**M (SD)**	-0.07 (1.21)	0.43 (0.85)	0.50 (1.51)	0.29 (1.98)	0.50 (1.51)	-0.43 (0.94)	0.07 (0.62)
	95% CI	-0.77, 0.63	-0.06, 0.92	-0.37, 1.37	-0.86, 1.43	-0.37, 1.37	-0.97, 0.11	-0.28, 0.43

Note: Larger values indicate better performance, except for the Maladaptive Behaviours subdomain which is denoted with †

6.4.3 Exploratory Analysis

Following analysis of the domain and subdomain level data required for primary Objectives 3 and 4, further exploratory analysis of the available Vineland-II data was conducted. As with the exploratory analysis of the content quiz data, the primary purposes of this exploratory analysis were identifying characteristics of individuals who most benefited from the software and topic areas that most improved, along with possible identification of any problems within the research methodology or software and where focus should be given to in future research. As with the content quiz data, the following analyses were exploratory in nature only, and as such no correction for multiple comparisons has been applied.

Analysis by Domain, Subdomain and Item

Following the primary analysis above, for interest the experimental group data was further broken down according to the response level subgroups identified during analysis of the content quiz data in Subsection

6.3.1, however Kruskal-Wallis analysis identified no significant differences between the response groups when applied to the Vineland-II data. Interestingly, when mean scores for each domain and subdomain are calculated for the response subgroups an encouraging trend can be seen where the high response subgroup outperforms the other subgroups on many domains and subdomains, and the average response subgroup likewise outperforms the low response subgroup. A summary of domain data can be seen in Table 26 and subdomain data in Table 27.

Table 26: Summary of Vineland-II domain change from pre-test to post-test by response subgroups

		Communication		Socialization		Combined	
		v-Sum	Standard	v-Sum	Standard	v-Sum	Standard
High	**M (SD)**	1.00 (1.41)	1.50 (2.38)	3.25 (5.44)	6.00 (9.93)	4.25 (6.85)	7.50 (12.29)
	95% CI	-1.25, 3.25	-2.29, 5.29	-5.40, 11.90	-9.81, 21.81	-6.65, 15.15	-12.05, 27.05
Average	**M (SD)**	-0.29 (2.06)	-0.57 (3.82)	2.29 (4.03)	4.00 (7.12)	2.00 (4.62)	3.43 (8.04)
	95% CI	-2.19, 1.62	-4.11, 2.96	-1.44, 6.01	-2.58, 10.58	-2.27, 6.27	-4.01, 10.86
Low	**M (SD)**	0.25 (2.22)	0.50 (3.70)	-0.25 (1.26)	-0.50 (1.73)	0.00 (2.94)	0.00 (4.69)
	95% CI	-3.28, 3.78	-5.38, 6.38	-2.25, 1.75	-3.26, 2.26	-4.68, 4.68	-7.46, 7.46

Note: Larger values indicate better performance.

Particularly notable is that the high response subgroup display larger improvements on Socialization domain measures and combined Socialization and Communication domain measures than average responders (M = 6.00 and M = 7.50 compared with M = 4.00 and M = 3.43 respectively for standard scores), who themselves make greater gains in these same areas than the lowest responding subgroup (M = -0.50 and M = 0.00 respectively).

Table 27: Summary of Vineland-II subdomain change from pre-test to post-test by response subgroup

		Communication		Socialization			Maladaptive Behaviours †	
		Receptive	Expressive	Interpersonal Relationships	Play and Leisure Time	Coping Skills	Internalizing	Externalizing
High	**M (SD)**	0.25 (0.50)	0.75 (1.71)	0.50 (2.08)	2.25 (2.63)	0.50 (1.29)	0.25 (2.63)	0.25 (1.71)
	95% CI	-0.55, 1.05	-1.97, 3.47	-2.81, 3.81	-1.93, 6.43	-1.55, 2.55	-3.93, 4.43	-2.47, 2.97
Average	**M (SD)**	0.29 (0.95)	-0.57 (1.81)	0.57 (1.13)	1.57 (1.51)	0.14 (2.27)	0.29 (1.70)	-0.14 (1.35)
	95% CI	-0.59, 1.17	-2.25, 1.10	-0.48, 1.62	0.17, 2.97	-1.95, 2.24	-1.29, 1.86	-1.39, 1.10
Low	**M (SD)**	0.50 (1.29)	-0.25 (0.96)	-0.25 (0.96)	0.50 (0.58)	-0.50 (1.29)	0.25 (1.89)	0.50 (1.00)
	95% CI	-1.55, 2.55	-1.77, 1.27	-1.77, 1.27	-0.42, 1.42	-2.55, 1.55	-2.76, 3.26	-1.09, 2.09

Note: Larger values indicate better performance, except for the Maladaptive Behaviours subdomain which is denoted with †

Prior to commencement of data collection, a set of Vineland-II survey items that were deemed most closely aligned with the content of the Social Tutor were identified. These can be seen in Appendix J. To determine if any of these pre-identified items had changed from pre-test to post-test, Wilcoxon rank sum tests were conducted comparing the change in scores for the experimental group against the change for the control

group. Four of the sixty four identified items were of marginal significance ($0.05 < p < 0.1$), all of which came from the Socialization domain, however no items reached it. For completeness, the remaining items from the Vineland-II behavioural assessment were also individually analysed using Wilcoxon rank sum tests. Four items that reached statistical significance were identified, along with two additional items of marginal significance. A summary of all items that reached significance or were of marginal significance and their descriptions can be seen in Table 28.

Table 28: Vineland-II items of marginal and statistical significance for change from pre-test to post-test

Domain	Sub domain	Item	Wilcoxon rank-sum	Experimental	Control	Description
Communication	Expressive	45	p = 0.058 r = -0.25	M = -0.13 95% CI [-0.42, 0.15] SD = 0.52	M = 0.29 95% CI [-0.07, 0.64] SD = 0.61	Says own telephone number when asked.
		51	p = 0.063 r = -0.24	M = -0.27 95% CI [-0.71, 0.18] SD = 0.80	M = 0.21 95% CI [-0.03, 0.46] SD = 0.43	Says complete home address (that is, street or rural route, apartment number, city and state) with or without zip code, when asked.
Socialization	Coping Skills	**6**	p = 0.054 r = -0.25	M = -0.07 95% CI [-0.40, 0.26] SD = 0.59	M = 0.36 95% CI [0.07, 0.64] SD = 0.50	Ends conversations appropriately (for example, says, "Good bye"; "See you later"; etc)
		14 **	p = 0.005 r = -0.37	M = -0.33 95% CI [-0.60, -0.06] SD = 0.49	M = 0.21 95% CI [-0.03, 0.46] SD = 0.43	Refrains from talking with food in mouth.
	Play and Leisure Time	13 *	p = 0.037 r = -0.27	M = 0.20 95% CI [-0.11, 0.51] SD = 0.56	M = -0.21 95% CI [-0.46, 0.03] SD = 0.43	Plays simple make-believe activities with others (for example, plays dress-up, pretends to be superheroes, etc.)
	Interpersonal Relationships	**20**	p = 0.093 r = -0.22	M = -0.27 95% CI [-0.60, 0.06] SD = 0.59	M = 0.14 95% CI [-0.24, 0.53] SD = 0.66	Has best friend or shows preference for certain friends (of other gender) over others.
		29	p = 0.066 r = -0.24	M = -0.20 95% CI [-0.43, 0.03] SD = 0.41	M = 0.14 95% CI [-0.17, 0.45] SD = 0.53	Meets with friends regularly.
		31	p = 0.089 r = -0.22	M = 0.20 95% CI [-0.17, 0.57] SD = 0.68	M = -0.21 95% CI [-0.55, 0.12] SD = 0.58	Places reasonable demands on friendship (e.g. doesn't expect to be a person's only friend, or have the friend always available, etc)
		37 *	p = 0.016 r = -0.32	M = 0.20 95% CI [-0.03, 0.43] SD = 0.41	M = -0.21 95% CI [-0.46, 0.03] SD = 0.43	Goes on group dates.
Maladaptive Behaviours †	Internalizing	3 *	p = 0.042 r = -0.27	M = 0.00 95% CI [-0.30, 0.30] SD = 0.53	M = -0.43 95% CI [-0.73, -0.13] SD = 0.51	Has eating difficulties (for example, eats too fast or too slowly, hoards food, overeats, refuses to eat, etc)

Note: Item numbers pre-identified as closely aligned with Social Tutor content are highlighted in bold with grey background. Larger values indicate better performance, except the Maladaptive Behaviours subdomain which is denoted with †

*** denotes statistical significance ($p < 0.01$), * denotes statistical significance ($p < 0.05$), unmarked denotes marginal significance ($0.05 < p < 0.1$)*

To investigate participant performance on pre-identified Vineland-II items in comparison to non-selected items, the means for both sets of items were calculated for the control group and the experimental group for each domain. A summary can be seen in Table 29. The Maladaptive Behaviours domain had fewer pre-selected items than non-selected items (N = 6 and N = 15 respectively), and the Communication domain

likewise had fewer selected items than non-selected items (N = 9 and N = 16 respectively). Being the core area of focus for the Social Tutor software, the Socialization domain had more included items overall and there were also more selected items than non-selected items (N = 49 and N = 37 respectively).

Table 29: Comparison of pre-identified and non-selected Vineland-II items

			Communication	Socialization	Maladaptive Behaviours †
Experimental	**Pre-identified items**	**M (SD)**	0.022 (0.62)	0.082 (0.67)	0.089 (0.59)
		95% CI	-0.16, 0.21	-0.05, 0.21	-0.09, 0.27
	Non-selected items	**M (SD)**	0.029 (0.58)	0.013 (0.71)	0.009 (0.69)
		95% CI	-0.08, 0.14	-0.08, 0.11	-0.18, 0.20
	Difference	***M***	*-0.007*	*0.069*	*0.080*
Control	**Pre-identified items**	**M (SD)**	0.048 (0.56)	0.051 (0.60)	0.000 (0.44)
		95% CI	-0.06, 0.16	-0.05, 0.16	-0.08, 0.08
	Non-selected items	**M (SD)**	0.045 (0.54)	-0.010 (0.63)	-0.052 (0.57)
		95% CI	-0.01, 0.10	-0.11, 0.09	-0.14, 0.04
	Difference	***M***	*0.003*	*0.061*	*0.052*

Note: Larger values indicate better performance,
with the exception of the Maladaptive Behaviours subdomain which is denoted with †

Analysis of Adaptive Levels

Adaptive levels for both the Communication and Socialization domains were determined for all participants at both pre-test and post-test according to the scoring system outlined in Table 30.

Table 30: Scoring for adaptive levels of domains in Vineland-II Caregiver Survey

Standard score		Adaptive level
Lower bound	**Upper bound**	
0	24	Profound deficit*
25	39	Severe deficit*
40	54	Moderate deficit*
55	70	Mild deficit*
71	85	Moderately low
86	114	Adequate
115	129	Moderately high
130	160	High

** A score of 70 or below is considered 'Low' with finer grained classification adapted from guidelines provided with Table C.4 of the Vineland-II Survey Manual (Sparrow et al. 2005b)*

For the Communication domain there were two participants in the control group who went up an adaptive level from pre-test to post-test, and two in the experimental group. One participant in the experimental group also went down a level. For the Socialization domain, in the control group two participants went up an adaptive level and one went down, while for the experimental group four participants went up a level and

one went down. Details of the changes for each of these individuals can be seen in Table 31. Change in standard score, and thus change in adaptive level, was not found to be statistically significant when comparison was made between the control and experimental groups, however it is nevertheless interesting to note that four of the fifteen experimental group participants improved their adaptive level in the Socialization domain (26.7%), the domain most closely aligned to the content of the Social Tutor. Given that these are standardised scores and calculated taking age into account, scores would be expected to remain stable over time if no interventions or learning opportunities were provided. However, data regarding other interventions and school based programs that participants may have been exposed to was not collected, and it is reasonable to expect that participants would also be involved in interventions outside of the Social Tutor evaluation. Thus, it is not possible to ascertain the causes for the observed changes in adaptive level.

Table 31: Summary of participant data where a change in adaptive level occurred from pre- to post-test

Domain	Group	Pre-test level	Post-test level	Change in standard score
Communication	Experimental	Low (mild deficit)	Moderately low	+3
		Low (mild deficit)	Moderately low	+2
		Adequate	Moderately low	-8
	Control	Moderately low	Adequate	+4
		Low (mild deficit)	Moderately low	+4
Socialization	Experimental	Moderately low	Adequate	+20
		Low (mild deficit)	Moderately low	+12
		Moderately low	Adequate	+11
		Low (mild deficit)	Moderately low	+7
		Adequate	Moderately low	-7
	Control	Low (mild deficit)	Moderately low	+7
		Low (mild deficit)	Moderately low	+2
		Adequate	Moderately low	-5

Analysis by Demographics

The change from pre-test to post-test for each subdomain and domain in the Vineland-II data was calculated, then both control group and experimental data collated and analysed according to both age and 'age bucket' as described in Subsection 5.2.2, with six to eight year olds in the youngest group, nine and ten year olds in the middle group, and eleven and twelve year olds in the oldest group. A summary of demographic data for Vineland-II responses can be seen in Table 32.

To check the appropriateness of parametric tests for analysing these datasets, Levene's test for Homogeneity of Variance and Shapiro Wilk normality tests were run on the data for the experimental group, control group, and both combined when separated by gender or into age buckets, however again multiple domains and subdomains failed to meet the assumptions, making non-parametric alternatives more suitable for the following analyses.

Table 32: Summary of demographic data for Vineland-II responses

Group	Gender		Age							Total
	Females	Males	Youngest			Middle		Oldest		
			6	7	8	9	10	11	12	
Experimental	5	10	1	3	3	2	2	3	1	15
Control	3	11	2	2	2	0	3	3	2	14
Total	8	21	3	5	5	2	5	6	3	29

Kruskal-Wallis tests were employed to compare change by age bucket for all data combined and for both the experimental and control groups separately. No domain or subdomain reached significance, however the Coping Skills subdomain of the Socialization domain was marginally significant for both the experimental group ($p = 0.096$) and the set of combined data ($p = 0.091$). Post hoc pairwise Wilcoxon rank sum tests were then conducted to further investigate where any difference in the groups may exist, and significance was reached for comparison between the youngest and middle age buckets for both the experimental and the combined data set ($p = 0.033$, $r = -0.45$ and $p = 0.024$, $r = -0.36$ respectively), however once Holm-Bonferroni correction was applied significance was no longer attained.

Wilcoxon rank sum tests were next conducted to investigate the presence of any effects due to participant gender. When data from both groups was combined a significant result was found for the Coping Skills subdomain of the Socialization domain ($p = 0.024$, $r = -0.23$), with males ($N = 21$, $M = 1.67$, $SD = 5.89$, 95% CI [-1.02, 4.35]) outperforming females ($N = 8$, $M = 0.88$, $SD = 9.27$, 95% CI [-6.87, 8.662]). When data is separated by intervention group, a similar finding is present for the experimental group ($p = 0.033$, $r = -0.39$) with males ($N = 10$, $M = 0.20$, $SD = 1.40$, 95% CI [-0.80, 1.20]) again outperforming females ($N = 5$, $M = -0.20$, $SD = 2.49$, 95% CI [-3.29, 2.89]). It should again be noted that in both cases there are significantly fewer females than males which could have a noticeable influence on the outcomes found.

For the experimental group the Play and Leisure Time subdomain of the Socialization domain also reached significance for comparison by gender ($p = 0.045$, $r = -0.39$) with females ($M = 2.40$, $SD = 2.51$, 95% CI [-0.72, 5.52]) outperforming males ($M = 1.00$, $SD = 1.05$, 95% CI [0.25, 1.75]). The Play and Leisure Time subdomain also reached significance for the control group ($p = 0.037$, $r = -0.39$) however in this case males ($M = 0.91$, $SD = 1.30$, 95% CI [0.04, 1.78]) outperformed females ($M = -2.00$, $SD = 2.65$, 95% CI [-8.57, 4.57]). The percentile score for the Socialization domain overall also reached significance for both the experimental group ($p = 0.047$, $r = -0.36$) and the control group ($p = 0.027$, $r = -0.42$), and again the same trends as the Play and Leisure Time subdomain were observed whereby females in the experimental group ($M = 3.80$, $SD = 15.17$, 95% CI [-15.04, 22.64]) outperformed males ($M = -0.10$, $SD = 3.87$, 95% CI [-2.87, 2.67]) but for the control group males (10, $M = 4.18$, $SD = 4.47$, 95% CI [1.18, 7.18]) outperformed females ($M = -5.67$, $SD = 4.93$, 95% CI [-17.92, 6.59]). A visual representation of this can be seen in Figure 21, where it becomes apparent that there is much more variation between the scores of female participants than

male participants and given the small sample size of females in comparison to males as indicated in Table 32, it is clear that caution must be taken when interpreting any results based on participant gender.

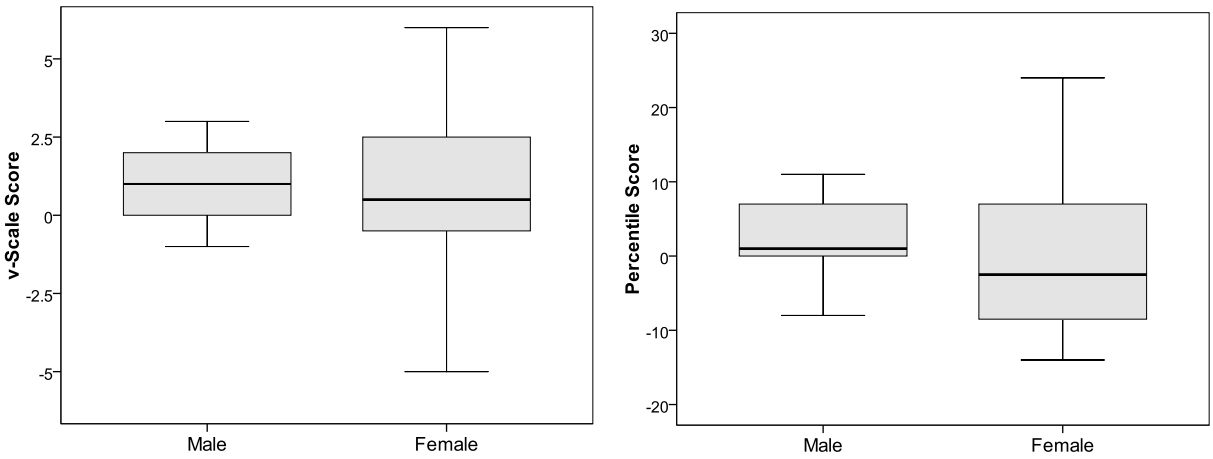


Figure 21: Box plots displaying variation in Vineland-II scores by gender for Play and Leisure subdomain v-Scale scores (left) and Socialization domain percentile scores (right) for the experimental and control groups combined

Additionally for the control group the v-sum of the Socialization and Communication domains combined reached significance ($p = 0.049$, $r = -0.37$) with the standard score being marginally significant ($p = 0.050$, $r = -0.37$). In line with existing trends for the control group, males again ($M = 2.73$, $SD = 2.94$, 95% CI [0.76, 4.70]) outperformed females ($M = -2.33$, $SD = 4.04$, 95% CI [-12.37, 7.71]). Once again, the small sample sizes indicated in Table 32 mean that these results must be interpreted with caution, as there were only three females in the control group and from inspection of raw data it is apparent that two of these were the lowest two performers for the Socialization domain over all participants, making this the likely cause of the unanticipated significant results.

6.4.4 Summary

Analysis of Vineland-II data indicates that participants in the experimental and control groups performed similarly across the intervention period, and while some promising trends were uncovered more data is required to determine if they are genuine trends or simply due to chance. The apparent improvement in behaviour some participants displayed is possibly explained by exposure to a range of activities at home and school, a placebo effect in the parent-reported data, or simply due to increasing maturity, rather than being due primarily to use of the Social Tutor.

6.5 Research Objective 4 - Maintenance of Skills

To address Research Objective 4 and determine if any changes that participants made to their social skill knowledge or behaviour during the intervention phase were maintained after software use ended, longitudinal data was collected at both two and four months after the conclusion of the software use period. At both of these data points participants were asked to complete the content quiz to assess their knowledge,

and their caregivers were asked to complete the Vineland-II to assess their behaviours. The analysis of the longitudinal data from both of these measurement tools is presented here.

6.5.1 Task Completion

As discussed in Section 5.4, a number of technical difficulties were encountered between the end of the immediate post-test and the two and four month follow up tests, necessitating some participants completing their content quiz via a Microsoft Word document rather than the software itself. In these cases only correctness data could be obtained, not accuracy or duration data. A summary of task completion for both the control and experimental group at each data collection point can be seen in Table 33. It should also be noted that there were two control group participants who did not complete the immediate post-test and therefore were not included in the earlier discussion of content quiz results, however they did complete the pre-test and one or both of the two and four month follow up tests. Their results are indicated separately in Table 33.

Table 33: Summary of longitudinal content quiz completion by group

	Experimental		Control	
	Count	**Percentage**	**Count**	**Percentage**
2 month - correctness	15	93.8%	10	83.3%
2 month - complete	14	87.5%	9 (1*)	75.0%
4 month - correctness	12	75.0%	11 (2*)	91.7%
4 month - complete	9	56.3%	9	75.0%

** Two participants did not complete immediate post test, but did complete pre-test and one or both follow up tests. They are indicated in brackets as additions to the listed count and not included in the percentage calculation.*

Vineland-II completion at the two and four month follow up data collection points also varied across participants, a summary of which can be seen in Table 34. It should be noted that there was one control group caregiver and one experimental group caregiver who did not complete the Vineland-II at immediate post-test and therefore their results were not included in the earlier discussion, however they both did complete the Vineland-II at the two and four month follow up points and are therefore included here.

Table 34: Summary of longitudinal Vineland-II completion by group

	Experimental		Control	
	Count	**Percentage**	**Count**	**Percentage**
2 month	12 (1*)	80.0%	10 (1*)	71.4%
4 month	14 (1*)	93.3%	12 (1*)	85.7%

** One control group caregiver and one experimental group caregiver did not complete the immediate post test but did complete the pre-test and both follow up tests. They are indicated in brackets as additions to the listed count. As percentage is calculated based on the total number of participants who completed both pre-test and post-test, they are not included in this value.*

6.5.2 Primary Analysis

As mentioned in Section 6.3, paired t-tests were used to analyse participant correctness scores at the whole-quiz level for each time point, with complete results for each comparison after applying correction for

multiple comparisons provided in Appendix G. Again as discussed in Section 6.3, it was found that for the experimental group alone, the change in scores from pre-test to immediate post-test was significant ($p = 0.01$, $d = -0.73$). As can be seen in Table 42 of Appendix G, it was also found that the change from pre-test to the final post-test conducted four months after software use ended was significant ($p = 0.004$, $d = -1.03$). This suggests that participants in the experimental group were able to maintain their change in knowledge after software use ended. Change in correctness scores from pre-test to the second post-test conducted two months after software use ended was marginally significant, but only prior to correction for multiple comparisons being applied ($p = 0.051$, $d = -0.55$).

As in Section 6.4, Wilcoxon signed rank tests were used to compare the Vineland-II longitudinal data for the Communication, Socialization and Maladaptive Behaviours domains across the four test periods, however no significant interactions were found, suggesting participants made no significant change in observable behaviour across the period of the study.

6.5.3 Exploratory Analysis

After primary analysis of the longitudinal content quiz data, it was shown that participants displayed a significant change in knowledge from pre-test to immediate post-test, and from pre-test to final post-test, yet did not display a significant change from pre-test to intermediate post-test. Further exploratory analysis was conducted to investigate if any factors could be identified to explain this finding. Further exploratory analysis was also conducted on both the content quiz and Vineland-II data to identify if any patterns were present that may inform the direction of future research.

Content Quiz

The change in correctness, accuracy and duration scores between the two and four month follow up tests and the pre-test and immediate post-tests were calculated for both groups, with mean change and a count of the number of participants who performed better than they had at the earlier time points presented in Table 35. Interestingly, when the data from the experimental group and control group are graphed, as can be seen in Figure 22, a trend is observed where there is a dip in scores at the two month post test for both groups. When the experimental group is further broken down into response subgroups it can be seen that this dip is present in the control group, as well as both the high and average responding subgroups.

Following from this observation, it can be seen in Table 35 that for both the control and experimental group the correctness and accuracy scores are better at the four month data collection point than the two month data collection point, and this is also true of duration data for the control group. For the experimental group their question answering speed slows down to a point that is still faster than their pre-test speed, but slower than their immediate post-test speed.

Table 35: Summary of longitudinal data for content quiz by group

			Better than pre-test				Better than post-test				Change from pre-test	Change from post-test
			N	*count*	*% of avail*	*% total*	*N*	*count*	*% of avail*	*% total*	*M (SD)*	*M (SD)*
Experimental	**2 month**	**Correctness**	15	11	73.3	68.8	15	7	46.7	43.8	5.19% (8.08)	-1.61% (7.12)
		Accuracy	14	8	57.1	50.0	14	4	28.6	25.0	5.03% (11.56)	-1.30% (7.23)
		Duration	14	13	92.9	81.3	14	10	71.4	62.5	-33.37s (29.37)	-2.20s (16.39)
	4 month	**Correctness**	12	10	83.3	62.5	12	8	66.7	50.0	10.85% (9.51)	4.07% (7.10)
		Accuracy	9	7	77.8	43.8	9	4	44.4	25.0	10.42% (14.06)	4.49% (9.11)
		Duration	9	7	77.8	43.8	9	6	66.7	37.5	-25.77s (36.00)	3.13s (31.85)
Control	**2 month**	**Correctness**	11	7	63.6	58.3	13	9	69.2	75.0	0.63% (4.95)	-0.71% (6.93)
		Accuracy	10	8	80.0	66.7	9	5	55.6	41.7	3.13% (4.65)	0.79% (3.99)
		Duration	10	7	70.0	58.3	9	6	67.0	50.0	-16.62s (43.92)	-7.73s (57.30)
	4 month	**Correctness**	10	4	40.0	33.3	11	7	63.6	58.3	3.30% (5.44)	2.02% (5.24)
		Accuracy	9	7	77.8	58.3	9	5	55.6	41.7	6.29% (4.62)	3.11% (6.73)
		Duration	9	8	88.9	66.7	9	7	77.8	58.3	-36.19s (35.75)	-26.00s (34.26)

Note: "% of avail" calculated based on number of responses (N) available at that data point for the intervention group specified and varies for each measure and time point, while "% total" is calculated based on total number of participants included in that intervention group overall (N = 16 for experimental, N = 12 for control)

Turning again to Figure 22 the difference of approximately 10% in pre-test correctness scores between the control and experimental group becomes apparent, however as discussed earlier independent samples t-test indicated that this was not statistically significant ($p = 0.226$, $d = 0.33$). When the experimental group are separated into response subgroups it can be seen that it is primarily the results from the high responding subgroup that pull this pre-test score down, with average responders only approximately 5% lower than the control group and low responders approximately equivalent with the control group at pre-test. Interestingly the low and average responding subgroups perform similarly to the control group at most time points with

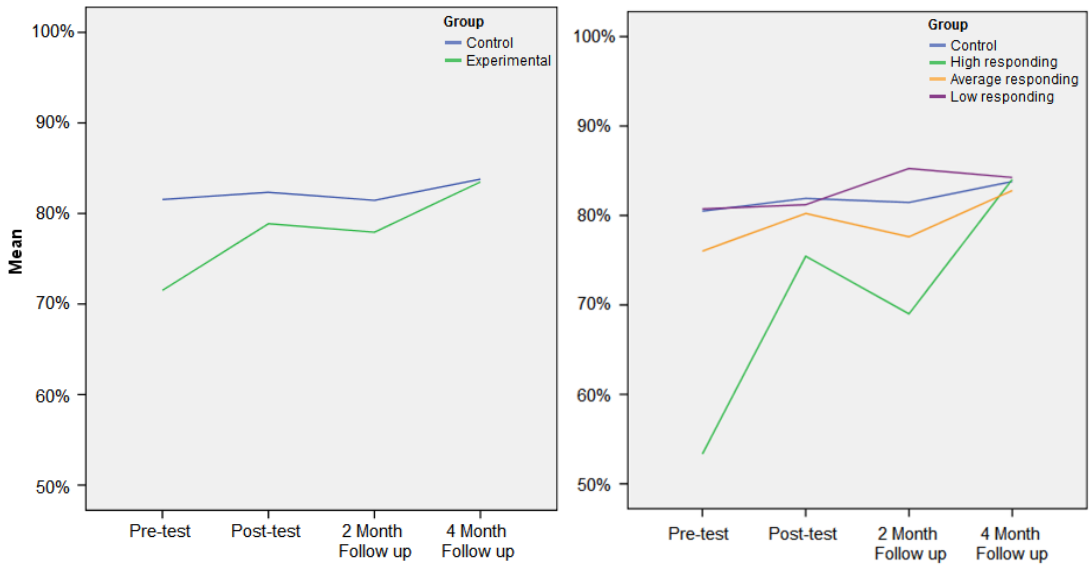


Figure 22: Mean correctness score at each data collection period by group and subgroup

fairly stable scores at each of the data collection points, with the exception of the two month follow up where the low responding group actually demonstrate an increase in correctness score on average.

Behavioural Assessment

The mean change in scores from pre-test and immediate post-test to two and four month follow up tests were calculated for the Communication, Socialization and Maladaptive Behaviours domains for both the experimental and control group, with a summary presented in Table 36.

Table 36: Summary of longitudinal data for Vineland-II by group

			Better than pre-test				Better than post-test				Change from pre-test	Change from post-test
			N	*count*	*% of avail*	*% total*	*N*	*count*	*% of avail*	*% total*	*M (SD)*	*M (SD)*
Experimental	**2 month**	**Communication**	13	5	38.5	33.3	12	2	16.7	13.3	-0.15 (2.12)	-0.67 (1.07)
		Socialization		7	53.8	46.7		5	41.7	33.3	-0.54 (7.04)	-2.58 (7.12)
		Maladaptive Behaviours		6	46.2	40.0		6	50.0	40.0	-0.38 (2.36)	-0.50 (1.83)
	4 month	**Communication**	15	11	73.3	73.3	14	6	42.9	40.0	1.40 (2.85)	-0.14 (3.44)
		Socialization		10	66.7	66.7		5	35.7	33.3	2.87 (6.67)	-0.83 (6.60)
		Maladaptive Behaviours		12	80.0	80.0		10	71.4	66.7	-1.93 (2.40)	-1.50 (2.91)
Control	**2 month**	**Communication**	11	8	72.7	57.1	10	6	60.0	42.9	0.91 (1.92)	0.90 (1.73)
		Socialization		8	72.7	57.1		5	50.0	35.7	2.09 (3.51)	1.80 (3.88)
		Maladaptive Behaviours		4	36.4	28.6		2	20.0	14.3	-0.91 (3.81)	0.80 (1.87)
	4 month	**Communication**	13	7	53.8	50.0	12	7	58.3	50.0	1.08 (1.98)	0.50 (1.78)
		Socialization		10	76.9	71.4		8	66.7	57.1	5.31 (5.23)	3.92 (5.02)
		Maladaptive Behaviours		8	61.5	57.1		7	58.3	50.0	-1.54 (2.93)	-0.50 (1.68)

Note: "% of avail" calculated based on number of responses (N) available at that data point for the intervention group specified and varies for each measure and time point, while "% total" is calculated based on total number of participants included in that intervention group overall (N = 15 for experimental, N = 14 for control)

It can be seen that at the four month follow up the majority of participants in both groups were performing more favourably than they had a pre-test on all three domains, with the experimental group outperforming the control group on all except the Socialization domain at this time point. A trend similar to that observed in the content quiz data can be seen at the two month follow up point, in particular for the Socialization domain of the experimental group, whereby there is a general decline in skills across all three domains. This is particularly clear when the results are graphed, as seen in Figure 23.

From further inspection of the graphs in Figure 23 it can be seen that for all three domains the average scores at each data point are displaying an encouraging trend, with the Communication domain increasing very slightly over time and the Socialization domain increasing more markedly, albeit with the previously mentioned dip at the two month follow up point. For the Maladaptive Behaviours domain a reduction in score reflects a reduction in problematic behaviours, and thus the downward trend observed here is likewise

a positive sign. To further investigate these trends the data was next broken down into the previously indentified response level subgroups, and graphs of this can be seen in Figure 24.

It can be seen that for the Communication and Maladaptive Behaviours domains the general trends are similar for all three response level subgroups and the control group. For the Communication domain average responders outperform the control group at all data points while the control group outperforms both high and low responders, whereas for the Maladaptive Behaviours domain all three experimental subgroups outperform the control group at the four month follow up point, with high responders displaying the largest decrease in problematic behaviours and thus the largest improvement. For the Socialization domain the trends are noticeably less consistent, with high responders displaying a much stronger improvement trend than any other group, which is consistent with the findings from analysis of the content quiz.

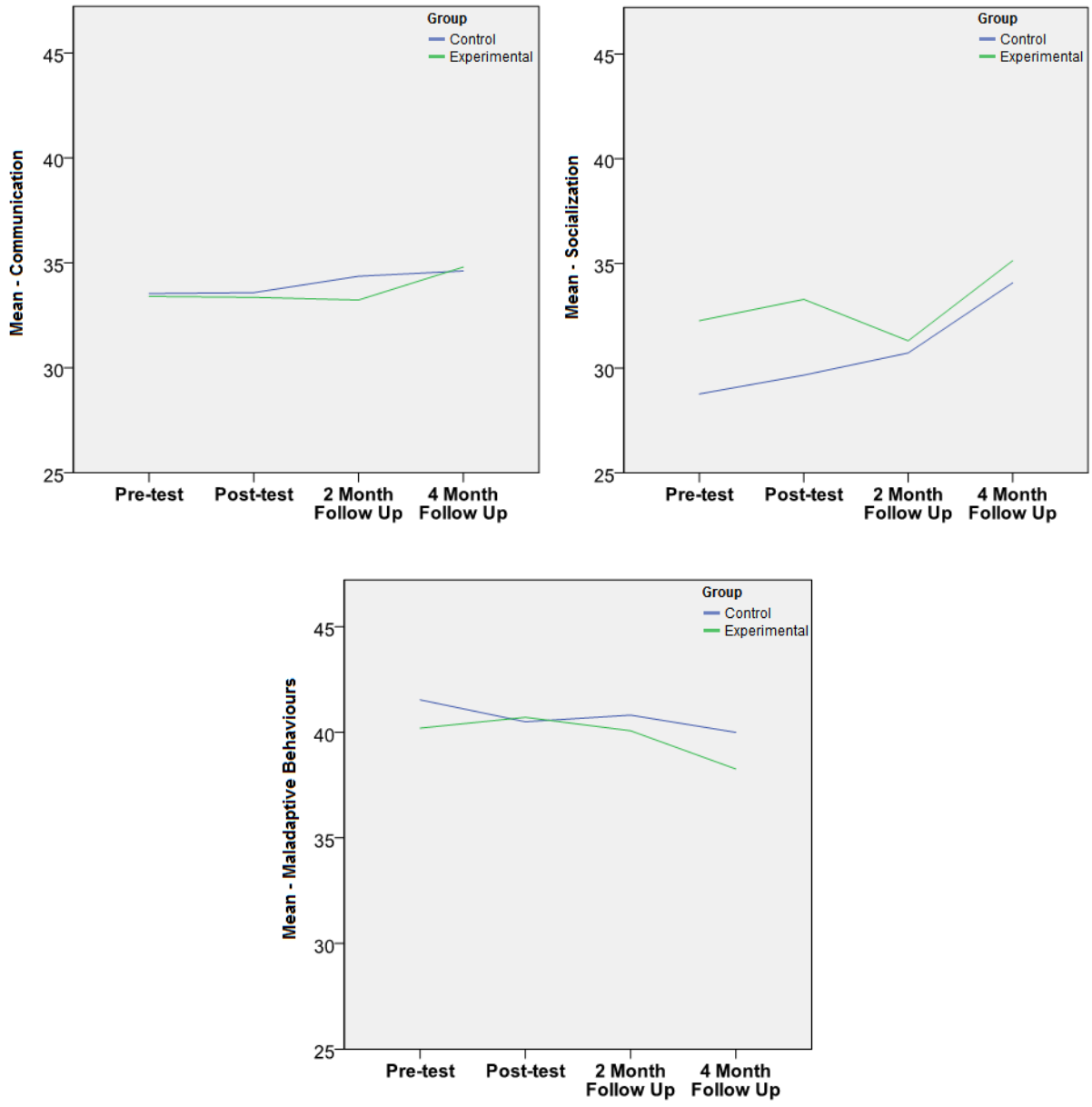


Figure 23: Mean v-sum scores at each data collection point by group for the Communication, Socialization and Maladaptive Behaviours domains of the Vineland-II

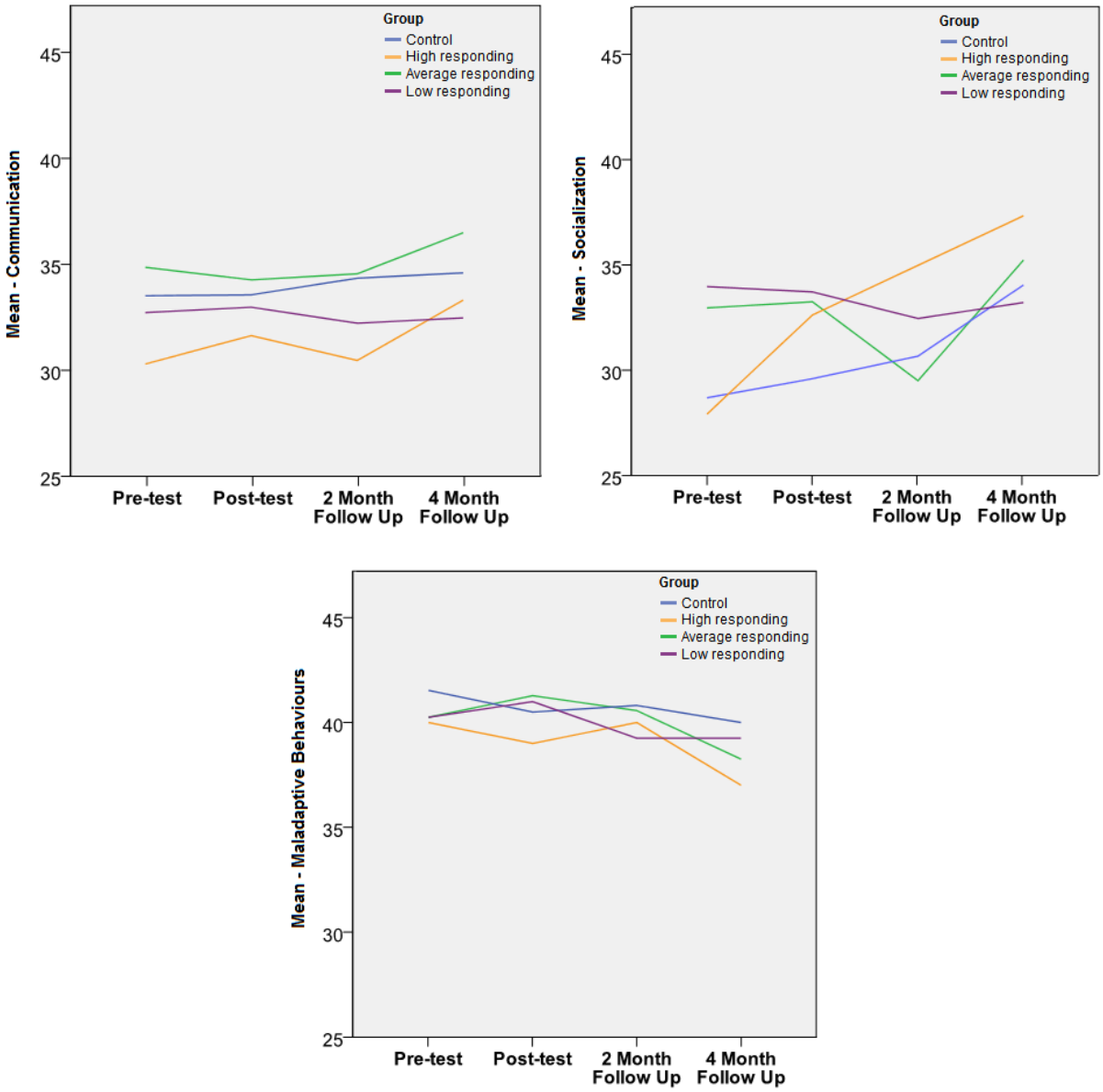


Figure 24: Mean v-sum scores at each data collection point by response subgroup for the Communication, Socialization and Maladaptive Behaviours domains of the Vineland-II

6.5.4 Summary

Longitudinal content quiz results suggest that participants in the experimental group were able to maintain the gains they made during the software use period, with exploratory post hoc analysis suggesting that, for the high responding subgroup in particular, they may even continue to build on their knowledge after software use has ended. Longitudinal Vineland-II outcomes showed that generally participants from both groups performed similarly at all four data collection points, with exploratory analysis again suggesting that the high responding subgroup may be an exception, displaying improvements in behaviour from pre-test to post-test and continuing this trend beyond the end of the software use period.

6.6 Research Objective 5 - Perceptions of Software

To evaluate both participant and caregiver expectations of the Social Tutor prior to use and their perceptions of it following the intervention period, thus addressing Research Objective 5, questionnaires were administered at both pre-test and immediate post-test data collection points. At pre-test participants alone were asked to complete a very brief questionnaire assessing their previous experiences with technology and their expectations of the software, while at post-test both participants and their caregivers were asked to complete a slightly longer questionnaire investigating their perceptions of the software and eliciting feedback for future development of this and similar software. Effort was made to ensure that the Social Tutor was both intuitive and enjoyable to use, and thus it was hypothesised that participants and caregivers would report positive experiences overall, however the surveys were primarily exploratory in nature given that this is the first evaluation of this unique Social Tutor software. The pre-test and post-test questionnaires themselves can be seen in Appendix B and Appendix C respectively, with responses to both presented here.

6.6.1 Pre-test Questionnaire

The pre-test questionnaire was completed by all thirty one included participants and consisted of one page of multiple choice and Likert-style questions, and one page of open-ended questions.

Rating Scale Responses

All participants indicated that they used either desktop computers or tablet devices regularly, with most using them daily (61.3%), and the remainder either using them several times a week (32.3%) or at least once a week (6.5%). A visual representation of participant responses to these Likert-style questions can be seen in Figure 25, with further detail available in Appendix K: Table 48. The assumption of homogeneity of variances was not met thus non-parametric Wilcoxon rank-sum tests were run, however no significant differences were found between the experimental and control groups for any Likert-style question.

Open Ended Responses

Participant responses to the open-ended question "What do you think it will be like using the Talking Head software?" support the responses obtained from the earlier Likert-style questions, with over half of participants indicating positive expectations for the software to be 'fun' or 'good' (51.61%) and many responding that they didn't know what it might be like (45.16%). Some participants felt the software would be educational (16.13%), a few expected it to be hard (6.45%) and only one suggested it would be bad or boring (3.23%). To gain insight into participants' prior experiences they were also asked "What do you normally use computers for?" All participants included 'playing' as one of the purposes, with a wide range of games cited including several mentions of 'Minecraft' and a few explicitly educational games, mostly in the areas of mathematics and literacy. Almost half of all participants indicated that they used computers for watching videos (48.39%), particularly using YouTube and often for the purpose of obtaining information about the games they are playing or other hobbies. A few indicated that they used the computer for browsing the internet and looking up information generally (19.35%). Other purposes included listening to music,

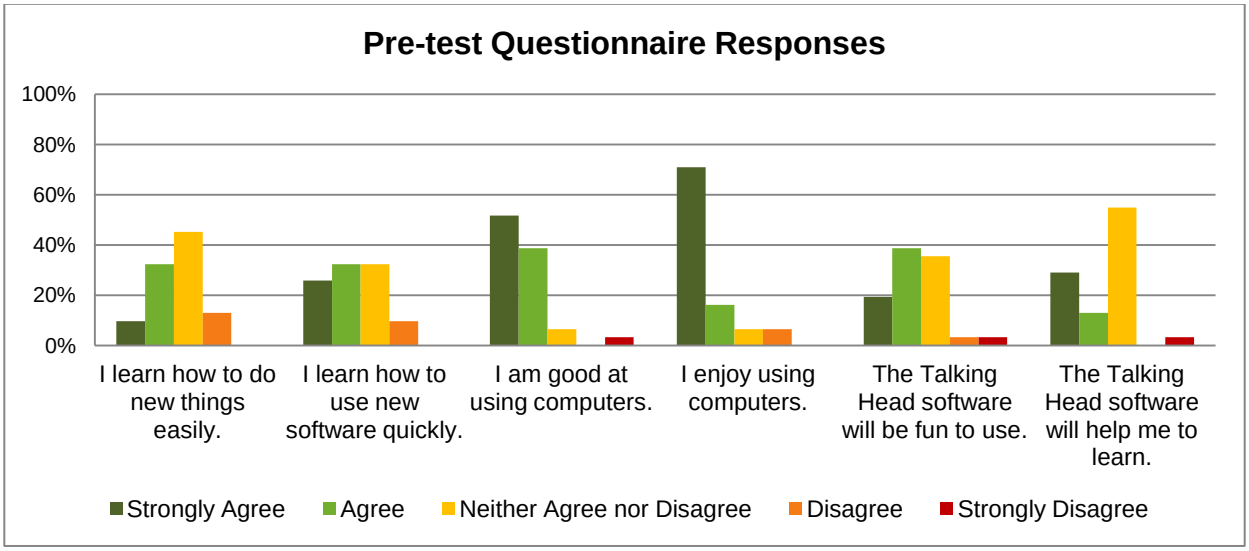


Figure 25: Visual summary of numerical pre-test questionnaire items

writing, drawing, creating videos, science-related tasks and writing their own computer code. Only two participants indicated that they used computers to email friends (6.45%). A summary of participant computer use can be seen in Figure 26.

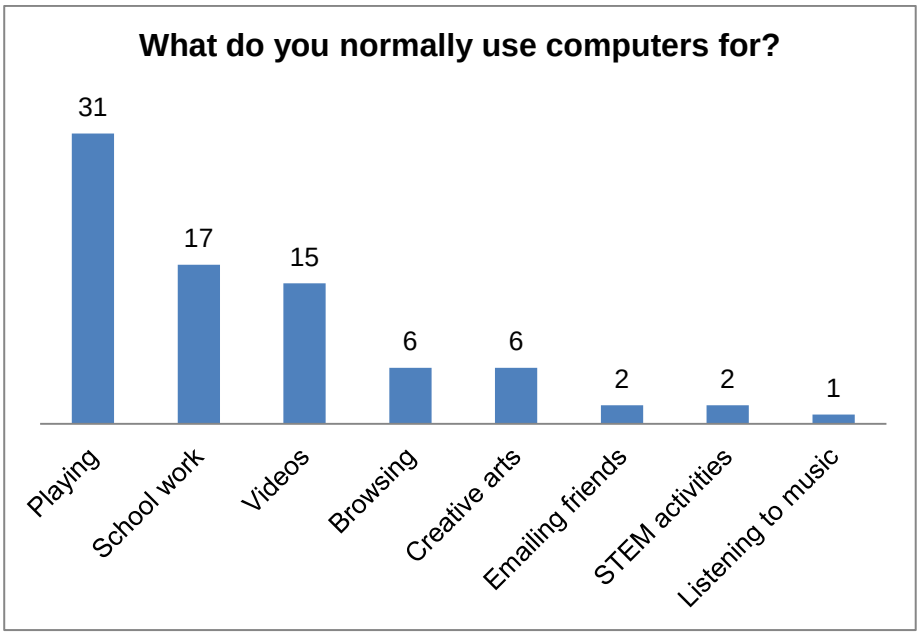


Figure 26: Participant purposes for computer use

6.6.2 Post-test Questionnaire

Immediately following the end of the three weeks of software use, a post-test questionnaire was completed by participants and their caregivers. Identical questions were presented to all respondents. The completion rate for the post-test questionnaire can be seen in Table 37. In both the control group and the experimental group there were two families with siblings participating, resulting in four families where there was one caregiver questionnaire and two participant questionnaires completed.

Table 37: Post-test questionnaire completion rates

Group	Complete	Total possible	Percentage complete
Control Participants	11	15	73%
Control Caregivers	9	13	69%
Experimental Participants	13	16	81%
Experimental Caregivers	13	14	93%

Spontaneous comments from a small number of caregivers indicated that they felt their children were providing answers based on what they believed the researcher would want to hear rather than their true feelings, thus social desirability bias exists in the responses from the participants in particular and caution must be taken when interpreting results.

Rating Scale Responses

The post-test questionnaire contained both Likert-style rating scales and open ended questions. A visual summary of mean responses to user experience, enjoyment and educational value focussed Likert scales can be seen in Figure 27, with further detail available in Appendix K: Table 49. Visual representations of the frequency of responses to these same statements can be seen in Figure 28 for the experimental group and Figure 29 for the control group.

All statements were framed in a positive sense, with higher responses representing more agreement with the given statement. There were few significant differences found between the responses obtained from the control group and the experimental group, with Wilcoxon rank sum tests indicating a significant difference for participant answers to the statements 'it was easy to choose the activity I wanted to do' ($p = 0.012$, $r = -0.36$) and 'the activities were fun' ($p = 0.011$, $r = -0.36$) only. For caregiver responses marginal

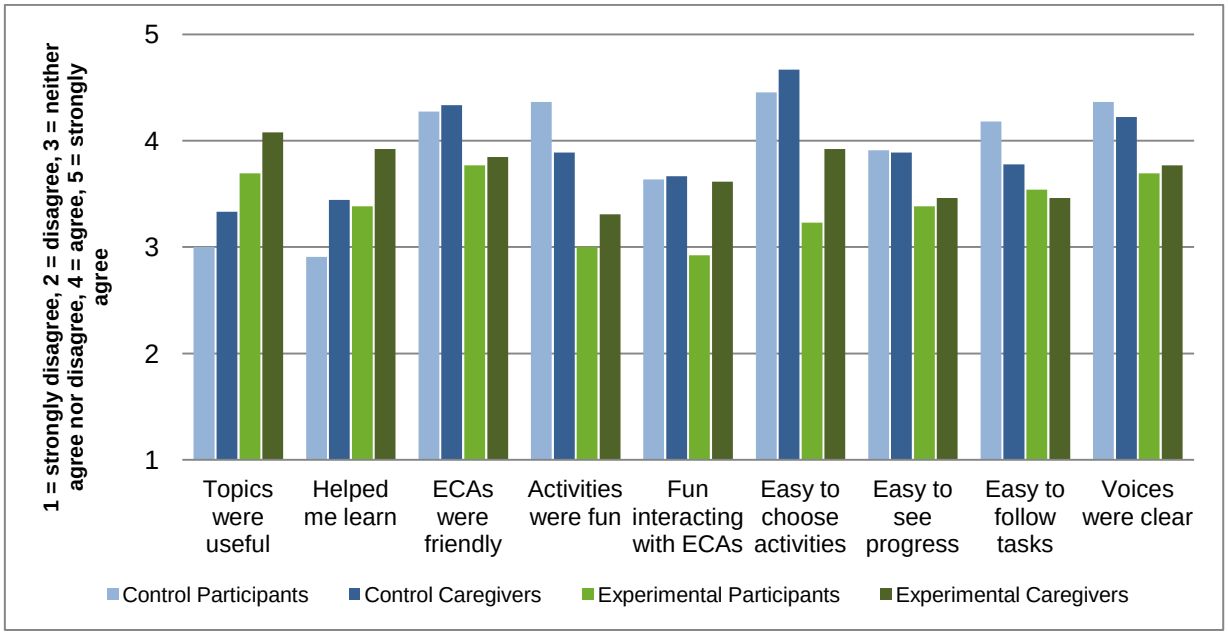


Figure 27: Mean responses Likert-style post-test questionnaire items by group

significance was reached for 'it was easy to choose the activity I wanted to do' ($p = 0.096$, $r = -0.28$) and 'the activities were fun' ($p = 0.096$, $r = -0.27$) but no item reached significance.

On average, control group participants responded approximately as expected, disagreeing slightly with the statement 'the software helped me to learn' ($M = 2.91$, $SD = 1.38$, 95% CI [1.99, 3.83]) and neither agreeing nor disagreeing with the statement 'the topics were useful' ($M = 3.00$, $SD = 1.61$, 95% CI [1.92, 4.08]). Interestingly, control group caregivers on average agreed slightly that 'the topics were useful' ($M = 3.33$, $SD = 1.23$, 95% CI [2.39, 4.27]) and that 'the software helped my child to learn' ($M = 3.44$, $SD = 1.13$, 95% CI [2.58, 4.31]). Neither of these statements reached significance for participants or caregivers separately, however when the data was combined a marginally significant difference between the control and experimental groups was found ($p = 0.064$, $r = -0.19$).

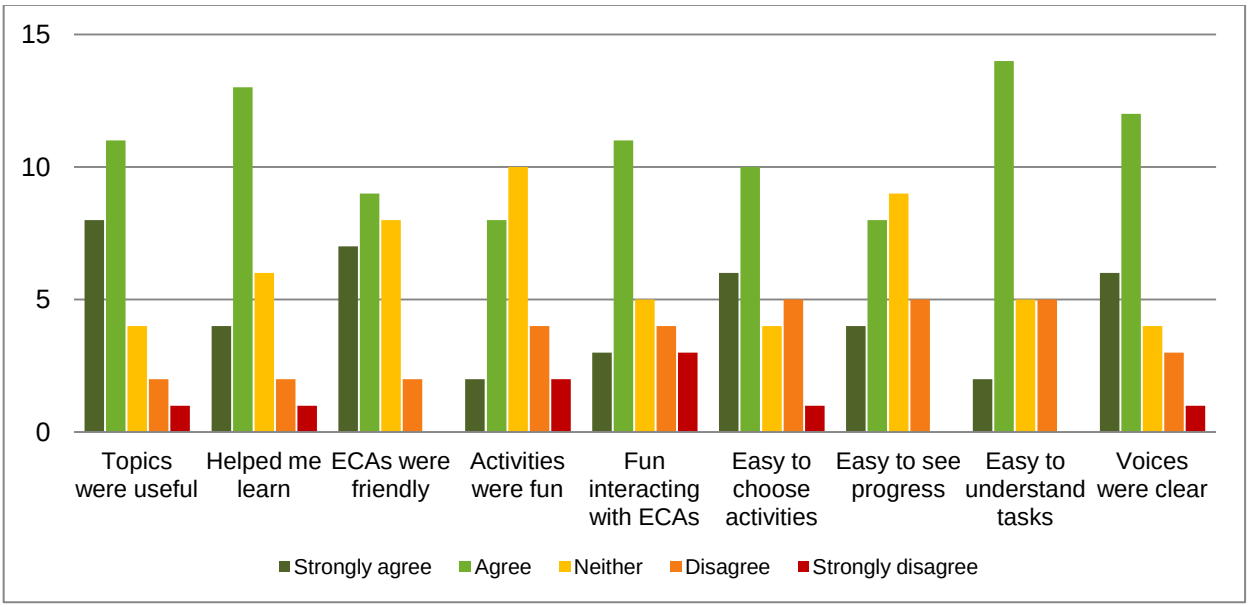


Figure 28: Frequency of experimental group responses to Likert-style post-test questionnaire items

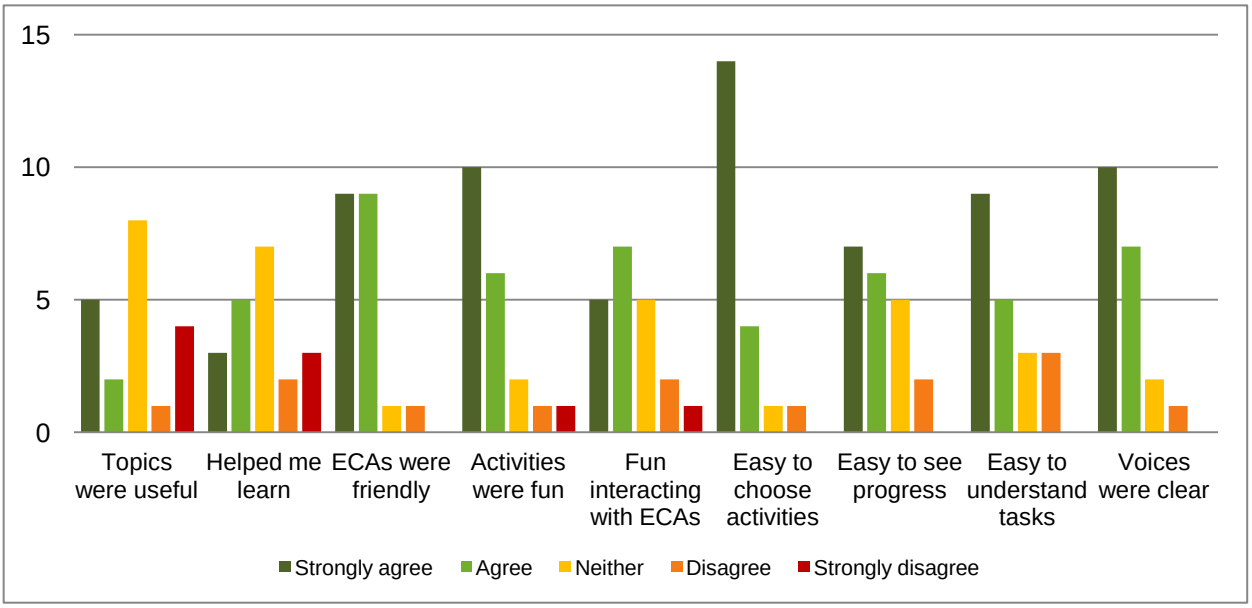


Figure 29: Frequency of control group responses to Likert-style post-test questionnaire items

Enjoyment was another key factor that the post-test questionnaire aimed to address. From Figure 27 it can be seen that control group participants and caregivers rated the software more favourably for every enjoyment related statement than did those in the experimental group, on average agreeing with every statement. In contrast, experimental participants disagreed on average with the statement are 'interacting with the virtual humans was fun' ($M = 2.92$, $SD = 1.26$, 95% CI [2.16, 3.68]) and neither agreed nor disagreed that 'the activities were fun' ($M = 3.00$, $SD = 1.29$, 95% CI [2.22, 3.78]). As mentioned previously, Wilcoxon rank sum test indicated that the difference between responses to 'the activities were fun' reached significance between the control and experimental groups for participants separately ($p = 0.011$, $r = -0.36$) and when participant and caregiver data was combined ($p = 0.002$, $r = -0.33$). Both the control and experimental groups agreed that 'the virtual humans were friendly' with the control group again rating this higher than the experimental group. Marginal significance was reached for this item when caregiver and participant responses were combined ($p = 0.060$, $r = -0.20$).

The next area assessed in the post-test questionnaire was that of usability, covering how easy it was to understand what was required and navigate through the software. Overall responses from participants and caregivers in both groups indicated agreement with the usability focussed statements, however once again caregivers and participants in the control group tended to respond more positively than those in the experimental group. Participants in the control group indicated higher agreement with both 'it was easy to understand what I needed to do in activities' and 'I could clearly hear and understand what the virtual humans said' than those in the experimental condition, and marginal significance was reached for both of these items when participant and caregiver responses were combined ($p = 0.063$, $r = -0.19$ and $p = 0.052$, $r = -0.20$ respectively).

Participants and caregivers in both groups indicated similar levels of agreement regarding the statement 'it was easy to see how well I was going with activities'. For the final usability related statement 'it was easy to choose the activities I wanted to do', Wilcoxon rank sum test indicated a significant difference between the responses from participants in the control and experimental conditions ($p = 0.012$, $r = -0.36$) and when participant and caregiver responses were combined ($p = 0.002$, $r = -0.33$), again with control group participants and caregivers agreeing more strongly with the given statement.

Participants and caregivers were also asked to rate both the time spent on lesson activities and the difficulty of these activities, the results of which can be seen in Table 38. In these cases the rating scale ranged from 'very easy' and 'very short' to 'very hard' and 'very long' with a central rating indicating the difficulty and time spent was 'just right'. There was no significant differences between groups for either statement.

Table 38: Mean ratings of 'time spent on lessons' and 'lesson difficulty'

	Time spent	Difficulty
Experimental Participants	3.08 (0.95)	2.38 (1.12)
Experimental Caregivers	2.77 (0.83)	2.85 (0.90)
Control Participants	2.73 (1.01)	2.00 (1.48)
Control Caregivers	2.44 (0.88)	2.33 (1.00)

1 = very short/easy, 2 = a little short/easy, 3 = just right, 4 = a little long/hard, 5 = very long/hard

As can be seen in Figure 30, median responses indicate that participants and caregivers overall tended to feel that the time spent on lessons was 'just right'. Responses regarding difficulty of lessons were more varied; however, there were no significant differences between groups. As can be seen in Figure 31, participants and caregivers mostly found the lessons to be either 'just right' or tending towards 'too easy'.

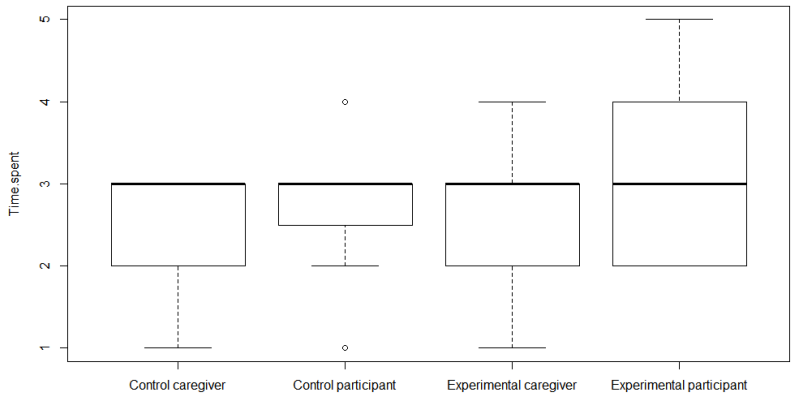


Figure 30: Ratings of 'time spent on lessons' by group

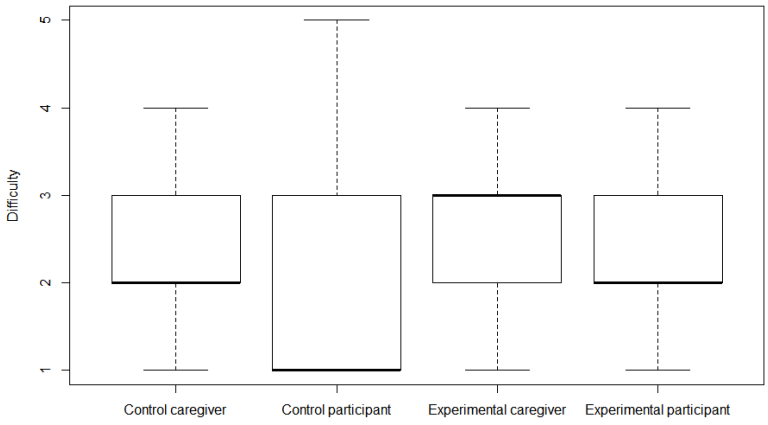


Figure 31: Ratings of 'difficulty of lessons' by group

Open Ended Responses

The post-test survey contained a small number of open-ended questions to give families a chance to provide feedback and elaborate on any of their selections in the previous section. Respondents were given

opportunities to state what they liked best and least, anything they found difficult to do or understand, and what they would change in the software. The primary researcher conducted a descriptive analysis of participant comments, first reviewing them and identifying primary themes. The responses were then divided into the broad categories of strengths, challenges, and suggestions, and further divided into four subcategories for clarity of analysis and discussion. The subcategories are ECA behaviours, lesson content and structure, non-lesson features, and general comments. Finally, the categorisations were inspected by the primary researcher's supervision team. Here the responses are presented descriptively.

Comments on Strengths

A summary of the strength-focussed comments from participants and caregivers can be seen in Table 39. The first category, ECA behaviours, covers positive comments regarding direct interaction with the virtual people. These comments included that users liked the virtual role-plays, they found the characters to be friendly, felt that the characters spoke slowly and clearly, enjoyed having the virtual characters talking and being able to ask them for help, and liked that the virtual characters praised the children using the software and provided suggestions to them regarding pro-social behaviours they could try.

Table 39: Summary of strength-focussed comments from open-ended post-test survey

Positive Comments		**Participants**		**Caregivers**		**Totals**
Subcategory	***Description***	***Experimental***	***Control***	***Experimental***	***Control***	
ECA behaviours	e.g. virtual role-plays, friendly characters, heads spoke clearly	1	2	3	4	**10**
Lesson content and structure	e.g. topics were relevant, repetition, variety, watching videos	4	3	5	2	**14**
Non-lesson features	e.g. getting stickers, rewards, timer, homework	5	2	2	2	**11**
General comments	e.g. it was interactive/fun/ challenging, used it independently	2	7	9	5	**23**
	Totals	**12**	**14**	**19**	**13**	**58**

The second category, lesson content and structure, covers comments relating specifically to the topic and lesson sequence and material. One experimental group caregiver noted that repetition and reiteration of key concepts was a positive, while two experimental group caregivers particularly liked the amount of variety presented to their learners. Two caregivers in the control group commented that the maze themes were particularly enjoyed by their children, and one caregiver in the experimental group noted that the topics presented were very relevant to their child. For the control group specifically, three participants stated that they enjoyed the maze activities, and for the experimental group four participants and a caregiver noted that watching videos within the lessons was a highlight.

The third category, non-lesson features, addresses comments relating to features outside of the core lesson activities. Six participants and one caregiver felt that the sticker and rewards system was a particular highlight, making this the most highly repeated positive comment across all categories. Additionally, one

caregiver from each condition felt that the timer that helped participants manage their software usage by letting them know when 10 and 15 minutes was up was particularly useful, one caregiver particularly appreciated the homework feature, and one participant noted that completing the content quiz was enjoyable.

The final category covers general comments, including two caregivers in each condition who appreciated the interactivity of the software, four participants in the control group who found the software fun to use, along with one who found it easy and another who enjoyed how challenging it was. Two experimental group participants noted that they enjoyed learning new things, while one control group participant said they liked everything about the software. Regarding caregiver comments, one liked how easy it was to navigate the software, three appreciated that it was helping children learn in an engaging and fun way, while two more noted how it was teaching needed skills in a way their children could relate to. Finally, three caregivers particularly appreciated that the software could be used independently by their children.

These comments are all useful when it comes to considering future development of the software, highlighting areas that were particularly appreciated by study participants and caregivers, and thus should be retained or expanded.

Comments on Challenges

A summary of comments relating to aspects of the study that participants and caregivers did not enjoy or would like to see changed can be seen in Table 40. In the first category, ECA behaviours, the most common complaint was that the heads spoke too slowly and that the pauses in their speech were too long, with seven participants and two caregivers noting this. Other issues relating to the virtual characters' speech include that the voices were too robotic and monotone, an issue cited by five caregivers, and that the voice was unclear, noted by one participant and one caregiver.

Another issue relating to the virtual people was their appearance, with six caregivers noting their general appearance, lack of idle motion, or lack of physical expression as problematic. It should be noted that the

Table 40: Summary of challenge-focussed comments from open-ended post-test survey

Negative Comments		**Participants**		**Caregivers**		**Totals**
Subcategory	***Description***	***Experimental***	***Control***	***Experimental***	***Control***	
ECA behaviours	e.g. heads spoke too slowly, were robotic or unclear, appearance of virtual people	8	2	11	3	**24**
Lesson content and structure	e.g. Instructions unclear, activities too hard, content too easy	11	7	8	8	**34**
Non-lesson features	e.g. Homework, couldn't interact while head talking	3	5	1	1	**10**
General comments	e.g. Too repetitive, boring, technical difficulties	5	5	8	3	**21**
Totals		**27**	**19**	**28**	**15**	**89**

virtual people did have idle motion enabled, however in a small number of installations it was necessary to deactivate it to reduce computational load due to the lower-end specifications of the family computer. Finally, one caregiver and one participant in the control group felt that the facility to ask the virtual characters for help was too limited; however, this was not an issue for the experimental group.

For the second category, lesson content and structure, the most common issue was related to activities being too hard, which cumulatively also represents the largest source of negative feedback overall. Six participants and four caregivers in the experimental group noted that mind maps were too difficult or confusing. An additional three participants in the control group felt that some of the mazes were too hard or confusing, which could be referring to the higher level mazes or may be connected to technical difficulties that some participants experienced. An additional seven participants and seven caregivers noted issues around activities being too hard or instructions being confusing. Another issue was that some respondents felt that the content was too easy or that they did not learn anything new, however only one of these participants was in the experimental group, while another one participant and two caregivers who offered this comment were in the control group where educational material was not provided. Finally, three control group caregivers desired more variety of activities; however, this did not impact the experimental group.

The third category covers non-lesson related features. Five participants found it frustrating that they could not interact with the software while a virtual person was talking, and the only other non-lesson feature that received negative feedback was homework, with three participants and two caregivers citing this as an issue, however only one participant and one caregiver were from the experimental group. Given that homework was also cited as a strength by one experimental group caregiver, it appears that opinions on the homework feature are reasonably neutral overall for the experimental group.

The final category of general feedback includes comments relating to engagement and enjoyment, with two participants in each condition reporting that the software was boring and one experimental group caregiver finding that it was not motivating for their child, one control group participant and one experimental group caregiver finding it too repetitive, one caregiver reporting that dragging a mouse was difficult for their child, and another reporting that the two-click method for selecting a lesson or topic was also problematic. Finally, a large source of issues in the general comments category covered that of technical difficulties, with five participants and seven caregivers reporting issues here. Several of these were not reported until the three weeks of software use were over and could not be addressed, while others were addressed mid-study where possible. The technical issues that were reported are easily resolvable with further development, while other issues may require tweaking of the curriculum sequence and lesson activities themselves to resolve.

Suggestions

One of the key purposes of the post-test questionnaire was to elicit suggestions for future development from participants and caregivers once they had experienced a taste of the existing software, and the feasibility of many of these suggestions is discussed in more depth in Chapter 8. Many of these suggestions are directly

related to the difficulties described previously, for example in the ECA behaviours category five participants and three caregivers suggest using clearer, more emotive voices, another four participants and one caregiver would like the virtual characters to speak more quickly, and two participants and four caregivers would like to see the virtual characters improved by either being made more realistic or replacing them with fun, light hearted characters such as animals or cartoons. Other suggestions specific to the virtual people are to allow customisation of the virtual characters or provide more virtual characters to choose between.

Regarding lesson structure and content, two participants and two caregivers suggest including more game-like and playful elements in the software. Two participants and two caregivers from the experimental condition suggest making the instructions clearer, while two participants and one caregiver from the experimental condition suggest providing more control over activity selection, both in terms of providing more options and in terms of skipping activities that are too easy or otherwise undesirable. One participant in the experimental group also suggested including more virtual role-play style activities, while an experimental group caregiver suggested that in addition to the topics their child attempted, lessons on conversation ideas, reciprocating, and interpreting and performing facial expressions and body language would be beneficial.

Non-lesson features and modifications requested by participants include unlocking mini games and rewards more easily or earlier, being able to customise the stickers offered and being able to skip instructions or interact with the software while a virtual person is speaking. Some adjustments to the content quizzes were also recommended, as some questions did not provide enough flexibility to capture all the answers that users wanted to input. A more in depth discussion of issues surrounding the measurement tools can be found in Chapter 7. In addition to these suggestions, one participant requested the ability to take a break, and an experimental group caregiver suggested providing feedback on how well learners performed after each task. In the final category of general comments, three participants wanted a longer trial with the software and one wanted a shorter trial, while three caregivers in the experimental condition recommended improving the graphics used and the visual aspects of the user interface to make them more attractive and engaging.

Table 41: Summary of suggestions from post-test survey responses

Suggestions		**Participants**		**Caregivers**		**Totals**
Subcategory	***Description***	***Experimental***	***Control***	***Experimental***	***Control***	
ECA behaviours	e.g. clearer/more emotive voices, improve character appearance, more customization	11	3	7	2	**23**
Lesson content and structure	e.g. include more game-like elements, clearer instructions, more control over task choice	3	1	3	0	**7**
Non-lesson features	e.g. skip instructions, unlock rewards earlier, more feedback	6	8	6	4	**24**
General comments	e.g. longer/shorter trial time, improved user interface	3	4	1	1	**9**
	Totals	**23**	**16**	**17**	**7**	**63**

Unsolicited Feedback

In addition to the above suggestions, a few interesting comments were obtained from caregivers through spontaneous comments during email or face to face conversations. Two parents commented on their children's disconnect between knowledge and execution of skill in real life, suggesting that more work is required to bridge the gap. Another parent noted that their child did the homework task but then did not see the need to do that behaviour again once the homework was complete. Finally, as was noted previously, a few parents mentioned that their children responded to the survey in the manner they believed was expected of them, for example at least one child stated that they found the software enjoyable despite their parent commenting that this was not actually the case. While participants were actively encouraged by the researcher to respond honestly, this social desirability bias clearly has an impact on results and care must be taken when interpreting responses.

Following completion of the study a small number of parents have taken the time to contact the researcher and comment that they felt their child had benefited from participating, with one parent particularly stating that they have observed an improvement in their child's social behaviours since beginning the program. All participating families were provided with the details necessary to continue using the software following completion of the trial if they wished, and a number of families have expressed their enthusiasm to keep using the software. To this end the researcher has visited a few families in person and supported others via email to enable them to update the software and continue its use.

Overall the results of the post-test questionnaire and spontaneous feedback have yielded much useful information to assist with future improvement of the software which will allow developers to target problem areas and enhance the strengths of the software efficiently. Both formal and spontaneous feedback has provided much encouragement for the future potential of this Social Tutor and related software.

6.6.3 Summary

Outcomes of the pre-test questionnaire indicate that participants showed a similar profile of software use prior to exposure to the Social Tutor, and in general were either positive or undecided about what to expect from it. The post-test questionnaire indicated that the Social Tutor was generally well-received and perceived as a worthwhile educational tool, but with some key issues identified, the resolution of which would greatly enhance the user experience. In particular these centred around the voices and behaviours of the virtual characters, the speed at which participants could move through the Social Tutor content, and the desire for more game-like elements to be included.

CHAPTER 7. DISCUSSION

The current research aimed to develop software that utilises virtual humans to teach social skills to children with autism in a way that is non-judgemental, engaging and can primarily be used independently by learners. At the time of development this was the first known attempt to apply the use of autonomous virtual tutors to the domain of social skills education for this target population. After implementation of the Social Tutor was underway, the ECHOES program by Bernardini et al. (2014) was also developed which uses an autonomous virtual character to teach social skills to children with autism, however ECHOES and the Social Tutor utilise disparate approaches, target different specific skills and the ECHOES program was designed to be optimally used on a large touch screen, making it more suited to a school or clinic environment rather than personal home use. Further, the evaluation of the ECHOES program did not incorporate a control group, use standardised measures, or evaluate generalisation or maintenance effects. Thus, both ECHOES and the Social Tutor in the current research contribute their own unique insights and lessons to this domain.

Given the novel nature of the Social Tutor developed here, directly relating to existing work is challenging. However, the work of Bosseler and Massaro (2003) has similarities in that it focuses on using an autonomous virtual agent, Baldi, for teaching children with autism and the software evaluation is conducted over a comparable duration. In contrast to the Social Tutor, the focus of Baldi is on developing vocabulary, which is easier to evaluate accurately given the quantifiable nature of measuring learned or recognised words, while social skills can be quite complex and context-dependent. Comparisons between the Social Tutor and Baldi are drawn where applicable, with works by other authors also discussed as appropriate.

Following implementation, the Social Tutor software was evaluated to ascertain if participants used it in the intended manner, and whether it was effective in leading to a change in knowledge and behaviour of the targeted social skills. Finally, both caregiver and participant perceptions of the software were determined. From the analysis of results presented in Chapter 6 it appears that the software was indeed used in the intended manner and that its use led to a significant change in knowledge of social skills, although evidence of its use leading to a change in everyday behaviour is limited. Feedback from caregivers and participants indicate that, while there is room for improvement, the software was generally well received and perceived as a valuable learning tool with much future potential. In this chapter the findings associated with each research objective and the implications of such are discussed in more detail.

7.1 Research Objective 1 - Software Implementation

As previously discussed, software was developed to meet the first aim of the current research, specifically to 'design and implement an evidence-based Social Tutor software program that can be used by children with autism'. The Social Tutor automatically collected log data of participant interactions with the system. Participants were asked to use the software three to five times a week for three weeks, resulting in an expectation that they would use the software nine to fifteen times across the study. This duration was

intended to be long enough to collect a meaningful amount of data without being so long as to be an unnecessary burden on families and participants. From the data presented in Section 6.2 and in particular Table 10, it can be seen that on average participants did use the software for the requested number of days, however there were some individuals who used it more or less often than required. This may be due to a combination of factors including families' already busy schedules, participant engagement and willingness, and whether caregivers perceived the software as beneficial for their children and thus encouraged and enforced its use or not. While there was considerable variation between individuals, on average the profile of software use was similar for both groups in terms of the number of days the software was accessed, and these patterns confirmed that the software was utilised appropriately and as intended.

When designing the lesson content for the Social Tutor, inspiration was drawn from the concept of Shute and Towle's (2003) 'Learning Objects' whereby a large number of very short lesson activities were created with the aim of being able to 'mix and match' these to ensure a dynamic and responsive learning environment. Following the recommendations of Shute and Towle (2003), the goal was to design lessons that would each take only a few minutes to do, thus allowing students to complete three to four lessons a day. The data in Section 6.2 Table 11 indicates that these goals were met for both the control and experimental groups, with the experimental group on average completing a larger number of shorter lessons per day, and the control group on average completing a smaller number of longer lessons per day, but both sets of content fitting within the scope of the original study aims.

On average, participants in the experimental group spent less time per day interacting with lesson content than participants in the control group, suggesting that students in the experimental group were more likely to exit once ten minutes was reached, whereas those in the control group were more likely to continue until the software forcibly closed. This behaviour appears to be indicative of engagement levels. While experimental group participants, receiving educational content, typically completed the minimum required of them, it appears several control group participants, receiving game-like content, used the software for the maximum time they could. Data from the post-test questionnaire presented in Subsection 6.6.2 supports this notion, with control group caregivers and participants rating all three 'enjoyment' related Likert-style questions higher than experimental group caregivers and participants did.

Interestingly, in both groups there was one participant who on average continued using the software for several minutes longer than the fifteen minute cut off. Spontaneous discussion with one participant indicated that they enjoyed using the software and would try to quickly choose a new lesson immediately before the fifteen minutes was up so they could keep going for as long as possible. This is a possible explanation for the individual in the experimental group, who also completed the most number of lessons across the experimental period. Another explanation for going over the fifteen minute cut off could simply be that the user took a long time to complete their last lesson for the day or left their last lesson open for a long period, either by choice or due to technical difficulties. The forced close only occurs once the current lesson is exited and the user is returned to the lesson selection screen, to ensure students are not interrupted mid-activity.

From further inspection of the log data, this appears to be a possible explanation for the individual in the control group, who also had the lowest number of recorded days of use.

Topic choice was provided to give users some control over their learning and enable them to select the topic that was most appropriate for them, so it is reassuring that the data presented in Section 6.2, particularly that in Table 12 and Table 13, appears to indicate that participants have made a choice between the topics rather than simply clicking the first item in the list that was offered to them. From closer inspection of the data it can also be observed that several participants completed the overview lesson for a particular objective, which tells them what is covered in that objective, and then chose not to proceed with that objective and instead selected an alternative. This is also reassuring, indicating that these participants are attempting to make informed decisions about what activities they wish to engage with. However, from Table 13 it can also be seen that some participants were quite erratic in their topic choices, jumping backwards and forwards between objectives across the three week period.

While Spearman's rank-order test did not indicate a statistically significant correlation between response level subgroups and number of objective switches, it is interesting to note that three of the four participants identified as 'high response' appeared less erratic and fell into the lower 50% of switches made, while three of the four participants identified as 'low response' appeared to be more erratic and fell into the upper 50% of switches made. This could be indicative of the level of focus or engagement of these individual learners, with possible explanations including that the Social Tutor as an intervention was simply more appropriate for 'high response' individuals, perhaps due to their personal learning styles, interests or current needs, leading them to maintain better focus on the software across the intervention period. In contrast 'low responding' participants appear less engaged and more likely to 'jump around' in an attempt to find an activity that piques their interest or challenges them more. However, an alternative explanation may be that, instead of a lack of focus or interest driving the erratic selection of topics, these participants wanted to experience different parts of the software, but doing so led to a lack of consistency and disrupted the intended scaffolding process that occurs within topics, where lessons are unlocked in a sequential manner to ensure new content reinforces and builds upon previous content. Scaffolding has been repeatedly shown to be a critical element of effective learning in a multitude of contexts, not just within software (Kerr 2002, Herrera et al. 2008, van de Pol et al. 2015, Crook and Sutherland 2017), and so without this scaffolding process occurring as intended learners may not have received the optimal educational experience from the Social Tutor software. Thus, it may be advisable that in future iterations of the Social Tutor participant choice of topic is managed differently, possibly having the software default to the last chosen topic rather than always offering all available topics up front when the learner logs in for the day so that users have to make a conscious choice to move to a different topic.

Related to this, it can be seen in Table 12 of Section 6.2 that the third core topic 'Beginning, Ending and Maintaining Conversations' was not unlocked by any participant due to the fact that no participant completed enough of the prerequisite topic 'Listening and Turn Taking'. This suggests that the intervention period may have been too short and that more time would be needed to fully assess the available content of the Social

Tutor as it currently stands. Determining an appropriate intervention period for a software evaluation of this type is challenging given the relative novelty of the domain and the variety of approaches in the existing literature. For example Bosseler and Massaro (2003) had participants use their virtual tutor Baldi for 10 minutes a day until they reached 100% success at the language tasks they were presented with. On average this took participants 3 to 5 days for each of the three sets of tasks, resulting in approximately the same amount of active software use as participants in the current research. However, the tasks presented by the Social Tutor are more complex than those presented by Baldi, so it is logical that a longer intervention period would be required to reach content mastery in the current research. The evaluation of Andy and the ECHOES program (Bernardini et al. 2014) likewise deals with social content, and the evaluation ran for a six week period. Rather than setting a specific time limit for the evaluation, following the approach of Bosseler and Massaro (2003) and having participants use the Social Tutor until a predetermined performance criterion is met may provide more insight into the length of time needed for participants to benefit from the software.

The failure to unlock Topic 3 also suggests that the automated assessment and dynamic lesson sequencing algorithm may benefit from adjustment so that more capable students are able to skip basic content and move through the lessons more quickly. As previously mentioned, it was a deliberate design choice to include more content than participants were expected to complete in three weeks, both to ensure that participants would have a choice of activities at all times, and to ensure those who worked quickly would not run out of activities to access. However, given that no participants accessed Topic 3, a longer evaluation period and some adjustments to the automated assessment and dynamic lesson sequencing system are recommended. While only one experimental group participant reported in the post-test questionnaire that they felt the content was too easy, it may be that those participants in the 'low response' group who displayed little change in content quiz score from pre-test to post-test would have benefitted from the more advanced content of the 'conversations' topic. By adjusting the lesson sequencing system to allow them to skip ahead more quickly these students may have had a better educational experience and achieved improved outcomes during the intervention period.

From analysis of the log data it is clear that the first research objective has been successfully met, with software developed that can be used successfully by study participants. Overall the software was used in the manner intended, with the number of days used and time spent per session aligning with the study goals.

7.2 Research Objective 2 - Changes in Knowledge

To address research objective two which was to 'determine if knowledge of targeted social skills changes due to interaction with the Social Tutor' a software evaluation was undertaken as described in Chapter 5. As part of this evaluation participants completed a content quiz before and immediately after using the software for three weeks. This interactive quiz was delivered using the Social Tutor software and directly addressed the content presented to the experimental group, allowing participants to demonstrate their knowledge of expected behaviours in a range of scenarios and the steps involved in performing particular target skills. Outcomes from primary analysis of content quiz correctness data indicate that use of the Social Tutor has led

to an improvement in social skill knowledge for participants in the experimental group. Exploratory analysis of accuracy and duration data further supports this, and post-hoc categorisation of participants suggests that there is a small subgroup of participants who respond particularly well to the software and have some possible shared characteristics that may assist to inform future research and development.

7.2.1 Content Quiz as an Assessment Tool

The content quiz was designed to directly assess student knowledge of the content presented via the Social Tutor software, and as such the quiz consisted of twelve mini activities that very closely reflected the style of activities within the tutor, with four questions specifically addressing each of the three topics. In contrast to activities presented during the intervention period, no feedback was provided during content quiz activities. Anecdotally from observing participants completing the pre-test content quiz and from parental feedback, it appears that there were a few flaws within the content quiz design. First, while activities were designed to minimise reliance on strong reading or typing skills and therefore consisted largely of 'drag and drop' style activities in various guises and 'yes or no' choices, this appears to have led to some quiz questions being too restrictive in their possible answers. This means it may not have been as sensitive to subtle differences in understanding as it otherwise could have been, and therefore may not provide an optimal representation of participant knowledge. However, the purpose of administering the content quiz was to determine if changes in knowledge occurred across the intervention period, and since the same quiz was delivered at each data collection point, it still met this goal sufficiently. However, given that the same questions were used at all four data collection points, albeit in a randomised order, it is possible that test scores may also be influenced by familiarity with the quiz questions themselves, and thus similar but differently worded questions at each data collection point may be preferable in future. The second identified issue with the quiz was that it was delivered via the Social Tutor software, and therefore if participants did not enjoy the somewhat lengthy content quiz process, it may have negatively impacted their perceptions of the software overall. Delivering the quiz via the Social Tutor was intended to ensure consistency and reduce the burden on caregivers by being automated and accessible, however some additional changes to the presentation of the quiz to make it appear more distinct from the core software may be advisable for future iterations of this research.

7.2.2 Response Subgroups

Initial analysis of the intervention data from the content quiz appeared to indicate the presence of three response level subgroups, with 'high responders' using the software for a below-average amount of time yet achieving large improvements in correctness score from pre-test to post-test, 'low responders' achieving little to no improvement over the intervention period, and 'average responders' falling in between. More in depth analysis of both the content quiz and Vineland-II data provides further support of this idea, and after inspection of the raw data some patterns regarding subgroup attributes become apparent.

In the low responding subgroup there are three males and one female, with participants tending to be slightly younger: three fall into the youngest age bucket and one into the oldest. From inspection of their content quiz

scores it can be seen that they fall around the middle of overall correctness scores for both pre-test and post-test. There does not appear to be a single obvious cause for their lower response to the software captured in the current data, however it may indicate that the content of the Social Tutor was not a good fit for their current needs or learning style, they may simply not have enjoyed or engaged with it, or it may have been too easy or slow-paced. Given that no participants unlocked the third topic containing the most advanced content, if the evaluation had proceeded for longer and these participants had reached that content, more improvement may have been observed. As discussed previously in Section 7.1, erratic selection of topics may also have played a part in their lower performance. Anecdotally, it was observed by the researcher that one of these students found the virtual characters unsettling at first, but then completed a high number of sessions and their caregiver reported that they said goodbye to the virtual characters at the end of the study, while another participant was expecting a more game-like experience and did a relatively low number of sessions. This suggests that engagement varied across the low responding subgroup and, while it may be a contributing factor, it is not the only one.

In comparison, the high responding subgroup consisted of two females and two males, and tended to be slightly older: three participants fell into the middle age bucket and one into the oldest. Observation of content quiz correctness scores shows that three of the high responding participants had the lowest three pre-test scores, while the last performed around the middle of scores overall. In contrast, at immediate post-test one of these participants achieved the second highest score overall while the others still fell into the bottom half of post-test scores but displayed marked improvement. From the log data it can also be observed that high responders used the software a below-average amount as measured by number of lessons completed and total time actively engaging in the educational content of the Social Tutor, and tended to be more consistent with their topic choices, returning to and continuing with their previous topics more often than they selected a new one. These trends may indicate that the Social Tutor was a good fit for their needs at the time of the evaluation, or could indicate that the software may be more suitable for slightly older students. These results do not appear to be related to engagement or enjoyment given the below-average amount of time these students spent using the software, particularly as anecdotally one of these participants expressed that they found the software boring. Another possibility is that these participants may have better generalisation skills to start with and as such gained more from the software with less exposure. This is supported by the Vineland-II longitudinal data where it can be seen that the high responding subgroup are the only group to display an obvious and consistent upwards trend in any Vineland-II domain across all four data collection points, namely the Socialization domain. While these trends provide some interesting initial insight, more research would need to be conducted to determine if any factors can be identified that would help predict which students would be most likely to benefit most from software of this style.

7.2.3 Correctness

The difference in total percentage correctness from pre-test to post-test was calculated, with the control group displaying only a very small mean change that did not reach statistical significance, while the experimental group displayed a larger mean change that was statistically significant. These results indicate

that participants in the experimental group as a whole were able to improve their knowledge of social skills concepts by a small but significant amount through use of the Social Tutor software and demonstrate 'near transfer' of their improved knowledge when completing the content quiz. This is particularly encouraging given the short three-week intervention period, and aligns with the results of existing software-based social skills interventions such as Emotion Trainer (Silver and Oakes 2001) and FaceSay (Hopkins et al. 2011), although it should be noted that both Emotion Trainer and FaceSay did not incorporate virtual humans and addressed a different set of social skills, namely facial expression recognition.

Interestingly, at pre-test the control group achieved mean overall correctness scores approximately 10% higher than the experimental group, however this was found to be statistically insignificant. This difference could be due to the control group having a mean age approximately a year higher than the experimental group, suggesting that the older children have had more educational and life experiences and demonstrate this more advanced knowledge with a higher pre-test score, or it may simply be due to chance.

After breaking the data down by response subgroups it can be seen that participants in the high responding subgroup noticeably outperformed all other groups, and average responders also displayed higher mean change in test scores as compared to low responders and control group participants. This is to be expected given that experimental group participants were grouped based on quiz correctness scores, but it is encouraging that both high and average responding participants, 75% of the experimental cohort overall, displayed higher means than the control group as a whole. As anticipated, participants in the control group did not make any significant changes to their knowledge of social skills concepts.

Notably it can be seen in Figure 18 of Subsection 6.3.2 that the bottom of the lower whisker for the experimental group box plot falls within the inter-quartile range of the control group, whereas the lower whisker for the control group extends down much further still. This indicates that even the poorest performer in the experimental group achieved a higher change in correctness score than the bottom 25% of the control group. This is further supported when the experimental group is separated into response levels, with both high and average responding groups obtaining mean and median scores above that of the control group, and the low responding group obtaining mean and median scores below the control group.

Interestingly four control group participants, around a third of the group, actually performed worse at post-test than they did at pre-test. This is possibly due simply to chance, with these participants making poorer 'guesses' the second time around. Alternatively, it may be due to a mislearning effect, where the lack of feedback and thus lack of positive reinforcement throughout the first content quiz was perceived as negative reinforcement and led these participants to believe their chosen answers were incorrect even when they may not have been. Because of this, at post-test these participants may have chosen different answers and obtained a lower overall score. It is possible that this same overt reinforcement effect contributed to the superior performance of the high responders from the experimental group, as all four participants identified as high responding were also within the bottom 44% of scores for their group at pre-test, and within the bottom 36% of scores at pre-test for all participants as a complete cohort. This effect of no reinforcement

being interpreted as negative reinforcement may have led these 'high responding' participants to select different answers at post-test, however unlike the control group participants, in their case it may have resulted in corrections and therefore a better score. However, given the magnitude of improvement that the high responding participants displayed, it is unlikely to be the only factor in their superior performance.

From observation of Table 15 in Subsection 6.3.1 it can also be seen that for the control group there was a high standard deviation associated with the change in correctness score, but this same trend was not observed for the experimental group. This indicates a high level of variation in change of correctness scores from pre-test to post-test for participants in the control group, while for the experimental group the change in scores is more stable. These results further support the notion that for the experimental group any changes in score from pre-test to post-test are primarily explained by use of the Social Tutor, while for the control group the changes are more likely due to outside factors, such as other intervention and educational programs they may have been participating in, and the negative performance possibly explained by the mislearning effect previously discussed.

As reported in Section 6.2 the third topic in the experimental group content titled 'Beginning, Ending and Maintaining Conversations' was not accessed by any participants, as no participant completed enough of the prerequisite topic 'Listening and Turn Taking' during the experimental period for it to become unlocked and available. In the content quiz, questions nine to twelve were specifically designed to assess the content that students would have been exposed to had they unlocked and interacted with lessons in that topic. Given this, it was anticipated that students would demonstrate greatest gains from pre-test to post-test in questions one to eight, while questions nine to twelve were expected to demonstrate negligible difference from pre-test to post-test. The content quiz data were broken down into these sections, and for completeness questions one to eight were also broken down by topic, with questions one to four aligning with the content taught in the 'Greeting' topic and questions five to eight aligning with the content in the 'Listening and Turn Taking' topic. It should be noted that the standard deviation was quite high for most subsections, indicating a large amount of variation between individual performance across all groups.

The results displayed in Table 16 of Subsection 6.3.3 indicate that the experimental group overall displayed higher mean change in correctness across all topics and combinations of topics (range: 5.63% - 9.67%) than the control group (range: 0.60% - 1.50%) further supporting the notion of the Social Tutor educational content being responsible for the improvement. Experimental group participants as a total cohort, likely influenced by the scores of the high response subgroup in particular, appear to have performed best in Topic 1 when compared to Topics 2 and 3, with this reaching statistical significance for the overall cohort and being of marginal significance for high responders.

In contrast to the expectations previously discussed, for Topic 3 the experimental group as a whole, again strongly influenced by the high responding subgroup scores, achieved mean change in correctness scores notably higher than the control group and on par with those they achieved for Topic 2, although the results did not reach statistical significance. Given that Topic 3 was not accessed at all, and that the control group

did not show this same improvement trend, it appears that the high responding participants may be generalising the knowledge they learned through doing the prerequisite content in Topic 2 and applying it to the questions aimed at evaluating their knowledge of Topic 3. This is logical given that the skills involved in being a good listener and taking turns with others are directly applicable to maintaining good conversations. It is also especially encouraging given that individuals with autism are known to have difficulty with generalisation, and it appears that they have done so in a 'near transfer' situation here.

For average responding experimental group participants a small improvement in scores occurred for Topic 1 and a somewhat larger improvement occurred for Topic 2, but only a negligible improvement on par with that of the control group was observed for Topic 3. This suggests that Topic 1 content may have been too basic for these participants, and further supports the idea that a mechanism to allow learners to move more quickly through content they already understand may have been of benefit to this response subgroup. It also suggests that average responders did not generalise their knowledge from Topic 2 to the Topic 3 questions like the high responders did, and therefore would require more support for this process to occur.

The low responding subgroup displayed negligible change in Topic 1 scores, unexpectedly displayed a mean decrease in scores for Topic 2, and yet achieved a moderate improvement in Topic 3 scores. This was the only subgroup to display a decrease in correctness scores in any topic. Given that low responders typically started with higher pre-test scores indicating better pre-existing knowledge, one possible explanation is that the scaffolding approach taken in the Social Tutor meant that the content presented was too simplistic for these learners and possibly confused them or contrasted with their existing understanding of how to approach or behave in the situations addressed. Another explanation may simply be that the results are due to individual variation or chance, particularly given the small group size ($N = 4$) and that the findings for this group do not align with any other subgroup. However, on closer inspection of the data it appears that the mean scores for this subgroup are quite reflective of the underlying individual scores rather than being the result of outliers, with all low responders displaying a mean decrease in scores for Topic 2 and three of four improving their scores for Topic 1 and Topic 3. A further follow up study with larger sample sizes may be beneficial to determine any underlying causes of such score decreases.

From the earlier investigation of the log data in Section 6.2 it can be seen that participants spent slightly more of their time working on Topic 1 than on Topic 2. Further, there were only two objectives in Topic 1 and these directly build on from one another ('Greeting' and 'Advanced Greeting') whereas in Topic 2 there are three objectives and, while closely related, the third of these objectives titled 'Turn Taking' is somewhat distinct from the first two objectives of 'Basic Listening' and 'Advanced Listening'. This suggests that for individuals beginning the software with less pre-existing knowledge, specifically those in the high response subgroup, the higher change in correctness scores for Topic 1 is quite logical, since more time was spent on this topic, the topic was by nature more coherent and a better implementation of the scaffolding approach, and it addressed more fundamental content. The corresponding lower scores for Topic 2 questions may likewise be a reflection of these interaction and content patterns, as participants spent less time working on

Topic 2 which contained more challenging content and was less cohesive. Given this, it is reasonable to expect that participants would require more time working on Topic 2 to reach the same level of achievement as they did for Topic 1. For those participants beginning the software with more pre-existing knowledge, specifically those in the average responding subgroup, it is likewise logical that little benefit would be derived from the basic content of Topic 1 and a larger improvement would instead be seen in Topic 2. As mentioned several times, a mechanism to allow students beginning the software with more pre-existing knowledge to move more quickly through the basic content or to guide them towards to the more advanced content rather than simply offering the topics as equivalent choices may help learners use their time more efficiently and derive more benefit from using the Social Tutor.

7.2.4 Accuracy

As previously discussed, the method used to measure accuracy is specific to the particular task type and incorporates the number of incorrect moves made prior to task completion. Most learning activities are designed to guide the student towards reaching 100% correctness before they leave the lesson, and accuracy is not applicable for some tasks such as watching a video or observing the virtual people model a social skill, therefore combining correctness and accuracy provides a better picture of the learner's mastery of a particular task. In contrast to typical learning activities, within the content quiz no feedback was given while the quiz activities were being attempted and participants moved onto the next quiz question as soon as they felt they had completed the current one.

Following observation of participants completing the pre-test quiz it became apparent that students often moved the labels in drag and drop activities around while they were deciding on their answers, not necessarily because they had made a 'wrong' choice but simply as part of their thinking process. As the measure of accuracy takes into account how many incorrect moves were made before reaching task completion, it is clearly influenced by this type of behaviour. Low accuracy scores are therefore not necessarily indicative of a lack of knowledge relating to the task at hand, but could indicate low self-confidence with the task. It was hypothesised that as students increase their mastery of a task, they will also increase their confidence, engage in fewer of these 'thinking' behaviours, and as a joint result display improved accuracy scores.

After analysis of the content quiz accuracy data it became clear that in general participants displayed a mean improvement in their accuracy from pre-test to post-test. While this trend held true for the control group, experimental group and all response level subgroups, it reached significance for the experimental group only and a high standard deviation in comparison to the mean was observed for the control group in particular. The magnitude of the improvement was in alignment with trends in the content quiz correctness data, with high responding experimental group participants displaying the most improvement and control group participants displaying the least. This suggests that increased confidence may lead to increased accuracy scores, with confidence potentially influenced by two factors - familiarity with the Social Tutor software itself, and increased knowledge. Familiarity with the software may explain the increased accuracy of the

control group, while this effect in addition to increased knowledge may explain the larger improvements of the experimental group.

7.2.5 Duration

Results indicate that on average participants answered content quiz questions faster at post-test than they did at pre-test regardless of which intervention group or response subgroup they belonged to, although statistical significance was reached only for the experimental group and the total cohort combined, and the magnitude of improvement in speed varied across groups.

The faster answering times at post-test may be due simply to participants' increasing familiarity with the software, as in both the control and experimental conditions participants would have a better understanding of how to input their answers when they are presented with the content quiz a second time. This is likely to be the main factor in the faster answering speed of the control group, who showed a smaller mean improvement than the experimental group and a noticeably larger standard deviation. The larger, more consistent and statistically significant improvement exhibited by the experimental group may be due in part to this familiarity effect, but in combination with participants' increasing knowledge and possibly their confidence in their answers. This is further supported by observation of Figure 20 in Subsection 0.0.0, where smaller quartiles and standard deviations are apparent on the box plot for the experimental group, indicating more consistency between participants at post-test.

When analysed according to response level subgroups, it was found that the high responding subgroup displayed a negligible improvement in answering duration, while all other groups and subgroups displayed a more marked improvement. Similarly, for the high responding subgroup two participants answered more slowly at post-test, which did not occur in any other response level subgroup but did occur for a third of the control group participants (see Table 17 in Subsection 0.0.0). When the data was further broken down by topic, it was observed that high responding participants slowed their answering speed in both Topic 1 and Topic 2, but displayed a very slight improvement in Topic 3.

Given the small sample sizes in each response subgroup these observations may simply be due to chance, however one alternative explanation is that the high responding participants were answering more carefully at post-test than pre-test and therefore taking their time. This could be for a number of reasons, for example these participants may simply have not answered very carefully at pre-test, whereas at post-test they may have felt more invested in the process. Alternatively, their increased knowledge from using the Social Tutor may mean they felt able to answer more questions correctly at post-test than they did at pre-test when they may have skipped or quickly guessed answers to questions they felt unable to answer properly, or their increased knowledge may have made them more aware of nuances within the questions, leading them to pay more careful attention to the options they were given at post-test.

From further inspection of the topic-level duration data presented in Table 18 of Subsection 0.0.0 it was observed that no topic reached significance for the high or low responding subgroups, likely due to their

small sample sizes (N = 4 in both cases), however the average responding subgroup reached statistical significance for change in answering speed from pre-test to post-test for all topics. Regarding magnitude of change, the low responding subgroup outperformed the average responding subgroup markedly on Topic 2, achieving the largest improvement in answering speed for any subgroup in any topic. Low responders also outperformed average responders on Topic 1, but the reverse was true for Topic 3, with average responders outperforming low responders. Interestingly, observation of individual data shows that two of the low responding participants improved their speed by approximately 70 to 80 seconds overall, while the other two displayed improvements in fitting with the average responding subgroup. A possible explanation for the improvement in answering speed by the low responding subgroup is that they remembered the answers they chose at pre-test and simply selected the same options again at post-test. Another possibility is that, while they did not show a marked improvement in their correctness scores, they did confirm and consolidate their existing knowledge and were able to answer the questions with less hesitation at post-test.

Finally, the duration data was broken down by individual question to identify if any specific features of the questions themselves could have contributed to changes in answering speed duration. For the experimental group it was found that all questions designated as being of 'high complexity', that is having a high number of variables for participants to respond to, reached statistical significance, along with three low complexity questions. For the control group two high complexity and two low complexity questions reached significance. This lends support to the notion that for the control group the key factor influencing reduction of speed when answering content quiz questions from pre-test to post-test is simply familiarity with the interface, while for the experimental group it appears that in addition to interface familiarity, confidence with the content material may also play a part, particularly in the later topics that cover more complex and nuanced material.

7.2.6 Demographics

During the group allocation process participants were matched based on age but were not explicitly matched on gender. Analysis confirmed that a reasonable balance for both age and gender was however obtained, as described in Section 6.1. Analysis of content quiz scores based on this demographic information found no effect of gender on content quiz correctness, accuracy, or duration scores for the experimental group, however a significant effect was uncovered for accuracy scores in the control group. Given that there are only three females in the control group (25% of the group overall) it is difficult to attribute this result to anything more than chance, however it is interesting to note that the three females ranked first, third and fourth in terms of control group accuracy improvement, with a noticeable jump between fourth and fifth place. A larger dataset would be needed to determine if this is a legitimate trend, however in the experimental group the female participants' performance was more varied, with three ranking in the lower half of scores for accuracy improvement, one being the poorest performer, and two in the upper half, supporting the notion that the trend in the control group is simply due to individual variation.

When the content quiz correctness, accuracy, and duration data was analysed by age bucket, no significant differences were found for the control group. However, for the experimental group significance was reached for both change in correctness and change in duration when comparing the youngest and middle age buckets, and for change in accuracy when comparing the youngest and oldest age buckets. Interestingly, it was found that the middle age bucket displayed most improvement in correctness scores, the youngest age bucket showed the most improvement in answering speed, and the oldest age bucket showed the most improvement in accuracy. The middle age bucket displaying the most improvement in correctness score is unsurprising given that three of the four participants in the high responding subgroup are members of this group. Given that the main utility of the accuracy and duration scores is as a reflection of individual confidence, and that the correctness score is the more accurate representation of efficacy of the Social Tutor, these findings suggest that participants in the middle age group likely benefited most from the software overall in terms of improved knowledge of the theoretical steps involved in performing the target social skills of greeting, listening and turn taking, and having conversations. However, whether this effect is truly a product of age or is simply due to chance is difficult to determine given the small number of participants in each age bucket.

7.3 Research Objective 3 - Changes in Behaviour

As part of the software evaluation described in Chapter 5 caregivers were asked to complete an assessment of their child's current social behaviour before and immediately after using the software for three weeks. This was conducted in order to address Research Objective 3, namely to 'determine if behaviour of targeted social skills changes due to interaction with the Social Tutor'. As previously described, the behavioural assessment tool used for this purpose was the Vineland-II, with caregivers specifically asked to complete the Socialization and Maladaptive Behaviours domains and the Receptive and Expressive subdomains of the Communication subdomain. Analysis of the resulting data uncovered some promising trends, however given the similarity of performance between the experimental and control group the apparent improvements are likely to be heavily influenced by chance, a placebo effect in the parent-reported data, increasing maturity across the intervention period, and exposure to a range of interventions at home and school, rather than primarily due to use of the Social Tutor software.

7.3.1 Vineland-II as an Assessment Tool

While the Vineland-II has been used to measure change in social skills of children with autism following human-led interventions on numerous occasions (Laugeson et al. 2012, Koning et al. 2013, Ng et al. 2016), it has not been used to measure the impact of software similar to the Social Tutor, making comparisons with existing work difficult. Adding further challenge, in the realm of using virtual tutors for children with autism, little existing research has progressed beyond initial pilot studies, as promising as these appear to be, and often the authors use bespoke measurement tools that are difficult to directly compare. Thus, in the current study a combination of approaches was implemented with the content quiz and Vineland-II used in tandem.

As discussed in Subsection 2.8.2, the Vineland-II was chosen because it was thought to be fine-grained enough to detect subtle changes in behaviour, but also for many pragmatic reasons including the ability to administer only the relevant sections, it not being too time-consuming for caregivers to complete, and it being appropriate for administration by the research team given their diverse skill sets. However, after analysis of the data one possible explanation for the lack of difference in performance found between the experimental and control groups could be that the Vineland-II was not sensitive enough to detect behavioural changes in the specific areas targeted by the Social Tutor software. Following this it was found that other recent work using the Vineland-II to measure social skills changes in children with autism following human-led interventions also encountered this issue, with Vineland-II results failing to reach significance even when multiple other measures did (Laugeson et al. 2012, Koning et al. 2013, Ng et al. 2016).

It should be noted that the Vineland-3 is now available and updates include an improved scoring system, consolidation and reorganisation of some items, shorter Domain-level forms in addition to the traditional comprehensive forms, and the ability to send the official form electronically to parents and teachers for completion (Sparrow et al. 2016). It may be that these changes address some of the limitations of the Vineland-II discussed here, however more research utilising the Vineland-3 to measure change in social skills of children with autism is needed before this can be determined.

Given the outcomes from the Vineland-II in the current study and the results from human-led intervention studies with comparable participant cohorts, it appears that a measurement tool focussing more on conversation skills specifically, such as ability to initiate, maintain and terminate conversations, may be more suitable for detecting generalisation from the Social Tutor to real-world interactions. Finding a tool that addresses these areas more closely but also meets the pragmatic considerations discussed above may prove challenging, and may even necessitate the development, validation and dissemination of a new measure suited to this domain of research and its unique goals, requirements, and features.

7.3.2 Outcomes by Domain and Subdomain

Change from pre-test to post-test for Vineland-II data was calculated and analysed for both the control and experimental groups, however no domain or subdomain reached statistical significance. When pre-test scores were compared to post-test scores, significance was reached for the Play and Leisure Time subdomain of the Socialization domain for the experimental group, however when further analysis was conducted it was found that the single item within this subdomain that reached significance did not align directly with the content taught by the Social Tutor, namely 'plays simple make-believe activities with others'. Thus this result is more likely due to other activities at home or school and not directly related to use of the Social Tutor, or could alternatively be more reflective of the fall in performance of the control group given that their mean change for this same item indicated poorer performance at post-test than pre-test.

Mean change in scores as seen in Table 24 and Table 25 of Subsection 6.4.2 show that, while a small improvement occurred over the intervention period for both the Socialization and Communication domains

and most subdomains, the control group and experimental group performed similarly, further suggesting that any improvements are likely heavily influenced by increasing maturity, exposure to a range of interventions at school and home, or a placebo effect in parent observations, than they are likely to be due to the use of the Social Tutor alone. Given that the intervention period was only three weeks long and that behaviour changes can be difficult and time consuming to enact, often including the necessity of breaking entrenched patterns, it is still promising to see that the experimental group showed a slightly higher improvement in the Socialization domain than the control group.

While no domain or subdomain reached significance when the data was broken down into response subgroups, from Table 26 of Subsection 6.4.2 a positive trend can be observed whereby the high responding subgroup outperforms the other subgroups on several domains and subdomains, and the average responding group likewise outperforms the low responding group. This is encouraging as it aligns with the content quiz results previously discussed, and may possibly indicate that with more time these improvements in knowledge could also manifest as improvements in target behaviours. This positive, yet not statistically significant, trend may indicate that the Social Tutor is a good candidate for inclusion in a hybrid approach to teaching social skills, such as that taken by Beaumont and Sofronoff (2008) in the Junior Detective program or Whalen et al. (2010) in TeachTown: Basics. In these examples the software is paired with a school or group therapy based role-play component which provides learners with an explicit opportunity to practice the skills they are developing via the software with their real-world peers in a guided manner. Both of these programs demonstrated good outcomes for learners. It was hoped that inclusion of virtual humans in the Social Tutor software would facilitate some of the same benefits, and it may be that with a longer intervention period this would be the case, however the current implementation appears insufficient and inclusion of more natural interaction methods and full-bodied virtual characters may help amplify this effect. Further discussion of the possibilities is undertaken in Section 8.6.

7.3.3 Outcomes by Item

As discussed in Subsection 6.4.2 and provided in Appendix J, a number of Vineland-II items were identified as being particularly relevant to the content taught by the Social Tutor. First, mean scores were obtained for both pre-selected items combined and non-selected items combined. When these mean scores were compared little difference was found for the Communication domain for either the control or the experimental group, however for the Socialization domain both groups made a greater mean improvement in the selected items than non-selected items. For the Maladaptive Behaviours domain the reverse trend occurred, with better performance for non-selected than pre-selected items. Again, it is encouraging that a positive trend was found for the Socialization domain, which aligns most closely with the core focus of the Social Tutor, but given the conflicting results of the other domains this must be interpreted cautiously.

Following on from this, analysis of the Vineland-II data was run at the item-level with a small number of items reaching statistical significance or being of marginal significance. These are presented in Table 28 of Subsection 6.4.2. From inspection of these significant items it appears that there is little overlap between

them and the content of the Social Tutor in most cases, for example the items of 'refrains from talking with food in mouth' and 'has eating difficulties' are not addressed by the Social Tutor in any way, however the items 'plays simple make-believe with others' and 'goes on group dates' are at least social in nature. Further, the experimental group outperformed the control group on only three items, these being 'places reasonable demands on friendship', 'plays simple make-believe activities with others' and 'goes on group dates', the latter of which many parents interpreted as group play-dates given the age of the children involved in the current study. While all three are social in nature, the skills associated with these items were not directly taught in any of the curriculum content that participants unlocked and accessed during their time using the Social Tutor software. Thus, while the fundamental skills that the Social Tutor did address may have improved participants' success at these social ventures, it is unlikely that using the software was the primary driving force behind this difference in scores from pre-test to post-test.

7.3.4 Adaptive Level and Demographics

Next analysis was conducted on participants' overall adaptive levels of functioning for both the Communication and Socialization domains. The lowest level of adaptive functioning obtained was 'mild deficit' and the highest was 'adequate' performance. This range indicates that the selection criteria for participants was appropriate and included individuals who were high functioning in the context of autism, with nothing more severe than a 'mild deficit' obtained, but also low enough functioning in these domains for the software to have the potential to benefit them, with nothing above 'adequate' functioning detected. As can be seen in Table 31 of Subsection 0.0.0, more experimental group participants than control group participants increased their adaptive level of functioning in the Socialization domain over the intervention period, with the improvements made by the experimental group typically being larger in magnitude. Again, this positive trend in the target domain of Socialization is encouraging, although it must be interpreted cautiously given the lack of statistical significance.

Finally, analysis of Vineland-II data was conducted according to age and gender, with no statistically significant effects found due to age. In terms of participant gender, some interesting effects were noted, although given the small ratio of female to male participants involved, which is reflective of the typical ratio found in the larger population of children with autism, care must be taken when interpreting any such results. A statistically significant interaction was found for the Coping Skills subdomain of the Socialization domain in the experimental group, with males outperforming females. Statistically significant interactions were also found for the Play and Leisure Time subdomain of the Socialization domain for both the control group and the experimental group, however in the control group males outperformed females while in the experimental group the trend was reversed with females outperforming males. For this subdomain in particular more variation in scores was observed between female participants than between male participants. Finally, for the control group the v-sum for both the Socialization and Communication domains reached statistical significance, with males again outperforming females.

While these results may suggest that the software was more appropriate for male participants, in reality there were only five females in the experimental group and from inspection of the data it is apparent that for the Socialization domain two of these fell into the bottom half of scores overall and three into the top half, with one being the highest performer overall, while in the control group there were only three females with two of these being the lowest two performers for the Socialization domain over all participants. Thus, the results are quite uneven and therefore unlikely to be reflective of genuine effects of gender and more likely to be reflective of differences in the individuals recruited for the current study, however a follow up study with larger sample sizes would be informative in investigating gender differences further.

7.4 Research Objective 4 - Maintenance of Skills

As previously discussed, collection of longitudinal data is still lacking in many intervention evaluations for children with autism (Rao et al. 2008, Neely et al. 2016). Studies that do collect this data typically only collect it once and rarely beyond three months after the intervention ends. In the current study, the content quiz and Vineland-II behavioural assessment previously discussed were also conducted at both two and four months following the end of the three week software use phase, addressing Research Objective 4, namely to "determine if any changes in knowledge or behaviour are maintained after software use ends".

From the content quiz results it was found that the experimental group not only maintained their post-test improvement but continued their upward trend after the software use period had ended, and this was particularly notable for the high responding subgroup. As anticipated, the control group remained stable at all four data collection points. Longitudinal data from the behavioural assessment showed that the control and experimental groups both performed similarly at all four data collection points with a slight improvement over time, however again the high responding subgroup appeared to not only retain but continue building on their improvement in scores beyond the end of the three week software use period. These results are encouraging and suggest that content learned with the Social Tutor is maintained beyond the intervention period, a known challenge when designing interventions for individuals on the autism spectrum.

7.4.1 Content Quiz

From the longitudinal content quiz data presented in Table 35 of Subsection 6.5.1 it can be seen that for the experimental group over 70% of participants who provided data at the time point in question performed better in terms of both correctness and question answering speed than they had at pre-test at both the two and four month follow up points, indicating that a majority of participants maintained at least some of the improvement they gained from using the software. Encouragingly, even when percentages are calculated based on the total number of participants included in the experimental group and missing values are treated as "not better" it is still found that over 60% of the experimental group achieve better correctness scores at two and four months post-test than they did at pre-test.

Similarly, over 60% of control group participants achieved better correctness scores at the two month follow up than they did at pre-test, but this drops to 40% at the four month point. However, for the control group the mean change in correctness from pre-test to the two month follow up point is negligible at only 0.63%, although it increases to 3.30% when comparing the pre-test to the four month post-test. This is in contrast to the experimental group where approximately the same number of participants maintained a correctness score better than their pre-test at both the two and four month follow up, but at the two month point the mean improvement is 5.19% which increases to 10.85% at four month follow up. Both the control group and experimental group show a higher mean change at four month follow up than at the two month follow up, although the experimental group shows a more marked improvement at both time points, supporting the notion that use of the Social Tutor has had a positive impact on their knowledge over time. This positive outcome regarding maintenance of knowledge is in line with the results obtained from use of the autonomous virtual tutor Baldi (Bosseler and Massaro 2003), where participants demonstrated maintenance of their newly developed vocabulary one month after the intervention ended. In the current study participants displayed this maintenance at both two and four months after the intervention ended, although here the targeted knowledge is social skills rather than language skills.

It should be noted that, given the upwards trend in scores even after Social Tutor use has ended, it appears that other factors may also be at play in the current results. A few likely candidates given the time frame are increasing maturity, exposure to social skills interventions beyond the Social Tutor software itself, and ongoing opportunities to practice their new knowledge in real-world situations and consolidate it. This same trend was not observed by Bosseler and Massaro (2003) however no comparison can be made in this instance, as participants used the language tutoring software until 100% success was reached, and as such there was no scope for participants to display an improvement after the intervention period ended.

In addition to correctness scores, for both the control and experimental group content quiz accuracy scores are also better at the four month data collection point than the two month data collection point, again supporting the notion of higher confidence and familiarity with the Social Tutor interface. While this is also true of duration data for the control group, for the experimental group their question answering speed at the four month point slowed down to less than their immediate post-test speed, although it remains above their pre-test answering speed. Given that experimental group participants' correctness at this point is still on average higher than it was at immediate post-test, this reduction in speed may indicate that participants are taking their time answering and applying new knowledge they have gained since use of the Social Tutor software ended, allowing them to answer more questions correctly. Alternatively it may reflect that they need more time to think about their answers or even how to input their answers since at this data collection point is has been several months since they used the software regularly and their familiarity with it may be fading, however given that the control group participants did not display the same slowing down trend, this second explanation is less likely.

Following the observation that a larger percentage of participants achieved higher than pre-test scores at the four month point than the two month point, the dates that data collection occurred on were inspected and it became clear that for many participants the two month follow up fell during the school holidays and in some cases fell very close to Christmas and New Year celebrations, with a number of families having reported that they were travelling on holidays during this period. This was not true for any other data collection point. There were also a small number of students for whom this data collection point fell early in the new school year. As discussed in Section 5.4, events such as these can be very disruptive and challenging for individuals on the autism spectrum due to their need for sameness and predictability, and since this is a period of time where their usual routine is particularly interrupted, it is not surprising that many participants' test performance appears to have been affected somewhat at this data collection point in particular.

When the data was broken down into response subgroups it was found that this 'two month dip' occurred for both high and average responders and the control group, but did not occur for participants in the low responding subgroup. Figure 22 of Subsection 6.5.1 provides a visual representation. This lack of 'dip' possibly suggests that the low responding participants were less impacted by the disruption of the school holidays. Given the small sample size (N = 4) this could simply be due to chance and the particular individuals in this subgroup, or their superior performance at pre-test may also be reflective of a more mature capability for dealing with such disruptions. From this response subgroup breakdown it can also be seen that all subgroups performed better on average at the four month follow up point than at both pre-test and immediate post-test, and the average responding subgroup also displayed a more marked improvement from pre-test to immediate post-test than the low responding subgroup or control group, instead displaying a similar trajectory overall to that of the experimental group as a complete cohort. The high responding subgroup by contrast displayed a marked improvement from pre-test to immediate post-test, and following the two month dip, displayed another sharp increase in scores from two to four month follow up.

The longitudinal data from the content quiz further supports the notion that the high responding subgroup gained most benefit from the software, with these participants not only improving the most between pre-test and immediate post-test, but at four month follow up continuing to increase their correctness scores beyond that achieved at immediate post-test. Since participants were not actively using the software after the immediate post-test, these results suggest that the knowledge they gained from the Social Tutor may have fed into their learning at other interventions and learning opportunities and helped them learn more rapidly, or may have been applied and practiced in real-world scenarios, consolidating and reinforcing the theoretical knowledge they gained with the Social Tutor, which was then reflected in higher content quiz correctness scores at follow up.

7.4.2 Behavioural Assessment

Analysis of the longitudinal data for the Vineland-II indicates that at four month follow up the majority of participants in both groups were displaying improvements across all three domains, namely Communication, Socialization and Maladaptive Behaviours, when compared to their pre-test scores, with the experimental

group slightly outperforming the control group on both the Communication and Maladaptive Behaviours domains. From the visual representation presented in Figure 23 of Subsection 0.0.0 it can be seen that for the Communication and Maladaptive Behaviours domains the trajectory of scores is similar for the control and experimental groups, and while for the Socialization domain the trajectory is somewhat different, ultimately at four month follow up both intervention groups performed at approximately the same level, suggesting that the Social Tutor had little impact on Vineland-II scores overall. This is in contrast to the findings of Bosseler and Massaro (2003) who did observe generalisation to real-world contexts after participants had used their autonomous virtual tutoring software, however it should be noted that improvements in vocabulary are more straightforward both to teach and to measure than the social skills targeted in the Social Tutor software.

Breaking the Vineland-II data down by response subgroups, a visual representation can be seen in Figure 24 of Subsection 0.0.0. From this the same two-month regression that was observed in the content quiz longitudinal data can also be observed in the Vineland-II data. For the high responding subgroup it manifested in the Communication and Maladaptive Behaviours domains, for the low responding subgroup in the Communication and Socialization domains, and for the average responding subgroup in the Socialization domain alone. As previously noted, this regression around the two month follow up is likely due to large disruptions in home and school life, as this data collection point aligned with the school holiday period, with some tests falling very close to Christmas and New Year celebrations and others very close to the start of the new school year. Events such as these are very disruptive and can be challenging for individuals with autism to manage given their strong preference for routine and predictability, and it is therefore not unexpected for such a decline in test performance to occur around this time.

Breaking down the longitudinal Vineland-II data into response level subgroups reveals further support for the suggestion that the high responding subgroup were most benefitted by the Social Tutor software. Confirming the trends observed in the content quiz longitudinal data, it appears that these participants not only increased their social skills while actively using the software but were able to continue to do so afterwards, possibly applying the knowledge gained in the Social Tutor to other learning opportunities or practicing in real-world scenarios, and thus improving their everyday performance of social skills as observed by caregivers. However, from the previous graphs in Figure 23 it is clear that, with the exception of high responders in the Socialization domain as discussed, the control group and experimental group perform very similarly overall and thus much of this improvement is likely to be due to factors outside of the Social Tutor software, in particular those already mentioned such as increasing maturity, exposure to other social skills interventions in addition to the Social Tutor, and opportunities to practice and consolidate these developing social skills in everyday life.

7.5 Research Objective 5 - Perceptions of Software

To address Research Objective 5 and "determine participants' and caregivers' perceptions of the software" participants were asked to complete a simple questionnaire before using the software, and both participants

and caregivers were asked to complete a more detailed questionnaire after the three weeks of software use was over. The goal of the pre-test questionnaire was to determine participants' previous experiences with related technologies and their expectations for the Social Tutor, and it was found that participants generally displayed a similar profile of prior technology experience and expected that the software would be both enjoyable and educational. The aim of the post-test questionnaire was to investigate both participant and caregiver experiences with and perceptions of the Social Tutor following its use, and to elicit any feedback to help inform future development of related software. While a number of areas that would benefit from modification were identified, overall feedback was positive and indicated that a tool such as the Social Tutor would be worthwhile to continue developing. Participant and caregiver suggestions specifically relating to recommendations for future development of the software are discussed in detail in Chapter 8, while all other questionnaire content is discussed here.

7.5.1 Pre-test

Almost two thirds of participants indicated that they use computers daily, with the rest using them several times a week or at least once a week, and a strong majority of participants either agreed or strongly agreed with the statement "I enjoy using computers" and indicated that they considered themselves good at using computers. This is in line with the survey by Putnam and Chong (2008) and experiences of Baron-Cohen et al. (2009), which suggest that children with autism often feel more comfortable interacting with technology than with their peers. It also echoes the findings of MacMullin et al. (2016) which show that individuals with autism have high daily usage patterns when it comes to technology. Together, this supports the idea that a software-based social skills tutoring program has the potential to both engage and benefit children on the autism spectrum as it harnesses their existing interests and can be easily incorporated into their established technology usage patterns. These responses also support the appropriateness of the inclusion criteria for the current study and suggest that the characteristics of the participants recruited match the intended target audience.

Regarding participants' perceptions of their own educational abilities, less than half indicated that they felt that learning new things came easily, with many neither agreeing nor disagreeing with the statement. The strength with technology was again highlighted, with over half of participants indicating that they learn how to use new software quickly, despite the lower rated responses for general learning. This suggests that for children who may be accustomed to finding new things challenging, teaching via a technology-based medium, something they report higher confidence with, may assist to reduce barriers by increasing willingness to engage with the new learning task, and making the experience more enjoyable overall.

In terms of expectations of the software, just over half were unsure what to expect, indicating that they neither agreed nor disagreed with the statement that the software would help them learn, however a large portion of the remainder were optimistic, either agreeing or strongly agreeing with the statement. For entertainment value, most participants anticipated that the software would be fun to use, about a third were undecided and only a few were not expecting to enjoy it. Participant responses to the open-ended question

"What do you think it will be like using the Talking Head software?" further support these Likert-scale based responses. While post-test questionnaire outcomes suggest that fewer participants found the software fun to use than hoped, this may be because the Social Tutor was designed and in development prior to the widespread rise of gamification in educational software (Hamari et al. 2014). As such participants may have been expecting the Social Tutor to be more game-like than it was, particularly given that gamified educational software is likely to account for much of what they have experienced in the past.

Finally, participants were asked "What do you normally use computers for?" and all participants indicated that 'playing' was one of the purposes. Around half the participants indicated that 'school work' or 'watching videos' were also common uses, followed by 'browsing' or looking up information, with a small number of participants also mentioning science, programming or creative arts related hobbies. Only two participants indicated that they used computers for social purposes, in this case emailing friends, and no participants indicated any other social purposes, although there may be social use within the context of playing games such as Minecraft. Given that all participants are primary school children, the lack of social computer use is likely to indicate appropriate internet use for their age rather than relating to autism specifically.

Interestingly, just over half the participants indicated that they use computers for school work despite computers and technology being a mainstay in most classrooms. Given that the researcher visits occurred within the family home and often on weekends, this may simply be because participants were not focussed on the school context at the time. Overall the participants displayed similar profiles of computer use across the cohort, typically including a variety of age-appropriate activities and in all cases using computers primarily for recreational activities.

7.5.2 Post-test

The post-test questionnaire sought to gauge participant and caregiver perceptions of the software, covering factors including usability, enjoyment, educational value and length and difficulty of activities. It also sought to elicit feedback on both the positive aspects of the software and ideas that families had for future improvement of the software. While responses relating to usability, enjoyment, and educational value and their implications are discussed here, a dedicated presentation of software-specific changes recommended for future development can be found in Chapter 8.

Rating Scale Responses

The Likert-style items of the post-test questionnaire were all framed in a positive sense, with higher scores indicating stronger agreement with the given statement. Encouragingly, the mean response for most statements is higher than the centre point, indicating that participant and caregiver responses are in most cases tending towards agreement with the positive statements and suggesting that in general caregivers and participants found the experience of using the Social Tutor a worthwhile one.

As anticipated, participants and caregivers in the experimental group agreed on average that the 'software helped [me] to learn' and that 'the topics were useful'. This is encouraging and suggests that the content developed for the experimental group software does target skills that are desired by learners on the spectrum and recognised as areas of need by their caregivers. Given that the content presented to the control group was not intended to contain educational material, it was predicted that the two statements relating to educational value would receive low ratings from the control group participants and caregivers. Most participants and caregivers stated that they neither agreed nor disagreed with the two 'educational value' statements, although more strongly negative responses were anticipated.

Unexpectedly, mean responses from control group caregivers actually indicated slight agreement with both educational value statements. This could be due to a number of factors, for example caregivers were not required to closely supervise their children while using the software and thus may not be fully aware of the activities they were undertaking, caregivers may have observed their children interacting with the virtual people by way of asking for maze hints and interpreted it as explicit social teaching, or they may be responding positively due to social desirability bias, believing this to be the 'desired answer' (Grimm 2010) or the expected outcome, despite the researcher encouraging them to answer completely honestly. In contrast to caregivers and in alignment with anticipated responses, control group participants on average disagreed slightly with the statement 'the software helped me to learn' and neither agreed nor disagreed that 'the topics were useful'.

The next set of statements related to enjoyment of the Social Tutor, and it was discovered that control group caregivers and participants rated the software more favourably in every case, including for the statement 'the virtual humans were friendly', although the experimental group did also agree with this statement on average. In contrast, the experimental group disagreed that 'interacting with the virtual humans was fun' and neither agreed nor disagreed that 'the activities were fun'. This highlights the impact that lesson content enjoyment has on user ratings of the virtual humans themselves. Interestingly, the virtual people were perceived as somewhat friendlier in the control condition even though their interactions were much more limited and repetitive than in the experimental condition. This may be because in the control condition there was no need for participants to be informed that they had an answer wrong, unlike in the experimental condition. Being corrected by the virtual humans may have been perceived as negativity or unfriendliness. While every effort was made to provide encouraging feedback, some participants may have felt that even an 'oops!' or 'not quite, try again' was negative. The work by Fletcher-Watson et al. (2016) suggests that ideally incorrect responses should be met with no response or simply guidance towards the right answer, so a minor adjustment of the feedback system may help reduce this perceived negativity.

Another possible explanation for the lower friendliness ratings by the experimental group could be that when users needed hints or help, it may have been less obvious how to get that information for experimental content due to the wide range of different activity types, or it may be that the support the virtual characters provided was not sufficient. Since immediate feedback is thought to be a core reason for the effectiveness of human one-on-one tutoring (Chi et al. 2001, Bowman-Perrott et al. 2013), adjusting the system so that the

Social Tutor can detect instances where the learner needs more help with the functional aspects of the software and not just with the educational content may help prevent learners losing momentum and improve their educational experience (VanLehn 2011). Overall the responses for participants and caregivers from both conditions were positive for perceived virtual human friendliness, however there is still room for improvement.

Usability statements were assessed next, and while caregivers and participants in both groups agreed with the usability-focussed statements overall, the control group again tended to agree more strongly with the positively worded statements that the experimental group. This provides some support for the notion that simpler is better, as the control group content was less complicated, less diverse, and more repetitive, but appears to have been more positively received. Interestingly, control group participants and caregivers rated 'the voices were clear' more highly than those in the experimental group even though identical voices were used for all participants, however this difference did not reach significance. This may be because in the content for the control group the spoken scripts were more repetitive and presented in smaller chunks, so it was easier for users to predict and interpret what was said even if the voices were perceived as being poor quality or there was distracting background noise in their learning environment.

For the usability statement 'it was easy to see how I was going with activities' both the control and experimental groups indicated similar levels of agreement, suggesting that the rewards system was helpful in providing learners with a sense of the progress they were achieving. It should be noted however that participants and caregivers were not explicitly directed towards the 'View Progress' function described in Section 4.6 as it was deemed non-essential and effort was made to avoid overwhelming participants when they first started using the software. Directing them towards this feature in future may assist to increase agreement with this statement.

Finally, a statistically significant difference was found between participant responses from the control and experimental groups for the usability statement 'it was easy to choose the activity I wanted to do'. This may be due to the algorithm of the automated assessment and dynamic lesson sequencing functionality described in Subsection 4.6.3, which can result in lessons that are attempted but unfinished not being made available again to the student until they have completed a few other unattempted lessons. If a student opens a lesson but decides to come back and try later, they may become frustrated if the lesson is not offered to them again immediately. In the control group the impact of this is expected to be less noticeable since all tasks at the same difficulty level are very similar, and given the nature of the maze activities presented to the control group, participants are more likely to simply complete the task the first time they open it rather than exiting and wishing to revisit it later.

Following the set of Likert-style questions, participants and caregivers were also asked to rate the difficulty of the lessons and the amount of time spent on each one. Responses indicate that the time spent on lessons was generally 'just right'. Since the aim during software development was to create lessons that typically took only a few minutes each to complete and student log data as described in Section 6.2 indicates that this aim was met, it is encouraging that the feedback from participants and caregivers reinforces that this was an

appropriate lesson length to aim for. Ratings of difficulty were more varied, although on average participants appeared to feel the lessons were 'just right' or 'too easy'. While this is somewhat preferable to lessons being too hard and causing frustration or disengagement, lessons being too easy is also problematic as learners may become bored and it may also indicate that learners are not being sufficiently extended.

Modifications to the dynamic lesson sequencing algorithm may assist to improve these ratings, for example by incorporating an assessment when a new account is created to determine where the learner should start rather than having all users start at the same point, and by giving students ongoing opportunities to 'fast forward' through content they are finding easy. Thus, overall responses to the rating scale questions were encouraging, but room for improvement remains.

Open Ended Responses

A detailed discussion of recommendations and possible directions for future development of the Social Tutor and related software is provided in Chapter 8, with much of that drawn directly from the open ended responses of the post-test participant and caregiver questionnaire. Here a summary of key feedback is provided, particularly relating to lessons learned through the evaluation process.

First, a number of strengths were identified by study participants and their caregivers. Participants typically found the characters to be friendly and supportive, and enjoyed interacting with them. Several users reported enjoying the virtual role-plays, and the videos presented to the experimental group were also flagged as a highlight by several participants and a caregiver. The lesson content was noted as being relevant, sufficiently varied, and appropriately repetitive so that learners had opportunities to practice and consolidate their knowledge. Several participants in the control group reported particularly enjoying their maze activities and, as discussed previously, the control group reported higher levels of enjoyment overall. Taken together this suggests that more gamification of the Social Tutor content could be worthwhile, and emphasises the benefits of incorporating a variety of digital media for both learning and general engagement purposes. This is also consistent with the current trend of gamification for educational software which was not so prominent when design of the experimental software was initially undertaken, however some relevant examples that inspiration could be drawn from include the Emotion Bubbles software (Madsen et al. 2008) which is approaching the realm of augmented reality, utilising cameras and software to assist learners to identify and demonstrate various facial expressions, and LIFEisGAME (Abirached et al. 2011) which has virtual people mimic the facial expressions of the learner. While not included in the Social Tutor due to the need for a webcam, deploying the software on a mobile device is likely to overcome this barrier and make such interactive elements more appropriate for future inclusion. For both Emotion Bubbles and LIFEisGAME, these game-like elements led to increased engagement by participants while intrinsically reinforcing the target social skills being taught, something that would likewise benefit the Social Tutor software.

A large number of participants reported that the sticker and rewards system was a particular highlight, this being the most highly repeated positive comment across all categories. This supports the inclusion of an

extrinsic rewards system to encourage users to persevere with their lessons. The software as a whole was perceived as easy to navigate and interact with, and caregivers appreciated that their children could use it independently. Related to interaction, one caregiver reported that dragging the mouse was difficult for their child. While tablets and other mobile devices were not in wide use when development of the Social Tutor began, the software was designed to be touch screen compatible. Therefore, porting the Social Tutor to a mobile environment may help to ease the fine motor control related challenges that some users experience.

Interestingly the character voices were reported as both a strength and a challenge, with some participants finding them clear, easy to understand and appreciating that they spoke slowly for clarity, while others found the voices too monotone and robotic, or found their slower speech and having to wait while they talked frustrating rather than helpful. As can be observed in Table 40 of Subsection 6.6.2, experimental group participants and caregivers flagged behaviours of the virtual characters themselves as problematic much more often than control group participants or caregivers. This may be due to higher expectations given the nature of the content being presented, or it may be that the relative complexity of the experimental group content caused these issues to be more noticeable or disruptive. While the slower speech with pauses to allow for comprehension was a conscious design decision during development to ensure that users would have enough time to listen and process everything the virtual people said, given the feedback adjustments are clearly required to strike a better balance. Similarly, many participants found having to wait for the virtual people to stop talking before they could interact with the software frustrating, however again this was a conscious decision designed to encourage children to pay full attention to what their virtual teachers and peers were telling them, as would be expected when interacting with real human peers and teachers. As likewise noted by Tartaro and Cassell (2006) and Piper et al. (2006) in the development of their technology-based social skills interventions, striking a balance between appropriate flexibility to encourage engagement and interaction and sufficient structure to support positive learning outcomes can be challenging. Finally, to address the monotone and robotic voice issues more realistic synthetic voices would need to be sourced, which is particularly difficult given the scarcity of good quality Australian child voices.

Incorporating customisation options around speech speed and allowed behaviours while the virtual people are talking, possibly with the option of these settings being password protected so that caregivers and educators could determine what combination is most educationally appropriate for their individual children, may help to improve the voice related difficulties. Increased personalisation of other aspects of the software was also suggested by participants and caregivers, and incorporating this may have the two-fold effect of ensuring that the presentation environment suits the sensory and learning needs of the individual, while also leading to increased feelings of ownership and engagement for users.

The appearance of the virtual humans was also identified as an area that could improve, with a number of participants and caregivers in the experimental group feeling that they would benefit from either a more realistic appearance or a more light-hearted cartoon-like appearance, along with both increased idle motion and increased expression. Given advancements in virtual human technology since development of the

software commenced, most of these appearance-related issues should be feasible to address in future development. The suggestion of making the virtual humans more realistic is more likely to be pursued than that of making them cartoon-like given that the current research aims to capitalise on the known benefits of video modelling as discussed in Subsection 2.2.1, and these benefits are predicted to be less likely to be replicated with cartoon characters.

Homework was also identified as both a positive and a challenge, and from anecdotal conversations and observations it appears that this feature has the potential to become a powerful tool for connection between the software and real-world experience. Particularly noteworthy was the participant who successfully performed their homework task once, but when prompted by their caregiver to repeat it on a different occasion, did not understand why they should since their homework was already done. This highlights that homework tasks as standalone, one-off activities are not sufficient, but by instead approaching homework as an ongoing activity, possibly earning a reward each time a target behaviour is enacted, it may have the potential to turn newly learned behaviours into habits.

Some of the learning tasks were identified as too hard or confusing, with a few participants experiencing this with the mind-map style activities. This may be because tutorial lessons explaining how to do mind-maps and other activity types were optional, and while explanations were available on the page of the activity itself, participants may not have realised how to access them or may have found the instructions insufficient. This suggests that further refinement of activities and more supportive hints when users are experiencing difficulties are needed. Related to this, a caregiver also suggested more feedback be included. Several mechanisms are in place with the aim of providing this kind of information to participants, however there is scope for more in-depth feedback to be incorporated.

There were also a number of suggestions from caregivers that were in fact already features of the software. The first was a request for more conversation skills content. As previously discussed, an entire topic dedicated to this was included, namely the 'Beginning, Ending and Maintaining Conversations' topic, however no participant completed enough prerequisite lessons to unlock it during the three week software use period. With longer term use of the software or improved mechanisms for more advanced students to fast forward through earlier content this would have become accessible. The ability to take a break was also requested, however this is already included in the timer system and was explained to caregivers when the software was installed, so it is just a matter of increasing awareness.

In general the software was well received by participants and caregivers alike, and would be even further enhanced with some relatively straightforward adjustments to the areas identified, in particular the inclusion of customisation options around the sound and speed of the virtual humans' voices and related behaviours, more guidance and feedback, and an improved dynamic lesson sequencing system. Spontaneous follow-up contact from a number of caregivers noting that they could observe a change in their child's social behaviours and that they felt their children had benefitted from use of the Social Tutor was very encouraging.

CHAPTER 8. SOFTWARE FUTURE DIRECTIONS

The preceding evaluation of the Social Tutor suggests that this application of virtual human technology not only has much promise for helping children with autism improve their social skill, but it addresses a need in a way that is accessible for families and engaging for users. The current evaluation was relatively small in both duration and sample size, and so further research is needed to identify the specific factors that most impact participant outcomes. Similarly, further development is needed to refine the software based on feedback from participants, caregivers, and educators. From the experience of developing and evaluating the software, and from the feedback received from the post-test questionnaires presented in Section 6.6, there are many recommendations for consideration should further development of the Social Tutor take place. These include features to retain, general modifications, modifications relating to educational content, additions to personalisation and customisation capabilities, improvements to the automated assessment and lesson sequencing algorithms, and suggestions for increasing the authenticity of interactions between users and the virtual characters.

8.1 Retained Features

From the post-test questionnaire feedback obtained from participants and caregivers (see 6.6.2) there were a number of features that were particularly appreciated by users and should be retained in future iterations of the Social Tutor. The most prominent of these was the sticker system, which many participants commented was their favourite aspect of the software. This system rewarded participants for completing tasks by giving them a gold star, and once five gold stars were earned a virtual sticker could be chosen. This gave participants a sense of progress and achievement, but was not very time consuming and so did not detract unduly from participants' lesson activities. Also connected to this rewards system was a mechanism to unlock reward games at 50% and 100% topic completion, however due to the short duration of the evaluation only a few participants unlocked any of these. Adding additional reward games and lowering the amount of completed content required to access these, for example unlocking games at 30% intervals, may be more motivating.

Another feature appreciated by caregivers in particular was the timer, which let participants know when their period of software use for the day was up. The timer was deactivated in the 'unlocked' version of the software so that participants choosing to continue using the software after the evaluation ended would not be limited in their usage, however it may be preferable to reintroduce the timer but remove the "forced shut down" element and retain only the "prompting" element, so that participants can choose when they close the software, but are still notified when they have reached the minimum recommend duration for the day.

Digital content such as videos and songs were also appreciated by both participants and caregivers, as was the variety of content and the repetition of content. While the precise nature of suitable content depends

heavily on the curriculum being implemented, it is recommended to include engaging digital media such as videos and songs, and a variety of activity types, in order to maintain user engagement.

The final feature that is particularly recommended for retention, albeit with refinements, is the homework system. While it was hypothesised that participants would choose not to engage with the homework activities given that they were optional, several participants did choose to do them and applied the skills they were learning in the software to simple interactions with their peers and family members. Thus, this appears to have promise for promoting generalisation beyond the software context. Refinement is needed however, as from anecdotal discussions with caregivers a few issues were identified. The first being that the homework activity asked participants to identify who they would try their skill with, when, and where. For some participants it was quite distressing if they could not complete their plan due to unforeseen circumstances, such as the identified peer not being at school that day. Thus, explicit preparation for dealing with these situations may be a necessary prerequisite skill before offering homework activities of this nature. From conversation with a second parent another issue was identified whereby the participant had successfully completed their homework task previously, and when the parent identified another opportunity to apply that skill and pointed it out to their child, the child replied that they had already done that homework and did not need to do it again. Modifying the homework activities so that rather than being one-off tasks, they instead keep a tally of the number of times the participant has used their skill in a real-world context and then reward participants for regular skill application, may be a promising alternative approach.

While it is expected that the general structure of the software could remain largely the same as the current iteration of the Social Tutor, the features discussed here were particularly appreciated by participants and caregivers and are therefore recommended for retention in some format in any future Social Tutor program.

8.2 General Modifications

A number of desirable general modifications have been identified, primarily in response to changes in software development trends and to accessibility of available technology. The most prominent of these is the recommendation to port the Social Tutor to a mobile platform. Given that the current Whiteboard software is written primarily in Java, porting it to an Android environment is expected to be most straightforward, however it would be desirable to port it to a more universal system where it can be made available in multiple formats easily, further increasing its accessibility for families, many of which have iPads and other non-Android devices. When development of the Social Tutor began tablet computers were not widely adopted by families and mobile devices had capacity only for simple programs, making them inappropriate targets for the Social Tutor given that accessibility has always been a core aim. With tablet computers and powerful mobile devices now being widely used by families, their portability and easy to use touch screen interfaces make them a particularly desirable platform for the Social Tutor.

In addition to this, porting the Social Tutor to a tablet computing context opens up opportunities for collaborative learning, for example by utilising the touch screens as shared surfaces for multiple

simultaneous users in a similar fashion to the approach used in the SIDES table top game (Piper et al. 2006). Further, as tablets are portable, they open up opportunities for in-situ learning, where the user can take the Social Tutor with its helpful virtual characters into the real-world situation they are learning about, and use it both for helping them to interpret the situation in front of them and for providing advice on appropriate behaviour for the context. One source of inspiration for this approach is the Emotion Bubbles trainer developed by Madsen et al. (2008), which involved a portable computer with a camera that could be pointed at faces to automatically assess the emotion being portrayed. Similar interactive activities could be incorporated into the Social Tutor, which would be both engaging and educational.

Building on this, another area that has been around for some time but gained particular momentum in recent years is that of 'serious games' (Michael and Chen 2005, Boyle et al. 2016). Anecdotally, in several cases participants expressed that they were expecting the Social Tutor to be a game, although the researcher was always careful to explain that it was 'educational software' with 'activities and lessons'. This emphasis on learning through play is particularly applicable to younger children, and still relevant and effective for older users, making inclusion of more game-like elements or gamification of the Social Tutor as a whole a recommendation for future development.

A number of smaller but still impactful recommendations relating to communication of information within the Social Tutor include making sure all spoken instructions are clear and concise, providing a summary at the end of each lesson to reiterate the main point of the task, and engaging a graphic designer or illustrator to provide clear and consistent visuals across the software. This would make the Social Tutor more cohesive, appealing and professional, but more critically would also ensure visuals are consistent and therefore more meaningful to learners who are relying on them for interpretation of the skill steps and activities. These recommendations are relatively straightforward but have the potential to greatly improve the user experience.

8.3 Educational Content

The educational content in the Social Tutor was by necessity drawn from a set of existing curricula, however these curricula were not purposefully designed to be implemented in a software context. This resulted in some difficulties fitting the content and activities to the context and meant that some of the most powerful learning opportunities, such as truly interactive open-ended role-plays and observational based learning, could not be fully utilised. Further, given the relatively small scale of this first Social Tutor evaluation, only a small section from each of the chosen curricula was implemented. A more powerful approach that aligns better with the ultimate intended purpose of the Social Tutor would be to work in collaboration with a curriculum or intervention development team and build the content of the Social Tutor directly around the content in the curriculum or intervention itself. This would ensure that the software fulfils its purpose of reinforcing content that users are learning through other avenues and would ensure that potential conflicts are also avoided, for example if the skill steps taught in the Social Tutor are slightly different to the steps taught in a school-based intervention it could cause confusion for learners, whereas developing these two resources

in conjunction would avoid such situations. As discussed in Subsection 2.2.2, existing research strongly indicates that this hybrid approach can lead to positive learning outcomes (Beaumont and Sofronoff 2008, Whalen et al. 2010, Jones et al. 2016).

Creating the Social Tutor content directly in conjunction with a curriculum or intervention developer would also result in the complete curriculum being addressed in the software, which would in turn open up more opportunities for tailoring the presented lessons to the learners' needs and possibly allow caregivers and educators to select what topics their learner should be focussing on. In the current version of the Social Tutor only a small set of three topics were implemented, which does not allow for as much personalisation as a broader set of topics would enable. In saying this, caregiver and participant feedback indicated that the topics chosen for inclusion were felt to match the needs of this learner group well, and given the functionality of the virtual humans these skills are likewise a good fit for the capabilities of the software. Thus, conversational social skills are recommended for inclusion in any future Social Tutor development. Additionally, caregivers and participants appreciated the variety of activity types presented and the repetition of core concepts across these various activities, so again these are features that should be retained.

8.4 Personalisation and Customisation

Following post-test questionnaire feedback, a number of settings are recommended for inclusion in future Social Tutor software to allow further personalisation of the virtual humans' appearance, voice and behaviour. The virtual humans' speaking speed was deliberately set quite slow to accommodate users with poor auditory processing abilities and those relying on reading subtitles, however many participants found the slow speed and the long pauses frustrating rather than helpful. Therefore, it is recommended to include a setting to allow the user to control the speech speed of the virtual characters independently. Related to this, while the synthetic voice for the virtual teacher was appropriate, being clear, relatively natural sounding, and with an Australian accent, the synthetic voices for the two child characters were both American English speakers and of poorer listening quality, with only one being a genuine child voice and the other being an adult female with the settings adjusted to be as child-like as possible. The variety and quality of synthetic voices available now has improved and is expected to continue to improve, so for any future Social Tutor development it is recommended to purchase more appropriate voices for the child characters in particular.

Feedback also indicated that the synthetic voices were considered too monotone, so obtaining newer, more emotive voices would also address this issue. For truly accurate and emotive voices the best solution may be to instead employ voice actors and pre-record all content, however this would introduce severe limitations in terms of personalisation and extension of the content of the Social Tutor. For example it would be very difficult to facilitate caregivers and educators modifying or creating their own lesson content. Also, having the virtual characters use a participants' name to address them is a powerful engagement tool, as seen in the preceding work by Milne et al. (2009), and very appropriate given the content being taught, but this would be challenging if using a pre-recorded voice actor, particularly for uncommon names. Synthetic voices are

therefore still the solution of choice for this context, and given that sensory difficulties are a common co-morbid condition of autism spectrum disorder and that preference for voices varies with individuals, a variety of voices for all of the virtual characters could be offered for users to select from. Also related to speech, the software disables interaction with the Whiteboard if a virtual character is actively talking in order to ensure users are listening to the instructions they are being given. This was a deliberate design choice intended to reinforce positive social behaviours, however many participants found it frustrating that they could not skip instructions. Given this, a setting to turn this function on or off may be appropriate, but optionally password protected so that the caregiver or educator is the one to make the decision rather than the user themselves.

Another suggestion from the post-test questionnaire was to offer a choice of virtual characters, possibly including cartoon, animal, or fantasy characters that children might find more engaging. Existing research supports this suggestion, for example work by Mei et al. (2015) with adolescents on the autism spectrum showed that being able to customise the virtual character resulted in improved engagement, user experience and even better performance on the target task. Given that the benefits of using virtual humans are hoped to be similar to those of using video modelling (see 2.2.1 for more detail) it is expected that retaining the current human appearance of the characters would be most beneficial, however more research is needed to investigate this. Taking the video modelling concept a step further and focussing on self-modelling, using FaceGen software (Singular Inversions 2017) it is possible to develop virtual characters based on a small set of photographs of an individual. This model could then be used in the Social Tutor as the 'socially skilled' child, allowing the user to see a virtual version of themselves demonstrating good social skills. An example of a model created using this method can be seen in Figure 32. Some issues exist with this idea, including finding a synthetic voice that suits the individual being modelled, and sourcing or creating suitable hair models to complement the generated face, since existing research indicates that coherence between an agents' appearance and behaviours impacts users' interactions with and perceptions of the agent (Astrid et al. 2010, Skarbez et al. 2017), along with possible difficulties with a self-model being perceived as unsettling and possibly falling into Uncanny Valley territory (Mori 1970), however it is a potentially promising area for future research.

Certainly offering some customisation of the appearance of the virtual characters would be appropriate and achievable, particularly if done in conjunction with customisation of character names and a selection of voices so that authentic male and female characters could be created by users. Allowing such control over the learning environment has the potential to increase the sense of ownership that users experience, hopefully leading to an increase in positive perception and willingness to engage spontaneously with the software more often.

Another aspect worthy of further investigation around virtual character personas would be determining if the learner gains most from the models being similar to or different from themselves, and possibly utilising both in different learning tasks to enhance outcomes. While the benefits of self video-modelling are an area of

interest as previously discussed, it has also been shown that using models different to the learner can assist neurotypical individuals to learn empathy (Yee and Bailenson 2006, Peck et al. 2013). By challenging biases and providing novel interaction scenarios, individuals with autism may be able to develop more effective communication and social skills.

Figure 32: Example of a self-model created using FaceGen, with a photo of the target on the right.

Building on the theme of virtual character appearance, in future Social Tutor development full-bodied virtual characters are recommended in place of the current head-and-shoulders-only models. While these were not easily accessible when this iteration of the Social Tutor was being developed, they are now and would allow for more authentic display and explanation of nonverbal communication, such as hand gestures and body language, which would be highly beneficial to this learner group.

A feature that has been included in the Thinking Head Whiteboard but not utilised by caregivers or educators in the current evaluation is the XML-based lesson authoring system described in Subsection 4.4.1. This has been designed to facilitate personalisation of the lesson content by non-programmers, for example allowing them to easily replace default images with images that their learner would find more engaging, or modifying the speech of the virtual characters to directly reflect a phrase the user is learning in other interventions. The XML used to create lesson files has been designed with robust default behaviours so that basic activities can be created with a minimum of code, allowing educators and caregivers to add lessons themselves even without extensive programming experience. However, the lesson and curriculum files are still created by manually writing XML, so it is expected that this would be off-putting for many educators and caregivers who lack relevant technical experience. An ideal solution would be to develop a companion program that supports modification and management of curriculum files and provides a drag and drop lesson builder interface for creating new activities and editing existing ones.

Moving away from manual methods of customising content and the virtual characters themselves, another way to personalise the learning experience would be to introduce adaptive student models. Wittwer et al. (2010) emphasise the benefits of adapting instruction to the individual learner, including encouraging

researchers to consider implementing detection of nonverbal cues. In the Social Tutor, adaptive student models could be provided with initial information when the user creates an account, but then be continually fed data about the individual user's interaction with the system. The Social Tutor already logs direct interaction with the system and information about user achievement levels by lesson, objective, and overall topic, so it is well situated to be developed further in this manner. Further, the pre-test quiz that was used for the software evaluation could be retained as a feature and the results fed directly into the student model, or inspired by the suggestions of Sansosti (2010), an input mechanism could allow educators and caregivers to provide an initial overview of their learner's abilities. Possibly in addition to these mechanisms, participants could periodically be prompted to complete a mini quiz, with strong performance fast tracking the learner through their current material and poor performance indicating they need to be presented with additional support material, much like the approach taken by Jones et al. (2016) in their TeachTown: Basics software. All of these approaches would enable students to skip past lesson content that they have already demonstrated mastery for in other contexts. This initial information combined with ongoing log data and quiz data would be used to determine user preferences and areas of strength and weakness, and from this a more dynamic and targeted user experience could be provided. For example, different lesson activities could be offered depending on whether the user is a more visual or more aural learner, or if signs of frustration were building up the software could backtrack and offer simpler lessons or prerequisite lessons as a refresher, before offering the challenging content again.

8.5 Assessment and Sequencing Algorithms

Building on the idea of personalisation and incorporation of a student model discussed above, it is recommended that the lesson sequencing algorithm be updated to be more adaptive and to draw from a student model, rather than simply being based on a set of heuristics and static rules as it currently is. This could enable it to adapt to individual learning styles, do a better job at determining when a student is ready to progress to the next complexity level, and potentially be more targeted when a student is struggling, for example by giving them content that addresses their specific difficulty rather than just generically back-tracking them and repeating content they have already progressed through. To support inclusion of a rich student model, continuous assessment would be required. This is already built into the Social Tutor to an extent, with log data being recorded for every interaction a student has with the software. This log data includes performance information such as the time they spent on the activity (duration), how many correct and incorrect moves they made (accuracy), and what the final state of the activity was (correctness).

In addition to this log data, both reflective and affective information could be gathered and incorporated into the student model. One simple way to gather reflective data could be to periodically prompt users to complete quick self-assessments rating how they feel they are progressing. While this is not a reliable measure of social skill competency, self-reflection is a valuable skill in itself, and reflective practice has been shown to support deep learning and long-term retention of skills (Hattie and Timperley 2007, Nicholas et al. 2015). It should be noted that the style of reflection has been shown to impact the quality of retention and

memory, and thus the recommendation is to use open-ended questioning involving high levels of detail, possibly as an interactive replacement for the current virtual teacher-led 'recap' at the end of each lesson (Nicholas et al. 2015). Additionally, data from self-assessments could be integrated with other measures of affect detection to help identify and respond to emotions that are known to impact the learning process, such as frustration and engagement levels. Incorporating affect detection and having the virtual teacher pick up on and respond to these emotions could be very valuable in both building a relationship between the user and the virtual characters, and for managing the educational experience. By detecting happiness or satisfaction after a lesson is successfully completed, the virtual characters can celebrate in a meaningful way with the learner, building rapport. By detecting frustration the virtual characters can support the student, possibly redirecting them to a simpler activity and ensuring that their prerequisite knowledge is mastered before challenging them with the next level of content. Detecting boredom could trigger the software to present a pop quiz, with good results in the quiz allowing the learner to fast track through the content they were finding unstimulating.

From the post-test questionnaire, it has been identified that the dynamic lesson sequencing algorithm can actually be a cause of user frustration. The algorithm prioritises new, previously unseen activities and, if available, offers them before re-offering activities the user has tried before. For some participants this is frustrating, as they may feel ready to retry a particular activity, but if the software has three or more 'new' activities it will offer them rather presenting the activity the student may be looking for. Modifying the way this works to have two lists of activities, one being the three 'suggested' activities the current algorithm identifies, and a second easily accessible but perhaps less prominent list of 'previously attempted' activities may be one possible solution.

Also from the post-test questionnaire, caregivers have indicated that they would appreciate more targeted feedback, specifically suggesting that a summary of the skill steps at the end of each lesson would be beneficial. In addition to this worthwhile suggestion, ensuring that feedback is provided not only at task-level but also process-level is recommended to give users a chance to consolidate their learning more effectively, and combining this with self-assessment and reflection as discussed above may assist to further develop learners' reasoning and problem solving skills, as well as consolidating knowledge (Nicholas et al. 2015). Care must also be taken when considering what type of feedback to present to students. For example, existing research has shown that when students are presented with marks, they use these to compare themselves with others, but when they are provided with comments and suggestions instead, they use the feedback to improve their skills (Black 2015). Unsurprisingly, students given comments outperform those given marks alone.

8.6 Authentic Interaction

Many technologies that lend themselves to authentic human-computer interaction have advanced significantly over the time in which the current Social Tutor was developed, and this presents us with new

and exciting options for incorporating more realistic learning scenarios into the Social Tutor. As already discussed, simply using better quality synthetic voices and full-bodied virtual characters has the potential for increasing the educational value of the software. The idea of emotion recognition and response has been discussed in terms of its potential benefits for guiding the educational process, for example detecting frustration and supplying additional support or redirecting the student to more fundamental activities, however it also has much potential both for increasing rapport between the user and virtual characters, with the virtual characters being able to comment on and respond to the user's expressions, and for developing unique interactive learning activities. Recent research has seen the development and validation of a number of computer vision based approaches to engagement and emotion detection that appear suitable for future incorporation in the Social Tutor provided a camera is made available, for example see Grafsgaard et al. (2013), Whitehill et al. (2014) and Monkaresi et al. (2017). Activities could cover understanding one's own emotions, as well as detecting and responding to the emotions of others. Coupled with a mobile device, this could present some very interesting learning opportunities for users to explore the world around them and the people in it.

Potentially tying together the ideas of implementing a student model and incorporating emotion detection, Krämer (2006) suggests that implementing a 'theory of mind' for the virtual characters may go a long way towards improving their likeability, while also helping the software to make better judgements of the learner's current state and needs. Existing work has shown that a virtual character that can empathise with the student leads to students becoming more interested and displaying higher self-efficacy. To do this effectively, the virtual tutor must be able to detect the learner's emotion and respond appropriately (Krämer and Bente 2010). As discussed previously, learning is intertwined with emotion, so being able to respond to learner emotion can potentially improve educational outcomes.

While the first technique that often comes to mind when emotion detection is mentioned is that of visual detection from camera feeds, this is not the only possibility. Robison et al (2009) compared the use of task-based and affect-based feedback during user interactions in the exploratory narrative-based learning environment Crystal Island. When users interact with agents in the environment, they are prompted for a self-report of affective state wherein the user selects from nine available emotions: anger, anxiety, boredom, confusion, curiosity, delight, excitement, flow and frustration. The agent then provides a response to this which is either task- or affect-based, such as a hint or empathy. The user is then prompted with '... and you respond' and is required to use a Likert scale to evaluate the agent's response in terms of effectiveness and appropriateness. When the user has finished the session, they are prompted to select their affective state for the last time. It was found that the induced model used to determine feedback was the most effective, which consisted of a decision-tree that incorporated student characteristics, affect, and situation data, and made accurate predictions about appropriate feedback 96% of the time. By including a model such as this in a social tutoring application, it may be possible to maximise both motivation and learning gains. However, some adaptations may be required due to the difficulty individuals with autism can experience when deciphering their own affective states and the reliance of the model on self-reports. Possibly combining

techniques, for example incorporating affective cues extracted from speech and visual data in the decision making process, may increase robustness.

Other technologies that have improved and become more accessible in various ways since the initial development of the Social Tutor are speech recognition, natural language processing, and gesture recognition. While the discussion in Subsection 2.6.4 remains highly relevant, if these continue to improve and become more accessible, they present many exciting opportunities for authentic interaction and practice of social skills within the Social Tutor software. Being able to speak with the virtual characters instead of merely pressing buttons, and having them respond to natural gestures like waving hello, raises the interaction to a level that is much closer to human-human interaction, reducing one possible barrier and potentially making it more likely for skills practiced in the software context to be generalised to real-world situations. Of course, care must still be taken to ensure that inappropriate social behaviours are not inadvertently reinforced, for example having an unfriendly hand gesture interpreted as a wave and encouraged, so there are many challenges in implementing this level of natural human-computer interaction. Generalisation remains a significant challenge when developing interventions for individuals with autism spectrum disorders, and a multifaceted approach is recommended to ensure the best chance of positive outcomes for learners. Building on from this idea, it should also be noted that peer assessment and collaboration is a very important tool for assessment, feedback, and learning (Black 2015), with collaborative story telling already being shown to lead to improvements in social skills for children on the spectrum (Tartaro and Cassell 2008). While not directly measured, it is hoped that some of the benefits of peer assessment were realised in the current Social Tutor iteration due to the human-like nature of the virtual characters, and increasing the virtual characters' abilities to interact naturally may serve to enhance this effect, however the concept of collaboration rather than simply being taught and supported by them is an interesting one, and may also present unique opportunities for improving educational outcomes in an engaging way.

8.7 Participatory Design

For the current research decisions about target skills, content inclusion, and interface design were drawn primarily from existing literature, including large-scale surveys of caregivers and individuals on the spectrum, and feedback gathered from the earlier, smaller scale study by Milne et al. (2009). Input from end users, in this case children with autism, is incredibly important when designing software if uptake and outcomes are expected to be positive. To this end, feedback was again gathered in the current study to inform future development of the Social Tutor, including explicitly requesting recommendations for which features to retain, remove, or add. However, this process only involves end-user input at the very beginning and end of the development cycle, so instead for future development a more learner-centred approach is recommended. In line with current trends in the intervention and assistive technology research community, a participatory design approach is recommended, whereby children with autism, their caregivers and educators, and experts in the field are all included throughout the software design and development process as much as possible (Fletcher-Watson et al. 2016, Parsons et al. 2017).

CHAPTER 9. CONCLUSION

The aim of this research was to create software for teaching basic social skills to children with autism that harnesses virtual human technology and draws on evidence-based techniques and tools, then to evaluate the created software to assess its effectiveness for leading to changes in knowledge of social skills and performance of these skills in everyday situations. It is anticipated that the lessons learned through the software development and evaluation process will provide insight into the potential that this unique approach to social skills education holds and will assist to guide the direction of future related research. The key findings, study limitations, and an overview of recommendations are presented here.

9.1 Key Findings and Significance

The Social Tutor software developed for this research was in general used in the manner intended, with software log data and participant and caregiver post-test questionnaire responses indicating that it was used productively by participants and that the intended amount of time was spent on both individual lessons and on the software as a whole both per session and over the intervention period. This indicates that Research Objective 1 was effectively met and that the software could be used successfully by learners. Following from this and addressing Research Objective 5, post-test questionnaire responses also indicated that participants felt the virtual characters were friendly, the software was beneficial to their learning, and that it was overall easy to use, meeting the goal of providing a non-judgemental learning environment, although more game-like elements and personalisation capabilities were requested for future iterations of the software. Further details of recommended software features to retain, refine, remove and add are outlined in Chapter 8.

In addressing Research Objective 2, analysis of the content quiz data indicated that experimental group participants made a statistically significant improvement in correctness scores from pre-test to post-test while control group participants did not. This suggests that use of the Social Tutor software did directly lead to gains in social skills knowledge, a very encouraging finding which indicates that future research into the development of virtual humans as social skills tutors for children with autism is worthwhile and has the potential to greatly benefit this learner group in the manner intended.

This analysis of content quiz data also led to the post-hoc categorisation of participants into response subgroups based on their content quiz correctness scores, where it was found that a quarter of experimental group participants responded particularly well to the Social Tutor software, markedly improving their correctness scores from pre-test to post-test while also completing fewer lessons on average than their peers and spending less time in total using the software. In contrast, half of the experimental group made modest improvements in their correctness scores while completing more lessons on average and spending more time with the software than the high responding subgroup. The remaining quarter of experimental group participants fell into a low responding subgroup who made negligible improvements in correctness scores despite completing the most lessons and spending the most time on the software of all subgroups.

Inspection of the log data and pre-test quiz scores associated with these response subgroups indicated that high responders typically had lower pre-test scores and were less erratic in their topic and lesson choices during the intervention period, while low responders typically had higher pre-test scores and were more likely to jump back and forth between topics when choosing which activities. While the sample sizes are small for both groups (N = 4 respectively) and this could simply be due to chance, it may also indicate that the Social Tutor was a better fit for the high responders who began with less pre-existing knowledge as indicated by their pre-test scores, and a poorer fit for the low responders who had more pre-existing knowledge. The difference in lesson choice patterns may likewise support this hypothesis, with low responders being more erratic as they attempted to find activities that were sufficiently challenging, yet still managing to complete the most lessons on average of any subgroup, again supporting the notion that these participants had greater pre-existing knowledge and needed more challenging content from the Social Tutor.

Pairing the observation of response subgroup behaviour with the fact that during the three week intervention period no participant managed to complete enough prerequisite activities to unlock the third and final topic of 'Beginning, Ending and Maintaining Conversations' indicates that the current implementation of the automated assessment and dynamic lesson sequencing system is insufficient to meet user needs. Improving the system to ensure that participants with higher pre-existing knowledge can move more quickly through the content and skip activities that are too basic for them has the potential to improve educational outcomes.

Along with the correctness data from the content quiz, both accuracy and lesson answering duration were analysed. Both were hypothesised to be reflective of participant confidence, with results supporting this hypothesis. Accuracy and duration improved in both the control and experimental groups, with the smaller improvements of the control group accounted for by increased familiarity with the software interface and the larger improvements of the experimental group accounted for by this familiarity effect in combination with increased confidence with the content itself. When the data was analysed by response subgroup, some interesting trends were found for answering duration, with low responders greatly speeding up their answering times while high responders actually slowed down in a few instances, an observation that was not found for any other group or subgroup. For low responders this is thought to be reflective of them simply consolidating their existing knowledge and being able to answer without hesitation at post-test, while for high responders it is thought to indicate that they are now recognising that they can answer more questions correctly and are taking their time to do so, instead of simply guessing like they may have at pre-test.

When analysis of content quiz data was conducted at the topic level, it was expected that experimental group participants would show an improvement in the content quiz questions designed to directly assess content from Topics 1 and 2, the two topics that participants did access throughout the intervention period, but negligible improvement would be seen in the questions designed to address Topic 3, the topic that no participant unlocked. Contrary to these expectations, it was found that the experimental group as a whole in fact displayed improvements in the Topic 3 questions on par with the improvements they displayed for the Topic 2 questions. This is encouraging and indicates that participants may be displaying near-transfer

generalisation of the skills they learned in the activities from Topic 2, which teaches prerequisite skills relevant to the content that would have been presented in Topic 3 had it been reached. When these findings were investigated according to response level subgroups, this trend held true for both high and low responding participants, with average responders performing on par with the control group for Topic 3. This suggests that, while some participants were able to exhibit generalisation from one topic to another, for many others more support is still required for such generalisation to occur.

On the topic of generalisation, Research Objective 3 aimed to determine whether increased knowledge of social skills as obtained from the Social Tutor translated into improved performance of these skills in everyday life. The Vineland-II was used to measure this based on caregiver observations of their child's social behaviour. It was found that overall the control and experimental groups performed similarly at all data collection points, indicating that generalisation to novel contexts did not occur for most participants. Interestingly, when the data was once again broken down into response level subgroups, a trend was observed where the high responding subgroup outperformed all other subgroups on several domains and subdomains of the Vineland-II, and the average responding subgroup likewise outperformed the low responding subgroup. While no domain or subdomain reached statistical significance, these trends align with the content quiz results and possibly indicate that, given that changing behaviour and breaking and creating habits is time consuming and challenging, the intervention period may simply not have been long enough for these changes to occur, or the curriculum provided may not have been rich enough for learners to make behavioural changes sufficient for the Vineland-II to detect, given it is a broad measure of social functioning.

Following on from this, to address Research Objective 4 and investigate whether any changes in knowledge or behaviour detected over the three week intervention period were maintained once software use ended, content quiz and Vineland-II data were again collected both two and four months after the immediate post-test. Analysis of this longitudinal data for the content quiz found clear evidence that experimental group participants not only maintained the gains they achieved during the intervention period but continued to improve their knowledge after software use ended, and this was especially true for high responding participants. For the Vineland-II data, overall the control and experimental groups performed similarly at all data collection points, with the exception of the high performing subgroup who again showed this trend of continued improvement, especially so in the Socialization domain. This unexpectedly positive result may simply be due to chance or indicate that the participants in the high responding subgroup were going through a period of improvement due to other activities and interventions outside of the Social Tutor, however given that these trends were not observed in control group participants it certainly appears that use of the Social Tutor software did directly contribute. This suggests that not only were participants benefitting during the active software use period, but they were able to apply the knowledge and skills they learned using the Social Tutor to situations outside of this, practicing what they had learned, continuing to build on their knowledge and, for high responding participants, even translating this into real-world behavioural changes.

9.2 Study Quality, Limitations and Recommendations

The evaluation presented in this thesis aimed to address several of the limitations of existing studies identified by Rao et al. (2008) and Neely et al. (2016). In particular, care was taken during software development to specifically support and encourage the generalisation of knowledge gained through the Social Tutor to other contexts, and the evaluation itself specifically measured maintenance of knowledge and behaviours twice following the intervention period, at both two and four months after the intervention period ended. Additionally, the evaluation included a control group who used the same software interface but were provided with non-social content so that changes in knowledge or behaviour that occurred over the intervention period could be appropriately attributed to the explicit social skills teaching provided by the Social Tutor. A matched-pairs procedure was used to allocate participants into these two groups, and analysis of the data indicated that the experimental and control groups were successfully balanced with no statistically significant differences in age, gender, or socio-economic status between the two groups. Further, analysis of the Vineland-II results indicated that all participants recruited had an adaptive behaviour level appropriate to inclusion in the study.

Despite these strengths in the evaluation, there were also a number of challenges. As is common to most evaluations of this nature, a number of participants withdrew in the later stages of the evaluation or failed to complete some of the tasks at some of the data points. This was accounted for as much as possible within the data analysis. There were also a number of technical difficulties throughout the evaluation period that had various impacts on the study. In relation to collecting longitudinal content quiz data, a small number of participants had to resort to completing their answers via an electronic Microsoft Word document. While this was preferable to no data being collected, it did mean the method was inconsistent across participants at the two and four month follow up points, and that for the affected participants accuracy and duration data could not be obtained. This did not impact any of the other measures used in the study, and did not impact the pre-test or immediate post-test data.

In relation to technical issues, due to variation between individual families' home computers, a few minor adjustments to the software occasionally had to be made on-site at the installation visit. Specifically, a small number of home computers were not powerful enough to run the Social Tutor optimally, resulting in the virtual characters' idle motion having to be disabled or the software running more slowly than desired. Some other sporadic technical difficulties were also encountered by a small number of participants. Together this means that participants did not all receive a completely consistent experience with the software, however effort was made to ensure consistency and address such technical difficulties quickly as much as possible.

Another limitation of the current study is that of sample size. While thirty one participants completed the evaluation, approximately half were in the control group and half in the experimental group. Power calculations indicate that to achieve an effect size of 0.5 with = 0.05 and 80% power, a minimum of 27 participants per group would be ideal. Once the existing groups are further broken down for fine-grained exploratory analysis, it is difficult to claim sufficient statistical power to draw meaningful conclusions

beyond observing trends. Combined with these small sample sizes, the large number of analyses performed on the data also raises the issue of potential false positives. Holm-Bonferroni correction for multiple comparisons has been applied to the primary analyses that directly address the study's objectives to assist in addressing this issue, but has not been applied to analyses that are exploratory in nature only (Holm 1979). Thus, an expanded evaluation with a wider audience would be ideal to allow for more investigation into the interesting trends identified.

Given the necessity for the researcher to physically visit the family home or school to install the software, a larger sample size is challenging to achieve in the current environment. One possible avenue for both ensuring increased consistency and addressing the issue of sample size may be to port the Social Tutor to a mobile device platform, such as making it iPad or Android compatible. The Social Tutor in its current iteration is also reliant on a number of third-party libraries and proprietary components, such as the virtual character voices, which limits its ability to be disseminated. Resolving this issue by making it fully self-contained or addressing the related licensing issues would mean that families could download and install it themselves without a researcher visit being necessary, and given that all data collection was done electronically, this would mean that participant recruitment would no longer be limited to the physical location surrounding the research team. It would also potentially enable the software to be made more widely available and allow its benefits to be enjoyed by families outside of the current study. As discussed in Chapter 8 there are a number of other potential benefits to porting to a mobile device, for example being able to make use of the touch screen and physical portability, making this a particularly enticing pathway for future development.

The current evaluation was also relatively short in duration, with the active software use period being only three weeks. There are a number of indicators suggesting that this was not long enough, for example that no participant unlocked the third and final topic, and that some initial upwards trends were observed in the Vineland-II results but not enough data was obtained to know whether these were genuine trends or simply due to chance. Expanding the evaluation period, for example to six weeks, may assist in addressing these issues. A longer evaluation with more participants may also assist with identifying which factors most impact or predict participant outcomes.

In terms of the measures used, the content quiz was identified as being too limited in its possible answers. For the purpose of accessibility the content quiz activities were designed to be straightforward, however it appears that the side of simplicity has been erred on too much. Adding more distractor nodes, options, and increased flexibility may help to ensure the content quiz is a better reflection of participants' underlying knowledge. The content quiz was also observed to be insufficiently engaging and it took longer to complete for many participants than desired. Further, the same content quiz was presented at all four data collection points, albeit with the questions in random order, making it possible that test scores may be influenced by familiarity with the questions themselves. Thus, further refinement of this measure is recommended. Given that the content quiz was delivered directly via the Social Tutor and that it was the first exposure that

participants had to the software, it is possible that it may have inadvertently negatively impacted on their perceptions of the software as a whole before they even begun. It may therefore be advisable to present the content quiz in a format that is perceived as separate from the Social Tutor software. The Google Form approach used for the Vineland-II was reliable and consistent so this may present one possible option, although it does not lend itself to as much interactivity and flexibility in activity type as the current content quiz implementation.

The Vineland-II was used to measure behavioural changes, however it should be noted that only caregivers were asked to complete it. Given that the participants in the current study all attend mainstream school, to get a better overall picture of their social behaviours in context it may be advisable in future research to ask that participants' teachers also complete a behavioural assessment of participants at each data collection point. As discussed in Subsection 7.3.1, there were a number of limitations with the Vineland-II itself, thus for future research an alternative standardised measure of behaviour may be more suitable.

Finally, there were some difficulties regarding the content presented in the Social Tutor software itself. As discussed at length in Subsection 2.7.3, it was difficult to identify and obtain permission to use existing curricula that were evidence-based, validated for use with children on the autism spectrum, and that could also be adapted to a software context with high fidelity. However, developing a social skills curriculum in itself is a substantial undertaking that was deemed to be beyond the scope of this thesis. In future research the ideal situation would be to collaborate with curriculum designers and develop content for both the software context and the school or therapist-driven context simultaneously. This way the content addressed in the Social Tutor can be designed to best make use of the strengths of the integrated technology, while ensuring that it also fully complements and consolidates the content that learners are experiencing in their other intervention environments. Another content-related possibility for future development would be to expand the content to different subject areas, for example in the recent survey by Parsons et al. (2016) individuals with autism and those that support them indicated a strong interested in technology to help with academic skills too, not only social and daily living skills. Given that there is much existing research into using pedagogical agents for teaching academic skills to neurotypical children, this may provide an interesting synergy and a valuable avenue for future research.

9.3 Closing Statement

This research aimed to develop software for teaching basic social skills to children with autism using autonomous virtual humans, and then evaluated that software for its efficacy. The evaluation of the Social Tutor revealed that participants were able to improve their knowledge of social skills through use of the software, and that these gains were maintained up to four months after the period of software use ended. Results of the behavioural assessment showed that for a small subset of high responding participants only, these improvements in knowledge also translated into behavioural changes. While more needs to be done to support other learners to display this same generalisation from the software context to real-world scenarios,

the results are nonetheless very encouraging given the difficulties with generalisation that are a known challenge for any interventions targeted at individuals on the autism spectrum.

The development of the Social Tutor and its subsequent evaluation highlight the potential of virtual humans for improving the skills of children with autism in an engaging and judgement-free environment, and provide valuable insights into the challenges, strengths, and future opportunities of this unique and exciting approach.

GLOSSARY

Term	Definition
Correctness	A numerical value between 0.0 and 1.0 indicating whether the student has met all of an activity's requirements. It refers to the final state of the activity only and is calculated differently depending on the activity type.
Accuracy	A numerical value between 0.0 and 1.0 indicating how many mistakes the learner has made while completing an activity. To calculate accuracy, a tally is kept of both the total number of moves and the number of incorrect moves that the student makes during their interaction with an activity, accuracy is then calculated as: `(total_moves – incorrect_moves)/total_moves`
Duration	A numerical value calculated as the difference in total running time between the timestamp of the question at hand and the timestamp of the previous entry, and is presented as the time taken to complete that question in seconds.
High responder	Experimental group participant who displayed an overall improvement in correctness score of 10% or higher from pre-test to immediate post-test. High responders typically used the software for less time and completed fewer lessons than their peers.
Average responder	Experimental group participant who displayed an overall improvement in correctness score between 2% and 10% from pre-test to immediate post-test.
Low responder	Experimental group participant who displayed an overall improvement in correctness score of less than 2% from pre-test to immediate post-test. Low responders typically completed more lessons than their peers.

APPENDICES

Appendix A. Content Quiz Questions and Expected Answers

The content quiz consists of twelve short activities which are presented to the participant in random order. These are designed to span the three topics covered in the software:

- Greeting Others (Questions 1-4)
- Listening and Turn Taking (Questions 5-8)
- Starting and Ending Conversations (Questions 9-12)

While a title is displayed in the screenshots for convenience (e.g. "G1: Asking Someone's Name") that panel is disabled and no title shown to participants. Every activity page provides the following function buttons:

- **Instructions** - hear the original instructions (supplied as 'virtual tutor script' here) again
- **Read Aloud** - when activated, the user can click any button, box or other text-containing element and the virtual tutor will read the text content out loud.
- **Reset** - returns all interactive components back to their original locations and states

These buttons and their functionality are explained to the user prior to the content quiz starting.

All assessment activities are completed by dragging components into their correct locations, or clicking buttons to make a selection. None require typing text or recording audio-visual information.

The remainder of this appendix contains the following information about each assessment activity:

1. The 'virtual tutor script' spoken by the virtual teacher when the question is displayed
2. Screenshots
    a. If one screenshot, it shows the question with the desired answer state displayed
    b. If two screenshots, the first shows the original state, and the second shows the desired answer state

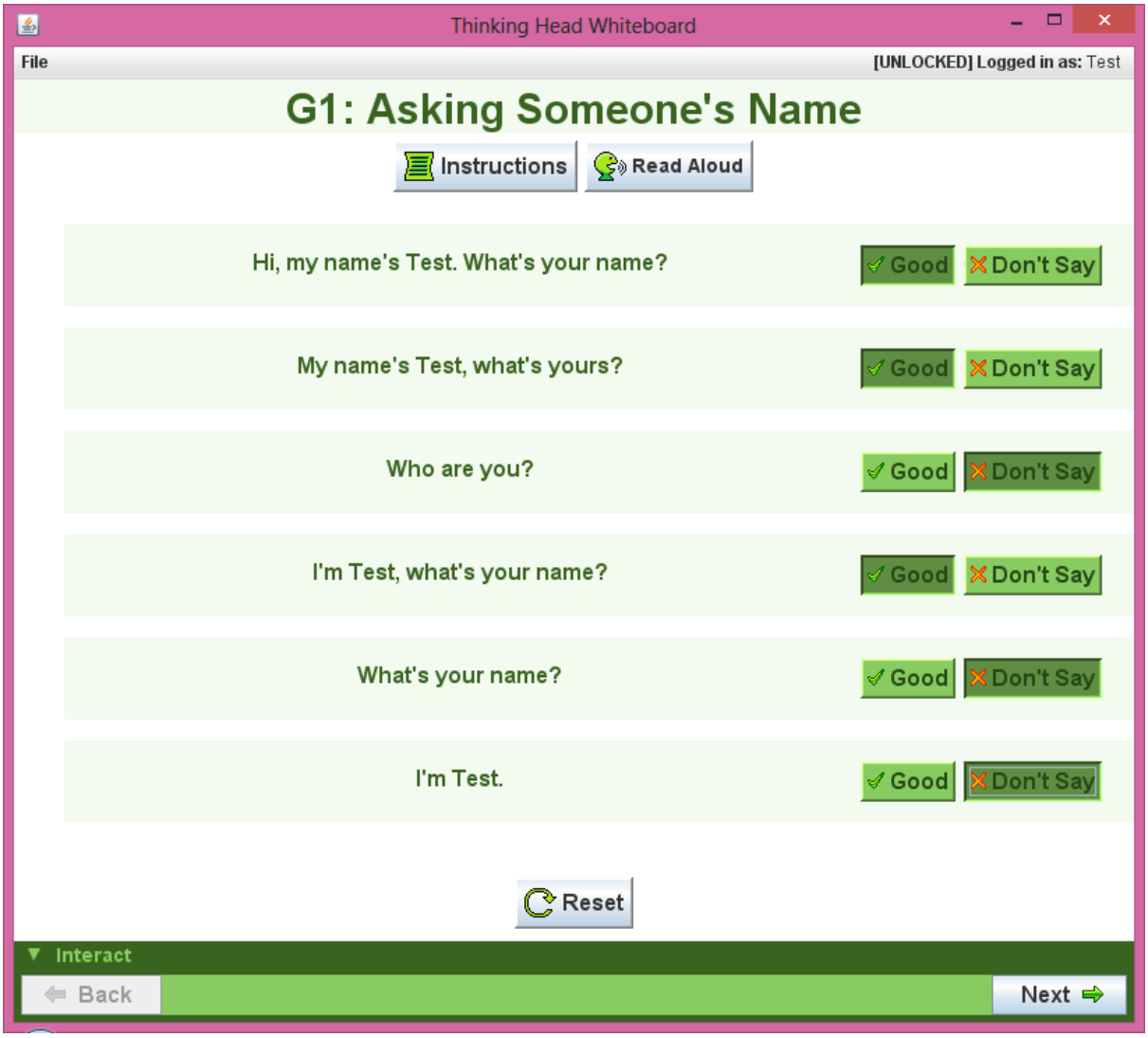
Task 1

Virtual tutor script:

This question is about asking someone what their name is.

Choose 'Good' next to good ways to ask someone their name.

Choose 'Don't Say' next to things you shouldn't say.



Note: *In these questions the user's own name will be displayed where 'Test' is used as a placeholder above.*

Task 2

Virtual tutor script:

This question is about when it's OK to greet someone.

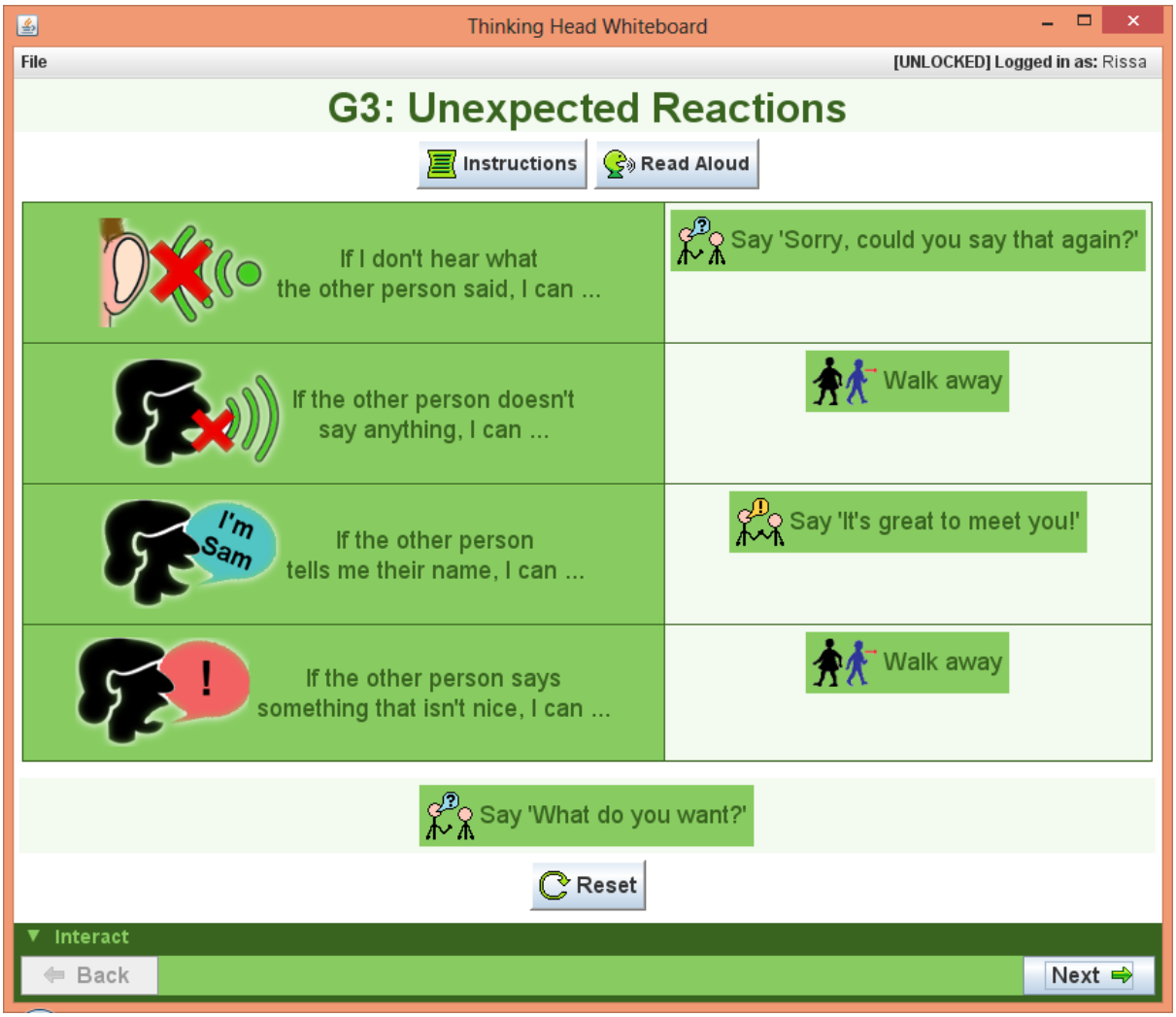
Match each scenario with its answer.

Task 3
Virtual tutor script:

Imagine you just asked someone. What is your name?

Here are some things that could happen next

Match each scenario with a good response.



Task 4
Virtual tutor script:

This question is about how to greet someone.

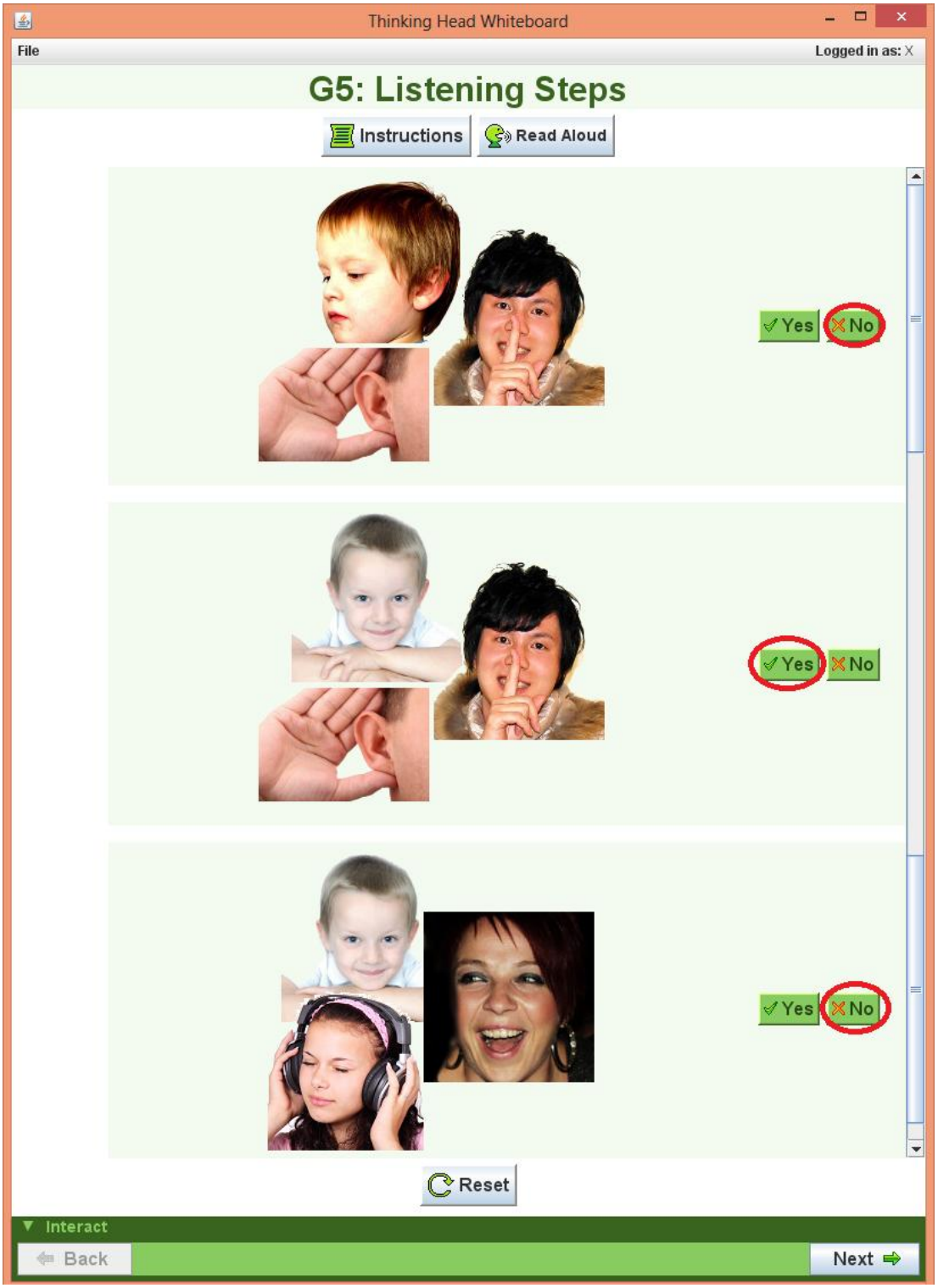
Click, YES next to the picture showing the right steps. Click, NO next to the others.

Task 5

Virtual tutor script:

This question is about listening to other people.

Choose YES, next to the picture showing the steps of good listening. Choose NO, next to the others.



Task 6
Virtual tutor script:

This question is about taking turns.

Click True, next to reasons why taking turns is important.

Click False, next to things that aren't correct.

Task 7

Virtual tutor script:

This question is about knowing when to stop talking

Here are some questions you might use to assess a conversation.

Match each question with one "Yes" and one "No" response.

Task 8

Virtual tutor script:

This question is about knowing when it's your turn to talk.

Here are some questions you might use to assess a conversation.

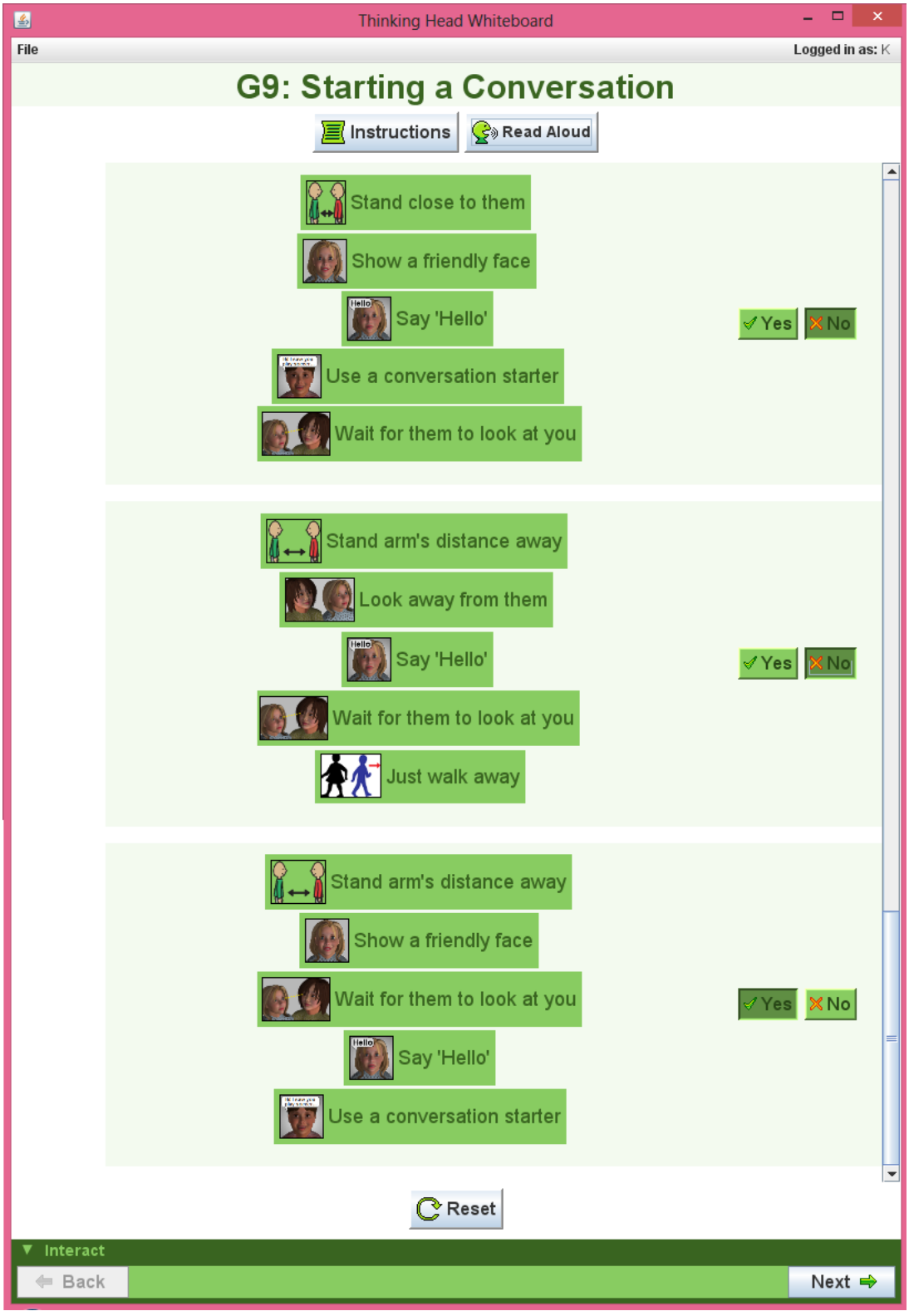
Match each question with one "Yes" and one "No" response.

Task 9

Virtual tutor script:

This question is about starting a conversation.

Choose YES, next to the picture showing the right steps. Choose NO, next to the others.



Task 10

Virtual tutor script:

This question is about starting conversations.

Organise the conversation starters into the right boxes.

Task 11

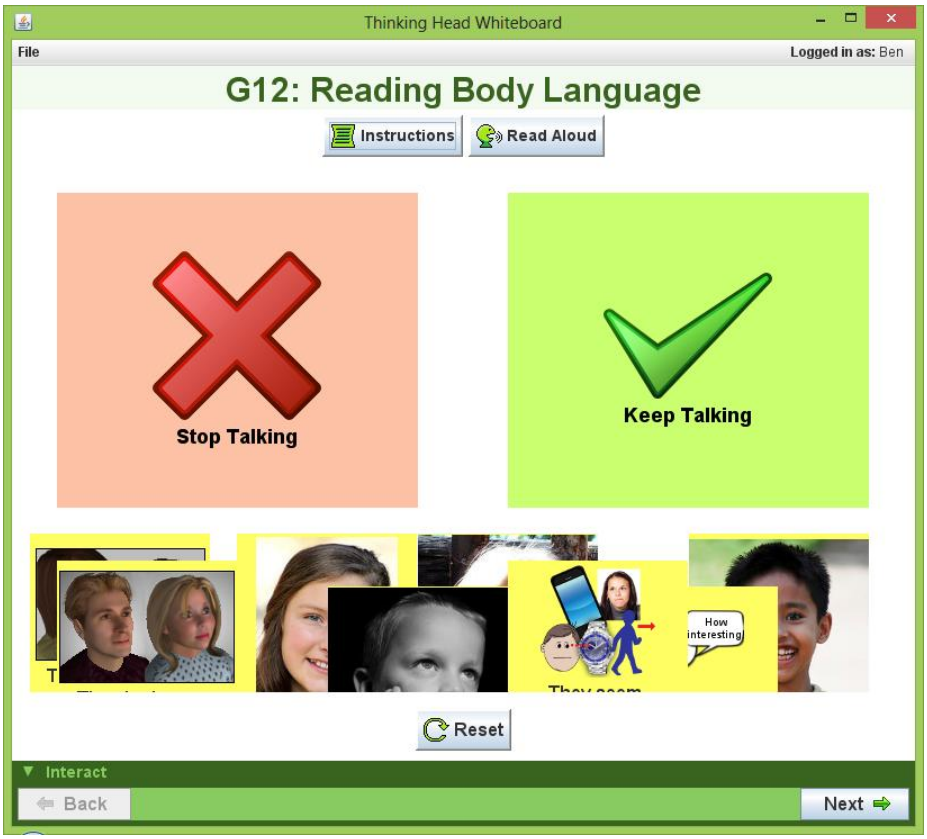
Virtual tutor script:

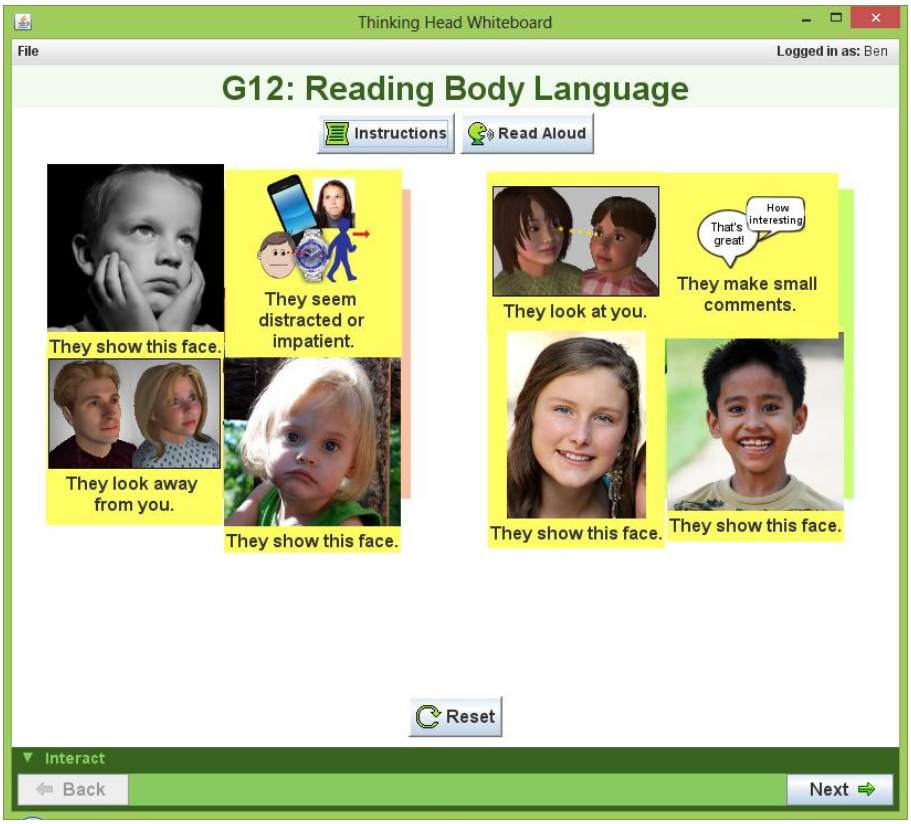
This question is about ending a conversation. Finish this table by adding the rest of the steps.

Task 12
Virtual tutor script:

This question is about using body language to tell when to stop talking.

Sort these body language pictures into the 'stop talking' and 'keep talking' boxes.





Appendix B. Pre-test Questionnaire

Multiple Choice Questions

Tick the box that best describes you:

I use a computer…	Every day	Several times a week	Once a week	A few times a month or less	Never

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
I learn how to do new things easily					
I am good at using computers					
I enjoy using computers					
The talking head software will help me to learn					
I learn how to use new software quickly					
The talking head software will be fun to use					

Open Ended Questions

1. **What do you normally use computers for?**

   e.g. Homework, playing games, browsing the internet, email, social networking (Facebook)

2. **What do you think it will be like using the Talking Head software?**

3. **Additional Comments:**

   Please write anything else you'd like to tell us about here.

Appendix C. Post-test Questionnaire

Multiple Choice Questions:

Tick the box that best describes you:

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
It was easy [for my child] to choose the activities [I/they] wanted to do					
It was easy to see how well [I/my child] was going with [my/their] activities					
It was easy [for my child] to understand what [I/they] needed to do in activities					
The virtual human(s) were friendly					
[I/my child] could clearly hear and understand what the virtual human(s) said					
The activities in the software were fun [for my child]					
Interacting with the virtual human(s) was fun [for my child]					
The topics covered in the software were useful to [me/my child]					
The software helped [me/my child] to learn					

The activities in the software were...	*Very easy*	*A bit easy*	*Just right*	*A bit hard*	*Very hard*

The amount of time spent on each topic was…	*Very short*	*A bit short*	*Just right*	*A bit long*	*Very long*

Open Ended Questions

1. **What did you like best about the software?**
2. **What did you like least about the software?**
3. **Was there anything [you/your child] found difficult to do or understand in the software?**
4. **Would you change anything about the software? If yes, what?**
   If you have any ideas to make the virtual tutor better, please write them here!
5. **Additional Comments:**
   Please write anything you'd like to tell us about your experiences with the software here.

Appendix D. Statistical Formulae and Conventions

Box plots were generated using SPSS and conform to standard SPSS box plot conventions, with the box extending to the 25th and 75th percentiles, the whiskers extending to 1.5 IQR of the lower and upper quartiles, and results beyond these represented as individual points. On all box plots the median is indicated with a solid line, and where appropriate the mean is additionally indicated with a dotted line. Given the relatively small sample sizes providing both the mean and median is intended to provide a better overall picture of the data. Other graphs were plotted using Microsoft Office Excel 2007 or SPSS as appropriate.

Unless otherwise noted, paired t-tests are accompanied by Cohen's d as the effect size measure, calculated as $d = \frac{t}{\sqrt{n}}$ where the t-value is obtained from SPSS directly and n is the number of subjects (Lakens 2013). Unpaired t-tests, due to the small sample sizes involved, are accompanied by Hedges' g as the effect size measure, which is calculated as $g \cong d\left(1 - \frac{3}{4(n_1+n_2)-9}\right)$ where Cohen's d is obtained using the previous formula and n_1 and n_2 are the number of subjects in each group. Where Wilcoxon signed rank or Wilcoxon rank sum tests are used, a lower limit effect size estimate is calculated as $r = \frac{Z}{\sqrt{N}}$ where the Z-value is obtained directly from SPSS and N refers to the total number of observations in the dataset (Rosenthal et al. 1994). For ANOVA, partial eta-squared is reported for effect size and calculated directly in SPSS, while for Chi-squared tests Cramer's V is the reported effect size measure and is calculated as $V = \sqrt{\frac{\chi^2}{n \cdot df*}}$ where n is the number of observations and degrees of freedom is calculated as $df^* = min(r - 1, c - 1)$ where r and c refer to the number of rows and columns in the contingency table.

Appendix E. Participant Completion of Data Collection Tasks

	Content Quiz				Vineland-II				Questionnaires		
Group	**Pre-test**	**Post-test #1**	**Post-test #2**	**Post-test #3**	**Pre-test**	**Post-test #1**	**Post-test #2**	**Post-test #3**	**Pre-test**	**Post Child**	**Post Parent**
E	X	X	X	X	X	X	X	X	X	X	X
E	X	X	X	X	X	X	X	X	X	X	X
E	X	X	X	X	X	X	X	X	X	X	X
E	X	X	X	X	X	X	X	X	X	X	X
E	X	X	X	X	X	X	X	X	X	X	X
E	X	X	X	written	X	X	X	X	X	X	X
E	X	X	X	written	X	X	X	X	X	X	X
C	X	X	X	X	X	X	X	X	X	X	X
C	X	X	X	X	X	X	X	X	X	X	X
C	X	X	X	X	X	X	X	X	X	X	X
C	X	X	X	X	X	X	X	X	X	X	X
C	X	X	X	X	X	X	X	X	X	X	X
C	X	X	X	X	X	X	X	X	X	X	sibling
C	X	X	X	X	X	X	X	X	X	X	X
C	X	X	written	written	X	X	X	X	X	X	X
E	X	X	written	X	X	X	X	X	X	-	-
E	X	X	X	X	X	X	X	X	X	-	X
E	X	X	X	X	X	X	X	X	X	-	-
E	X	X	X	X	X	X	-	X	X	X	X
E	X	X	X	X	X	X	-	X	X	X	sibling
E	X	X	-	X	X	X	X	X	X	X	-
E	X	X	-	X	X	X	X	X	X	X	X
E	X	X	-	X	X	X	-	X	X	X	-
E	X	X	-	-	X	X	-	-	X	X	X
C	X	X	X	X	X	X	X	X	X	-	X
C	X	X	X	X	X	-	X	X	X	-	-
C	X	X	-	written	X	X	X	X	X	X	sibling
C	X	X	-	-	X	X	-	-	X	-	-
C	X	-	-	written	X	X	-	X	X	X	X
C	X	-	X	X	X	X	-	X	X	-	-
C	X	-	-	-	X	X	-	-	X	X	-
Count	31	28	23	28	31	30	23	28	31	24	20
Percent	100.0%	90.3%	74.2%	90.3%	100.0%	96.8%	74.2%	90.3%	100.0%	77.4%	64.5%

Notes: 'X' indicates data was successfully collected at this point for the specified participant. 'Written' indicates that the content quiz at this data collection point was administered via electronic Microsoft Word document rather than via the Social Tutoring software. 'Sibling' indicates that this participants' caregiver completed the post-test questionnaire once under their sibling's identifier.

Complete data for all data collection tasks were obtained for participants above the dotted line. Participants below the dotted line have partial data sets for some or all tasks.

Appendix F. Additional Demographic Analyses

Balance of Intervention Group Socio-economic Status

Participant socio-economic status was not controlled for at recruitment or group allocation, however for completeness this factor was assessed to ensure no unintended imbalances occurred during the participant group allocation process. The measure used as an indicator of this was percentile ranking of participants' home postcodes as obtained from the Australian Bureau of Statistics Socio-economic Indexes for Areas (SEIFA) data (Australian Bureau of Statistics 2011). The mean percentile for the control group (M = 50.93, SD = 25.66, 95% CI [36.73, 65.14]) was higher than the experimental group (M = 46.50, SD = 23.82, 95% CI [33.81, 59.19]) but the difference was found to be non-significant by t-test ($p = 0.622$, $d = 0.09$).

Data Analysis by Socio-economic Status

For completeness a comparison of content quiz scores and Vineland-II scores based on participant socio-economic status was conducted for both the experimental group and the control group. For the content quiz, Spearman's rank-order correlations were run between socio-economic status and correctness, accuracy and duration data. A moderately strong positive relationship between socio-economic status and correctness reached marginal significance ($r_s = 0.480$, $p = 0.060$) for the experimental group alone, with no other significant relationships identified for either group. For the Vineland-II data Spearman's rank-order correlation tests revealed no significant relationships overall or for any domain or subdomain.

Appendix G. Primary Analysis Data

The Benjamini–Hochberg procedure has been conducted on the values below to correct for the false discovery rate due to multiple comparisons.

In the tables below "Pre" refers to the pre-test, "P1" to the immediate post-test, "P2" to the two month follow up post-test, and "P3" to the four month follow-up post-test.

Table 42: Significance values for experimental group whole-quiz correctness data

	Effect Size	Raw p-value	Threshold p-value
Pre to P3	-1.03	0.004*	0.010
Pre to P1	-0.73	0.010*	0.013
Pre to P2	-0.55	0.051	0.017
P1 to P3	-0.63	0.052	0.025
P1 to P2	0.03	0.900	0.050

* indicates corrected p-value is significant after applying FDR correction procedure.

Table 43: Significance values for control group whole-quiz correctness data

	Effect Size	Raw p-value	Threshold p-value
P1 to P3	-0.28	0.3473	0.0100
Pre to P2	0.23	0.4689	0.0125
P1 to P2	0.14	0.6421	0.0167
Pre to P3	-0.08	0.7770	0.0250
Pre to P1	0.08	0.7989	0.0500

Appendix H. Detailed Content Quiz Correctness Tables

Table 44: Comparison of pre-test and post-test content quiz correctness scores

		Experimental group N = 16	High responders N = 4	Average responders N = 8	Low responders N = 4	Control group N = 12
All questions	**M (SD)**	7.36% (9.05)	20.56% (8.98)	4.21% (1.47)	0.48% (1.38)	1.14% (7.16)
	95% CI	2.54, 12.19	6.28, 34.85	2.98, 5.44	-1.72, 2.68	-3.41, 5.69
	Wilcoxon signed rank	p = 0.001* r = -0.60	p = 0.068 . r = -0.65	p = 0.012* r = -0.63	p = 0.285 r = -0.38	p = 0.530 r = -0.13
Topic 1 & 2 (questions 1-8)	**M (SD)**	7.65% (10.77)	21.97% (9.68)	5.80% (4.27)	-2.97% (3.27)	1.05% (7.62)
	95% CI	1.91, 13.39	6.57, 37.37	2.23, 9.36	-8.17, 2.24	-3.79, 5.90
	Wilcoxon signed rank	p = 0.013* r = -0.44	p = 0.068 . r = -0.65	p = 0.017* r = -0.60	p = 0.068 . r = -0.65	p = 0.480 r = -0.14
Topic 1 (questions 1-4)	**M (SD)**	9.67% (16.01)	32.38% (7.15)	2.72% (10.77)	0.88% (6.42)	1.50% (14.10)
	95% CI	1.14, 18.2	21.00, 43.75	-6.28, 11.72	-9.34, 11.09	-7.46, 10.46
	Wilcoxon signed rank	p = 0.041* r = -0.36	p = 0.068 . r = -0.65	p = 0.499 r = 0.17	p = 0.715 r = -0.13	p = 0.677 r = -0.08
Topic 2 (questions 5-8)	**M (SD)**	5.63% (14.99)	11.56% (16.20)	8.88% (15.73)	-6.81% (1.68)	0.60% (8.64)
	95% CI	-2.36, 13.61	-14.21, 37.34	-4.28, 22.03	-9.48, -4.15	-4.88, 6.09
	Wilcoxon signed rank	p = 0.233 r = -0.21	p = 0.144 r = -0.52	p = 0.176 r = -0.34	p = 0.068 . r = -0.65	p = 0.480 r = -0.14
Topic 3 (questions 9-12)	**M (SD)**	6.80% (14.64)	17.75% (21.59)	1.03% (10.66)	7.38% (9.52)	1.31% (12.76)
	95% CI	-1.00, 14.60	-16.61, 52.11	-7.88, 9.94	-7.77, 22.52	-6.80, 9.42
	Wilcoxon signed rank	p = 0.140 r = -0.26	p = 0.109 r = -0.57	p = 0.866 r = -0.04	p = 0.144 r = -0.52	p = 0.767 r = -0.06

** denotes statistical significance ($p < 0.05$), . denotes marginal significance ($0.05 < p < 0.1$)*

Appendix I. Detailed Content Quiz Duration Tables

Table 45: Comparison of change in content quiz question answering duration by topic

		Experimental group N = 16	High responders N = 4	Average responders N = 8	Low responders N = 4	Control group N = 12
Topic 1 & 2 (questions 1-8)	**M (SD)**	**-26.03 (27.96)**	**2.72 (6.22)**	**-26.16 (17.14)**	**-54.53 (31.62)**	**-7.30 (55.40)**
	95% CI	-40.93, -11.13	-7.17, 12.61	-40.49, -11.83	-104.85, -4.21	-42.50, 27.90
	Wilcoxon signed rank	p = 0.023* r = -0.40	p = 0.273 r = -0.55	p = 0.012* r = -0.89	p = 0.068 . r = -0.91	p = 0.347 r = -0.19
Topic 1 'Greeting' (questions 1-4)	**M (SD)**	**-18.68 (26.93)**	**4.38 (23.43)**	**-20.44 (20.49)**	**-37.75 (29.91)**	**-14.56 (28.06)**
	95% CI	-32.91, -4.21	-32.90, 41.65	-37.57, -3.30	-85.34, 9.84	-32.39, 3.27
	Wilcoxon signed rank	p = 0.056 . r = -0.34	p = 0.715 r = -0.13	p = 0.017* r = -0.60	p = 0.068 . r = -0.91	p = 0.084 . r = -0.35
Topic 2 'Listening and Turn Taking' (questions 5-8)	**M (SD)**	**-33.50 (36.39)**	**1.06 (15.58)**	**-31.88 (19.37)**	**-71.32 (44.70)**	**-0.04 (100.87)**
	95% CI	-52.89, -14.11	-23.73, 25.86	-48.07, -15.68	-142.44, -0.18	-64.13, 64.05
	Wilcoxon signed rank	p = 0.020* r = -0.41	p = 1.00 r = 0.0	p = 0.017* r = -0.60	p = 0.068 . r = -0.91	p = 0.638 r = -0.10
Topic 3 'Good Conversations' (questions 9-12)	**M (SD)**	**-31.90 (27.56)**	**-9.31 (19.86)**	**-41.43 (25.84)**	**-35.44 (30.08)**	**-26.56 (36.83)**
	95% CI	-46.59, -17.22	-40.91, 22.28	-63.02, -19.83	-83.29, 12.42	-49.96, -3.16
	Wilcoxon signed rank	p = 0.023* r = -0.40	p = 0.461 r = -0.26	p = 0.012* r = -0.63	p = 0.144 r = -0.52	p = 0.023* r = -0.46

** denotes statistical significance (p < 0.05), . denotes marginal significance (0.05 < p < 0.1)*

Table 46: Change in content quiz duration in seconds from pre- to post-test by questions and group

			Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
Experimental	**Pre-test**		62.69	84.13	97.94	49.63	48.25	93.88	124.56	150.25	83.31	120.44	86.06	83.13
	Post-test		42.94	64.81	72.25	40.13	39.44	63.81	86.06	93.63	49.25	83.06	64.13	48.88
	Change		**-19.75**	**-19.31**	**-25.69**	**-9.50**	**-8.81**	**-30.06**	**-38.50**	**-56.63**	**-34.06**	**-37.38**	**-21.94**	**-34.25**
	Wilcoxon signed rank	p	0.009*	0.079 .	0.015*	0.379	0.301	0.008*	0.046*	0.006*	0.002*	0.008*	0.026*	0.002*
		r	-0.46	-0.31	-0.43	-0.16	-0.18	-0.47	-0.35	-0.48	-0.54	-0.47	-0.39	-0.54
Control	**Pre-test**		60.75	86.42	96.42	57.42	48.83	83.67	141.42	168.42	71.58	116.92	102.67	82.00
	Post-test		46.92	71.75	90.75	33.33	115.75	70.75	165.75	89.92	47.75	97.33	60.92	60.92
	Change		**-13.83**	**-14.67**	**-5.67**	**-24.08**	**66.92**	**-12.92**	**24.33**	**-78.50**	**-23.83**	**-19.58**	**-41.75**	**-21.08**
	Wilcoxon signed rank	p	0.182	0.209	0.530	0.018*	0.722	0.182	0.638	0.012*	0.136	0.182	0.034*	0.050*
		r	-0.27	-0.26	-0.13	-0.48	-0.07	-0.27	-0.10	-0.51	-0.30	-0.27	-0.43	-0.40

** denotes statistical significance (p < 0.05), . denotes marginal significance (0.05 < p < 0.1)*

Table 47: Change in content quiz duration from pre-test to post-test by question and response subgroups

			Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12
High responders	**Pre-test**		72.50	60.75	65.75	32.25	45.75	52.00	85.25	75.75	53.25	85.75	52.00	64.00
	Post-test		49.25	80.75	66.50	52.25	32.25	66.50	87.75	76.50	48.00	66.00	59.50	44.25
	Change		**-23.25**	**20.00**	**0.75**	**20.00**	**-13.50**	**14.50**	**2.50**	**0.75**	**-5.25**	**-19.75**	**7.50**	**-19.75**
	Wilcoxon signed rank	p	0.141	0.465	1.000	0.465	0.465	0.357	0.715	1.000	0.144	0.465	0.465	0.273
		r	-0.52	-0.26	0.00	-0.26	-0.26	-0.33	-0.13	0.00	-0.52	-0.26	-0.26	-0.39
Average responders	**Pre-test**		56.25	86.00	107.25	56.25	44.38	91.50	134.50	155.38	90.38	120.50	97.38	87.63
	Post-test		39.75	64.38	78.25	41.63	46.88	58.50	87.00	105.88	47.13	71.25	60.63	51.13
	Change		**-16.50**	**-21.63**	**-29.00**	**-14.63**	**2.50**	**-33.00**	**-47.50**	**-49.50**	**-43.25**	**-49.25**	**-36.75**	**-36.50**
	Wilcoxon signed rank	p	0.092 .	0.050*	0.050*	1.000	0.889	0.018*	0.093 .	0.050*	0.025*	0.012*	0.036*	0.025*
		r	-0.42	-0.49	-0.49	0.00	-0.04	-0.59	-0.42	-0.49	-0.56	-0.63	-0.53	-0.56
Low responders	**Pre-test**		65.75	103.75	111.50	53.75	58.50	140.50	144.00	214.50	99.25	155.00	97.50	93.25
	Post-test		43.00	49.75	66.00	25.00	31.75	71.75	82.50	86.25	54.75	123.75	75.75	49.00
	Change		**-22.75**	**-54.00**	**-45.50**	**-28.75**	**-26.75**	**-68.75**	**-61.50**	**-128.25**	**-44.50**	**-31.25**	**-21.75**	**-44.25**
	Wilcoxon signed rank	p	0.197	0.144	0.068 .	0.068 .	0.144	0.068 .	0.068 .	0.068 .	0.068 .	0.465	0.144	0.068 .
		r	-0.46	-0.52	-0.65	-0.65	-0.52	-0.65	-0.65	-0.65	-0.65	-0.26	-0.52	-0.65

** denotes statistical significance ($p < 0.05$), . denotes marginal significance ($0.05 < p < 0.1$)*

Note: all values are indicated in seconds

Appendix J. Pre-Identified Vineland-II Items

The following table provides a list of Vineland-II items identified as most likely to improve after use of the Social Tutor software. Grey text indicates less of a focus, but still potentially interesting due to tangential connection to material being taught. Normal text is directly relevant and taught in the software, while background highlighted text is particularly relevant, with material being explicitly addressed and a key focus in the social tutoring software.

Domain	Subdomain	Item	Item Description
Communication	Speech Skills	18	Understands sayings that are not meant to be taken word for word (for example, "button your lip"; "hit the road"; etc.)
Communication	Speech Skills	41	Modulates tone of voice, volume, and rhythm appropriately (for example, does not consistently speak too loudly, too softly, or in a monotone, etc...)
Communication	Speech Skills	42	Tells about experiences in detail (for example, who was involved, where activity took place, etc)
Communication	Speech Skills	43	Gives simple directions (for example, on how to play a game or how to make something)
Communication	Speech Skills	46	Easily moves from one topic to another in conversation
Communication	Speech Skills	47	Stays on topic in conversations, does not got off on tangents
Communication	Speech Skills	48	Explains ideas in more than one way (for example, "This was a good book. It was exciting and fun to read"; etc)
Communication	Speech Skills	49	Has conversations that last 10 minutes (e.g. relates experiences, contributes ideas, shares feelings, etc)
Communication	Speech Skills	53	Gives complex directions to others (e.g. to a distant location, a recipe with many ingredients or steps, etc)
Socialization	Communication/ Friendship	13	Uses actions to show happiness or concern for others (e.g. hugs, pats arm, holds hands, etc)
Socialization	Communication/ Friendship	14	Shows desire to please others (e.g. shares a snack or toy, tries to help even if not capable, etc)
Socialization	Communication/ Friendship	15	Demonstrates friendship-seeking behaviour with others the same age (e.g. says "Do you want to play?" or takes another child by the hand).
Socialization	Communication/ Friendship	17	Answers when familiar adults make small talk (for example, if asked, "How are you?" says "I'm fine", if told "You look nice" says "Thank you"
Socialization	Communication/ Friendship	19	Uses words to express own emotions (e.g. "I'm happy", "I'm scared")
Socialization	Communication/ Friendship	20	Has best friend or shows preference for certain friends (of other gender) over others
Socialization	Communication/ Friendship	22	Uses words to express happiness or concern for others (e.g. "Yeah! You won!"; "Are you alright?"; etc
Socialization	Communication/ Friendship	23	Acts when another person needs a helping hand (e.g. holds a door open, picks up dropped items, etc)
Socialization	Communication/ Friendship	24	Recognizes the likes and dislikes of others (e.g. says "Chow likes soccer"; "Susie doesn't eat pizza")

Socialization	Communication/ Friendship	25	Shows same level of emotion as others around him or her (e.g. does not downplay or overdramatize a situation, etc)
Socialization	Communication/ Friendship	26	Keeps comfortable distance between self and others in social situations (e.g. doesn't get too close to another person when talking, etc)
Socialization	Communication/ Friendship	27	Talks about shared interests (e.g. sports, TV shows, summer plans, etc)
Socialization	Communication/ Friendship	28	Starts small talk when meets people he or she knows (e.g. says "How are you?"; "What's up?"; etc)
Socialization	Communication/ Friendship	29	Meets with friends regularly
Socialization	Communication/ Friendship	30	Chooses not to say embarrassing or mean things or ask rude questions in public
Socialization	Communication/ Friendship	31	Places reasonable demands on friendship (e.g. doesn't expect to be a person's only friend, or have the friend always available, etc)
Socialization	Communication/ Friendship	32	Understands that others do not know his or her thoughts unless he or she says them
Socialization	Communication/ Friendship	33	Is careful when talking about personal things
Socialization	Communication/ Friendship	35	Demonstrates understanding of hints or indirect clues in conversation (e.g. knows that yawns may mean "I'm bored" or a quick change of subject may mean "I don't want to talk about that")
Socialization	Communication/ Friendship	36	Starts conversations by talking about things that interest others (e.g. says "Tyrone tells me you like computers"
Socialization	Playing/ Social Cues	5	Chooses to play with other children (e.g. does not stay on the edge of a group or avoid others)
Socialization	Playing/ Social Cues	6	Plays cooperatively with one or more children for up to 5 minutes
Socialization	Playing/ Social Cues	7	Plays cooperatively with more than one child for more than 5 minutes
Socialization	Playing/ Social Cues	9	Shares toys or possessions when asked
Socialization	Playing/ Social Cues	10	Plays with others with minimal supervision
Socialization	Playing/ Social Cues	12	Protects self by moving away from those who destroy things or cause injury (e.g. those who bite, hit, throw things, pull hair, etc)
Socialization	Playing/ Social Cues	14	Seeks out others for play or companionship (e.g. invites others home, goes to another's home, plays with others on the playground, etc)
Socialization	Playing/ Social Cues	15	Takes turns when asked while playing games or sports.
Socialization	Playing/ Social Cues	17	Shares toys or possessions without being asked
Socialization	Playing/ Social Cues	18	Follows rules in simple games (relay races, spelling bees, electronic games, etc)
Socialization	Playing/ Social Cues	19	Takes turns without being asked

Socialization	Playing/ Social Cues	22	Asks permission before using objects belonging to or being used by another
Socialization	Playing/ Social Cues	23	Refrains from entering a group when nonverbal cues indicate that he or she is not welcome
Socialization	Playing/ Social Cues	25	Shows good sportsmanship (that is, follows rules, is not overly aggressive, congratulates other team on winning, and doesn't get mad when losing)
Socialization	Coping Skills	2	Says "thank you" when given something
Socialization	Coping Skills	3	Changes behaviour depending on how well he or she knows another person (e.g. acts differently with family member than stranger, etc)
Socialization	Coping Skills	5	Says "please" when asking for something
Socialization	Coping Skills	6	Ends conversations appropriately (e.g. says "Good-bye"; "See you later"; etc)
Socialization	Coping Skills	9	Says that he or she is sorry for unintended mistakes (e.g. bumping into someone)
Socialization	Coping Skills	10	Chooses not to taunt, tease or bully
Socialization	Coping Skills	11	Acts appropriately when introduced to strangers (e.g. nods, smiles, shakes hands, greets them)
Socialization	Coping Skills	12	Changes voice level depending on location or situation (e.g. in a library, during movie or play, etc)
Socialization	Coping Skills	13	Says he or she is sorry after hurting another's feelings
Socialization	Coping Skills	15	Talks with others without interrupting or being rude
Socialization	Coping Skills	16	Accepts helpful suggestions or solutions from others
Socialization	Coping Skills	19	Says he or she is sorry after making unintentional mistakes or errors in judgement (e.g. unintentionally leaving someone out of a game)
Socialization	Coping Skills	20	Shows understanding that gentle teasing with family or friends can be a form of humour or affection
Socialization	Coping Skills	23	Controls anger or hurt feelings when he or she does not get his or her own way (e.g. not allowed to watch TV or attend party, suggestion rejected by a friend or supervisor, etc)
Socialization	Coping Skills	26	Controls anger or hurt feelings due to constructive criticism (e.g. correction of misbehaviour, discussion of test score or grade, etc)
Maladaptive Behaviours	Internalizing	5	Refuses to go to school or work because of fear, feelings or rejection or isolation, etc
Maladaptive Behaviours	Internalizing	8	Has poor eye contact (that is, does not look at or face others when speaking or spoken to)
Maladaptive Behaviours	Internalizing	10	Avoids social interaction
Maladaptive Behaviours	Externalizing	4	Taunts, teases or bullies
Maladaptive Behaviours	Externalizing	5	Is inconsiderate or insensitive to others
Maladaptive Behaviours	Externalizing	9	Says embarrassing things or asks embarrassing questions in public (e.g. "You're fat" or "What's that big red things on your nose?")

Appendix K. Detailed Questionnaire Responses

Table 48: Summary of raw count and percentage pre-test questionnaire data

	I learn how to do new things easily.	I am good at using computers.	I enjoy using computers.	The Talking Head software will help me to learn.	I learn how to use new software quickly.	The Talking Head software will be fun to use.
Strongly Agree	3 (9.68%)	16 (51.61%)	22 (70.97%)	9 (29.03%)	8 (25.81%)	6 (19.35%)
Agree	10 (32.26%)	12 (38.71%)	5 (16.13%)	4 (12.90%)	10 (32.26%)	12 (38.71%)
Neither Agree nor Disagree	14 (45.16%)	2 (6.45%)	2 (6.45%)	17 (54.84%)	10 (32.26%)	11 (35.48%)
Disagree	4 (12.90%)	0 (0%)	2 (6.45%)	0 (0%)	3 (9.68%)	1 (3.23%)
Strongly Disagree	0 (0%)	1 (3.2%)	0 (0%)	1 (3.23%)	0 (0%)	1 (3.23%)

Table 49: Mean responses to post-test questionnaire Likert-style items

	Educational value				**Enjoyment**						**Usability**							
Group	Topics were useful		Helped me learn		ECAs were friendly		Activities were fun		Fun interacting with ECAs		Easy to choose activities		Easy to see progress		Easy to follow tasks		Voices were clear	
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Experimental Participants	3.69	*1.32*	3.38	*1.19*	3.77	*1.25*	3.00	*1.29*	2.92	*1.26*	3.23	*1.24*	3.38	*1.04*	3.54	*0.78*	3.69	*1.09*
Experimental Caregivers	4.08	*0.76*	3.92	*0.64*	3.85	*0.93*	3.31	*0.75*	3.62	*1.12*	3.92	*1.04*	3.46	*0.97*	3.46	*1.05*	3.77	*0.8*
Control Participants	3.00	*1.61*	2.91	*1.38*	4.27	*0.92*	4.36	*1.03*	3.64	*0.92*	4.45	*1.04*	3.91	*1.04*	4.18	*1.25*	4.36	*1.01*
Control Caregivers	3.33	*1.22*	3.44	*1.13*	4.33	*0.83*	3.89	*1.27*	3.67	*1.41*	4.67	*0.5*	3.89	*1.05*	3.78	*0.97*	4.22	*0.5*

1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree

Appendix L. Ethics Approvals

Flinders University Social and Behavioural Research Ethics Committee

Initial approval notice

FINAL APPROVAL NOTICE

Project No.:	5703
Project Title:	Autonomous virtual agents as social skills tutors for children with autistic spectrum disorders
Principal Researcher:	Miss Marissa Milne
Email:	marissa.milne@flinders.edu.au
Address:	School of Computer Science, Engineering and Mathematics

Approval Date:	24 August 2012	Ethics Approval Expiry Date:	**1 December 2013**

The above proposed project has been **approved** on the basis of the information contained in the application, its attachments and the information subsequently provided.

Most recent approval notice

MODIFICATION (No.4) APPROVAL NOTICE

Project No.:	5703
Project Title:	Autonomous virtual agents as social skills tutors for children with autistic spectrum disorders
Principal Researcher:	Miss Marissa Milne
Email:	marissa.milne@flinders.edu.au

Modification Approval Date:	19 December 2016	Ethics Approval Expiry Date:	**1 September 2017**

I am pleased to inform you that the modification request submitted for project 5703 on the 7 December 2016 has been reviewed and approved by the SBREC Chairperson. Please see below for a list of the approved modifications. Any additional information that may be required from you will be listed in the second table shown below called 'Additional Information Required'.

Autism SA Approval Notice

AUTISM SA

Info Line 1300 288 476

Registered Office: 1/303 Main Road, Torrensville | Mailing Address: PO Box 906, Melrose Park SA 5039
NT Address: 71 Coonawarra Road, Winnellie | NT Mailing Address: PO Box 36595, Winnellie NT 0821
t (SA) (08) 8379 6970 | t (NT) (08) 8947 4800 | f (08) 8339 1216 | admin@autismsa.org.au | www.autismsa.org.au

4 August 2016

Marissa Milne
marissa.milne@flinders.edu.au

Dear Marissa

Re: PP201611 – MILNE – ***Using virtual humans to teach greeting and conversation skills to children with autism in mainstream schooling.***

I am pleased to confirm that approval has been granted by the Chair of the Professional Practice Committee of Autism SA for support of the research project named above.

By utilising Autism SA services you are acknowledging the following:

- Contact details for researchers and supervisors that are provided to the Professional Practice Committee may be displayed on the Autism SA website and accessible by the public
- Current and past research projects will be listed on the Autism SA website indefinitely
- Autism SA requires one copy of the final report for inclusion in the Autism SA Resource Centre
- A brief summary report of the outcomes and findings is required for distribution via the Autism SA website, Infomail and / or printed newsletter
- Researchers may be requested by Autism SA to present their outcomes and findings at conferences and seminars hosted by Autism SA at no cost.

Yours sincerely

Louise Davies
Acting CEO

Appendix M. Recruitment Materials

Autism SA Advertisement

Would you like your child to improve their social skills using a virtual tutor?

At Flinders University we are developing a software based tutor for improving social skills in children with high functioning autism and Asperger's Syndrome. The virtual tutor models desirable social behaviours and guides learners through a variety of interactive activities. The software is being evaluated for its potential as a teaching tool.

Who can participate: Children who

- are aged 6-12 years
- have an existing diagnosis of high functioning autism or Asperger's Syndrome
- currently attend mainstream schooling

You will need access to a laptop or desktop computer with:

- Windows operating system installed (Windows 7 or higher preferred)
- an Internet connection

What is involved:

Participants will be asked to interact with the software for 10-15 minutes each day, 3-5 days per week for three weeks. The software can be installed on the participant's home or school computer (with consent from the school). Participants and their caregivers will be asked to complete a number of assessment tasks before and after using the software. Families will be able to keep a reduced version of the software at the completion of the study, and will be reimbursed $30 total for involvement.

How to register interest:

Email marissa.milne@flinders.edu.au or call 0455 486 279 for more details.

This research is being conducted by Marissa Milne as part of her PhD and supervised by Professor David Powers from the School of Computer Science, Engineering and Mathematics at Flinders University of South Australia.

Appendix N. Family Information Pack

Letter of Introduction

Ms Marissa Milne
PhD Student
School of Computer Science,
Engineering and Mathematics
GPO Box 2100
Adelaide SA 5001
Tel: 0455 486 279
marissa.milne@flinders.edu.au
www.csem.flinders.edu.au
CRICOS Provider No. 00114A

Dear Sir or Madam:

This letter is to introduce Ms Marissa Milne (marissa.milne@flinders.edu.au, 0455 486 279) who is a PhD student in the School of Computer Science, Engineering and Mathematics at Flinders University. She will produce her student card, which carries a photograph, as proof of identity.

Marissa is undertaking research leading to the production of a thesis or other publications on using virtual talking heads to help children with autistic spectrum disorders learn to improve their social skills. This program is called a virtual tutor.

We would be most grateful if you would volunteer to assist in this project, by consenting to software being installed on your home computer or school computer and having your child interact with that software for 10-15 minutes per day, 3-5 days per week for 3 weeks. The study involves a researcher visit at the beginning of the study (approximately 1 - 1.5hrs), some assessment tasks at the end of the 3 weeks of software use (approximately 1hr, no researcher visit required), a follow up assessment 2 months after that (approximately 1/2hr, no researcher visit required) and a final follow up assessment 4 months after software use has ended (approximately 1/2hr, no researcher visit required). Any researcher visits can take place at your home, school or at a Flinders University campus, as agreed by yourself and the researcher.

Over the three weeks of software use during the study the total time commitment is expected to be a maximum of 3.75 hours interacting with the software across three weeks, plus a maximum of 4 hours meeting with the researcher and completing assessment tasks across four separate occasions. The software aims to improve participants' social skills through observation of and interaction with one or more virtual humans. There are two versions of the software and your child will be asked to interact with one of them. At the conclusion of the study you will be given a chance to use the software that your child was not exposed to previously and will be able to keep a reduced version of the software.

Be assured that any information provided will be treated in the strictest confidence and none of the participants will be individually identifiable in the resulting thesis, report or other publications. Your child is, of course, free to discontinue participation at any time or to decline to participate in particular activities.

Any enquiries you may have concerning this project should be directed to me at the address given above or by telephone on 8201 3663, by fax on 8201 2904 or by email (david.powers@flinders.edu.au).

Thank you for your attention and assistance.

Yours sincerely

Prof David Powers
Director, Artificial Intelligence and Language Technology Laboratory
School of Computer Science, Engineering and Mathematics

This research project has been approved by the Flinders University Social and Behavioural Research Ethics Committee (Project Number 5703). For more information regarding ethical approval of the project the Executive Officer of the Committee can be contacted by telephone on 8201 3116, by fax on 8201 2035 or by email human.researchethics@flinders.edu.au.

Parent Information Sheet

Ms Marissa Milne
PhD Student
School of Computer Science, Engineering and Mathematics
GPO Box 2100
Adelaide SA 5001
Tel: 0455 486 279
marissa.milne@flinders.edu.au
www.csem.flinders.edu.au
CRICOS Provider No. 00114A

Project Description
The aim of this project is to develop virtual tutoring software for improving social skills in children with high functioning autism and Asperger Syndrome. This virtual tutor is to be evaluated for its effectiveness as a teaching tool.

What is a Virtual Tutor?
A virtual tutor is a software program used for education that includes one or more virtual humans who are capable of speech, facial expressions and basic interaction. The software is used to teach children about some aspects of social skills by having them observe and interact with the virtual humans in the software. The faces of the virtual humans are shown on a normal computer screen next to a window displaying interactive tasks.

Your Role
There are two versions of the software being used in this experiment and your child will be asked to use one of them. The details of the software versions will be explained at the end of the study, at which point your child will be able to use the other version if desired. All participants will use software that includes virtual humans, but the exact nature of the software will depend on the version that your child is asked to use.

There will be an initial 1 - 1½ hour visit by the researcher to your home, school or a Flinders University campus, location to be agreed upon by yourself and the researcher. During this visit the researcher will:

- Confirm that you and your child understand the study and are happy to participate
- Install software on the chosen computer and explain its use
- Ask you to complete an assessment of your child's current social behaviour
- Ask your child to complete a test evaluating their social skills knowledge and a questionnaire evaluating past experience with computers and expectations for this software (both conducted via the software)

After this initial visit, your child will be asked to use the software for **10-15 minutes per day, 3-5 days per week for 3 weeks** (a maximum of 3.75 hours). After each session the software will automatically upload your child's interaction data, in anonymised format, to our secure server. After 3 weeks, the software will automatically perform a short content quiz assessing your child's current knowledge and a questionnaire to assess the experience your child had with the software. The software will then disable itself. At this point you will be asked to complete another assessment of your child's social behaviour as well as a feedback questionnaire, both delivered online. Alternatively, printed copies can be provided on request. You can complete these independently, or the researcher can be present to support you in completing these assessments if you prefer, and this can take place at your home, school or a Flinders University campus. *Between the software disabling itself and the final assessment, there are no special activities or tasks that your child or you are required to undertake.*

Two months after the software disables itself, the researcher will contact you to arrange for the next set of assessments to be undertaken. You will be asked to complete another assessment of your child's social behaviour following the same procedure as the last assessment. Your child will also be asked to access the software to complete another content quiz.

Another two months after the previous assessments (a total of 4 months after software use has ceased), the researcher will contact you again. Your child will be asked to access the software to complete another content quiz, and you will be asked to complete another assessment of your child's current social behaviour. This is the final assessment.

Once these assessments are complete the researcher will give you a code to enter into the software. This will enable the software again and allow your child to utilise both versions of the software if you wish - if your child was in an experimental group, the control group activity will be enabled, if your child was in the control group,

the experimental software will be enabled, along with the original content that your child was accessing throughout their participation in the study.

After 3 weeks the software will delete all copyrighted portions of itself. You will be able to keep this reduced version of the software permanently. The reduced version of the software includes the fully functional virtual human and a large selection of activities. Users can also write their own lessons to add to this software.

Any queries about the procedure should be forwarded to Marissa Milne.

Summary of assessment schedule

Assessment	Pre-Test	Post-Test (immediately at end of 3 weeks software use)	2 Month Follow Up Post-Test	4 Month Follow Up Post-Test
Caregiver assessments	Behaviour assessment	Behaviour assessment & questionnaire	Behaviour assessment	Behaviour assessment
Participant assessments	Content quiz & questionnaire	Content quiz & questionnaire	Content quiz	Content quiz

Assessments can take place at the child's home, school or at a Flinders University campus, as agreed by yourself and researcher, and with the permission of the school if conducted on school grounds. The researcher does not need to be present, but can be if you prefer.

Contact Details
You may contact Marissa Milne at anytime with any questions or issues relating to your child's participation in this project, including technical difficulties. See the letter head for contact details.

Privacy and Confidentiality
We respect the privacy and confidentiality of you and your child. Any and all information provided is kept in an anonymous format, and will not be personally identifiable from the stored data or in publication.

Right to Withdraw
Your child's participation in this study is entirely voluntary and you and your child have the right to withdraw from the study at any time without giving a reason. If you or your child decide not to participate in this study or if you or your child withdraw from the study, this may be done so freely, without prejudice.

Reimbursement
Families will be reimbursed $30 total for their involvement in the study, and will be able to keep a reduced version of the software following completion of the study.

Dissemination of Results
Results will be published in a thesis, peer-reviewed journals, conference presentations and other professional forums. In any publication, information will be provided in such a way that you cannot be identified. Results of the study will be provided to you, if you wish.

Benefit to the community
The success of this project is hoped to provide support for further development of and investigation into the use of autonomous virtual agents as social skill teaching tools, and may lead to the release of the software use in this project, or a further developed version of it.

Thank you for taking the time to read this information sheet and we hope that you will accept our invitation to be involved.

This research project has been approved by the Flinders University Social and Behavioural Research Ethics Committee (Project number 5703). For more information regarding ethical approval of the project the Executive Officer of the Committee can be contacted by telephone on 8201 3116, by fax on 8201 2035 or by email human.researchethics@flinders.edu.au

Child Information Sheet

Ms Marissa Milne
PhD Student
School of Computer Science, Engineering and Mathematics
GPO Box 2100
Adelaide SA 5001
Tel: 0455 486 279
marissa.milne@flinders.edu.au
www.csem.flinders.edu.au
CRICOS Provider No. 00114A

Project Description
The aim of this project is to build a 'virtual tutor' that can help you to learn social skills. We will look at how good it was at teaching these skills.

What is a Virtual Tutor?
A virtual tutor is software that includes one or more virtual people that have cartoon-like faces and can talk to you and help you learn things. We show the virtual tutor on a normal computer screen.

Your Role
There are two versions of the software being used in this experiment and you will be asked to use one of them. At the end of the experiment we will tell you the difference between the software versions and you will be able to use the other version as well if you like. It is hoped that by doing activities and watching the virtual people in the software, you will be able to understand more about social skills.

The virtual tutor software will be put on your home or school computer. You will use the virtual tutor for 3 weeks, and you will be asked to use it for 10-15 minutes, 3-5 days per week. After each session the software will automatically upload your answers to our computer.

There will be a visit from the researcher before you start using the software. This visit is to put the software on your computer and do a test and questionnaire about what you already know and what you think this software will be like. This will be done using the software.

When you start the software it will ask you to put in your name and a password. The first time you use it, it will give you some questions to answer. From then on when you use the software, it will give you activities to do instead. You are asked to keep doing activities for 10 to 15 minutes each day that you use the software. The virtual tutor will tell you when time is up.

At the end of the 3 weeks, the software will give you a test to find out what you know and a questionnaire to see what you thought about using it. The software will then stop working. This is so we can get you to do another test and questionnaire a few months later to see if you remembered what you learnt even after you stopped using the software. We will then get you to do one more test a few months after that to see if you can still remember what you learnt.

At the end of the study you will be able to keep a reduced version of the software. We have to remove some parts of it because we are still working on it – this experiment is helping us to make it better!

If you have any questions you should talk to a parent, teacher or caregiver who can contact Marissa Milne for further information (see the letter head for contact details).

What if I don't like it?
If you don't want to use the virtual tutor anymore you can take a break or stop any time. Being part of this study is completely up to you and you can stop at any time without giving a reason.

What happens afterwards?
We will use the information to make the virtual tutor better and to help other people understand how useful it is, but we won't tell anyone else your name or what you did. We hope that this project will eventually mean that you and other children will get to use a virtual tutor at home and at school, and that it will help you to feel more comfortable talking to people and being at school.

Thank you for reading this information sheet and we hope that you will accept our invitation to be involved.

This research project has been approved by the Flinders University Social and Behavioural Research Ethics Committee (Project number 5703). For more information regarding ethical approval of the project the Executive Officer of the Committee can be contacted by telephone on 8201 3116, by fax on 8201 2035 or by email human.researchethics@flinders.edu.au

Appendix O. Consent Form

PARENTAL CONSENT FORM FOR CHILD PARTICIPATION IN RESEARCH

(by experiment)

Autonomous virtual agents as social skills tutors for children with autistic spectrum disorders: Using a virtual person to improve the social skills of children with autism

I .....

being over the age of 18 years hereby consent to my child .....

participating, as requested, in the Letter of Introduction and Information Sheet for the research project on autonomous virtual agents as social skills tutors for children with autistic spectrum disorders.

1. I have read the information provided.
2. Details of procedures and any risks have been explained to my satisfaction.
3. I agree to the software automatically recording my child's information and participation.
4. I am aware that I should retain a copy of the Information Sheet and Consent Form for future reference.
5. I understand that:
   - My child may not directly benefit from taking part in this research.
   - My child is free to withdraw from the project at any time and is free to decline to answer particular questions or complete particular tasks.
   - While the information gained in this study will be published as explained, my child will not be identified, and individual information will remain confidential.
   - Whether my child participates or not, or withdraws after participating, will have no effect on any treatment or service that is being provided to him/her.
   - Whether my child participates or not, or withdraws after participating, will have no effect on his/her progress in his/her course of study, or results gained.
   - My child may ask that the recording/observation be stopped at any time, and he/she may withdraw at any time from the session or the research without disadvantage.

Parent/caregiver signature..... Date.....

C:\Users\Marissa\Desktop\PhD Ethics and Advertising\SBREC\Non-Mods\Consent_ParentChild.doc
Updated 28/6/07

I have discussed this study with my parent/caregiver and I am willing to participate.

Participant's signature..... Date.....

I certify that I have explained the study to the volunteer and consider that she/he understands what is involved and freely consents to participation.

Researcher's name.....

Researcher's signature..... Date.....

REFERENCE LIST

Abirached, B., Y. Zhang, J. K. Aggarwal, B. Tamersoy, T. Fernandes and J. Carlos (2011). Improving communication skills of children with ASDs through interaction with virtual characters. Serious Games and Applications for Health (SeGAH), 2011 IEEE 1st International Conference on.

Allmendinger, K. (2010). "Social Presence in Synchronous Virtual Learning Situations: The Role of Nonverbal Signals Displayed by Avatars." Educational Psychology Review **22**(1): 41-56.

American Psychiatric Association (2000). Diagnostic and statistical manual of mental disorders: DSM-IV-TR. Washington, DC, American Psychiatric Association.

American Psychiatric Association (2013). Autism Spectrum Disorder 299.00. Diagnostic and statistical manual of mental disorders: DSM-5. Washington, DC, American Psychiatric Association.

Amico, M., C. Lalonde and S. Snow (2015). "Evaluating the efficacy of drama therapy in teaching social skills to children with Autism Spectrum Disorders." Drama Therapy Review **1**(1): 21-39.

Astrid, M., N. C. Krämer, J. Gratch and S.-H. Kang (2010). "“It doesn’t matter what you are!” Explaining social effects of agents and avatars." Computers in Human Behavior **26**(6): 1641-1650.

Australian Bureau of Statistics. (2011). "Table 2. Postal Area (POA) Index of Relative Socio-economic Advantage and Disadvantage." Socio-economic Indexes for Areas (SEIFA) Retrieved 25 June, 2017, from http://www.abs.gov.au/websitedbs/censushome.nsf/home/seifa2011?opendocument&navpos=260.

Autism Speaks. (2013). "DSM-5 Diagnostic Criteria." from https://www.autismspeaks.org/what-autism/diagnosis/dsm-5-diagnostic-criteria.

Balakrishnan, S. and A. Alias (2017). "Usage of Social Stories in Encouraging Social Interaction of Children with Autism Spectrum Disorder." Journal of ICSAR **1**(2).

Baron-Cohen, S., O. Golan and E. Ashwin (2009). "Can emotion recognition be taught to children with autism spectrum conditions?" Philosophical Transactions of the Royal Society B: Biological Sciences **364**(1535): 3567-3574.

Bauminger, N. and C. Kasari (2000). "Loneliness and friendship in high-functioning children with autism." Child Dev **71**(2): 447-456.

Beaumont, R. and K. Sofronoff (2008). "A multi-component social skills intervention for children with Asperger syndrome: The Junior Detective Training Program." Journal of Child Psychology and Psychiatry **49**(7): 743-753.

Bellack, A. (1983). "Recurrent problems in the behavioral assessment of social skill." Behaviour Research and Therapy **21**(1): 29-41.

Bellini, S. (2006). "Autism social skills profile." Building Social Relationships: A Systematic Approach to Teaching Social Interaction Skills to Children and Adolescents with Autism Spectrum Disorders and Other Social Difficulties.

Bellini, S. (2007). Building social relationships: A systematic approach to teaching social interaction skills to children and adolescents with autism spectrum disorders and other social difficulties, AAPC Publishing.

Bernardini, S., K. Porayska-Pomsta and T. J. Smith (2014). "ECHOES: An intelligent serious game for fostering social communication in children with autism." Information Sciences **264**: 41-60.

Biswas, G., R. Roscoe, H. Jeong and B. Sulcer (2009). Promoting Self-Regulated Learning Skills in Agent-based Learning Environments.

Black, P. (2015). "Assessment: Friend or Foe of Pedagogy and Learning." Past as Prologue: 235.

Black, P. and D. Wiliam (2009). "Developing the theory of formative assessment." Educational Assessment, Evaluation and Accountability **21**(1): 5-31.

Blair, K., D. Schwartz, G. Biswas and K. Leelawong (2007). "Pedagogical agents for learning by teaching: teachable agents." Educational Technology **47**(1): 56.

Bock, M., M. F. Rogers and B. S. Myles (2001). "Using Social Stories and Comic Strip Conversations to Interpret Social Situations for an Adolescent with Asperger Syndrome." Intervention in School and Clinic **36**(5): 310-313.

Borjigin, A., C. Miao, S. F. Lim, S. Li and Z. Shen (2015). Teachable Agents with Intrinsic Motivation. Artificial Intelligence in Education: 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings. C. Conati, N. Heffernan, A. Mitrovic and M. F. Verdejo. Cham, Springer International Publishing**:** 34-43.
Bosseler, A. and D. Massaro (2003). "Development and Evaluation of a Computer-Animated Tutor for Vocabulary and Language Learning in Children with Autism." Journal of Autism and Developmental Disorders **33**(6).
Bowman-Perrott, L., H. Davis, K. Vannest, L. Williams, C. Greenwood and R. Parker (2013). "Academic Benefits of Peer Tutoring: A Meta-Analytic Review of Single-Case Research." School Psychology Review **42**(1): 39-55.
Boyle, D. and C. Hassett-Walker (2008). "Reducing overt and relational aggression among young children: The results from a two-year outcome evaluation." Journal of School Violence **7**(1): 27-42.
Boyle, E. A., T. Hainey, T. M. Connolly, G. Gray, J. Earp, M. Ott, T. Lim, M. Ninaus, C. Ribeiro and J. Pereira (2016). "An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games." Computers & Education **94**: 178-192.
Brown, D. J., P. J. Standen, T. Proctor and D. Sterland (2001). "Advanced Design Methodologies for the Production of Virtual Learning Environments for Use by People with Learning Disabilities." Presence: Teleoper. Virtual Environ. **10**(4): 401-415.
Butler, L. B. and V. Poedubicky (2006). "Social Decision Making/Social Problem Solving." The Educator's Guide to Emotional Intelligence and Academic Achievement: Social-Emotional Learning in the Classroom: 131.
Carter, A. S., F. R. Volkmar, S. S. Sparrow, J.-J. Wang, C. Lord, G. Dawson, E. Fombonne, K. Loveland, G. Mesibov and E. Schopler (1998). "The Vineland Adaptive Behavior Scales: Supplementary Norms for Individuals with Autism." Journal of Autism and Developmental Disorders **28**(4): 287-302.
Chen, C.-H., I. J. Lee and L.-Y. Lin (2015). "Augmented reality-based self-facial modeling to promote the emotional expression and social skills of adolescents with autism spectrum disorders." Research in Developmental Disabilities **36**: 396-403.
Cheng, Y., C.-L. Huang and C.-S. Yang (2015). "Using a 3D Immersive Virtual Environment System to Enhance Social Understanding and Social Skills for Children With Autism Spectrum Disorders." Focus on Autism and Other Developmental Disabilities **30**(4): 222-236.
Chi, M., S. Siler, H. Jeong, T. Yamauchi and R. Hausmann (2001). "Learning from human tutoring." Cognitive Science **25**(4): 471-533.
Chi, M. T. H. and R. Wylie (2014). "The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes." Educational Psychologist **49**(4): 219-243.
Clark, R. E. and S. Choi (2005). "Five design principles for experiments on the effects of animated pedagogical agents." Journal of Educational Computing Research **32(**3): 209-225.
Cline, B., C. Brewster and R. Fell (2010). "A rule-based system for automatically evaluating student concept maps." Expert Systems with Applications **37**(3): 2282-2291.
Conati, C. (2002). "PROBABILISTIC ASSESSMENT OF USER S EMOTIONS IN EDUCATIONAL GAMES." Applied Artificial Intelligence **16**(7): 555-575.
Conati, C. and M. Manske (2009). Evaluating Adaptive Feedback in an Educational Computer Game, Springer.
Connell, B. R., M. Jones, R. Mace, J. Mueller, A. Mullick, E. Ostroff, J. Sanford, E. Steinfeld, M. Story and G. Vanderheiden (1997). The Principles of Universal Design. The Center for Universal Design. Raleigh, NC, North Carolina State University.
Cooper, J., H. Goodfellow, E. Muhlheim, K. Paske and L. Pearson (2003a). PALS social skills program: Playing And Learning to Socialise: Resource Book, Inscript Publishing.
Cooper, J., K.-a. Paske and M. Zuzic (2003b). "Teaching social and problem-solving skills to reduce behaviour problems in early childhood."

Cooper, J. A., K.-A. Paske, H. Goodfellow and E. Muhlheim (2002). "Social skills training to reduce aggressive and withdrawn behaviours in child care centres." Australian Journal of Early Childhood **27**(4): 29-36.
Crawford, H., J. Moss, C. Oliver, N. Elliott, G. M. Anderson and J. P. McCleery (2016). "Visual preference for social stimuli in individuals with autism or neurodevelopmental disorders: an eye-tracking study." Molecular Autism **7**(1): 24.
Crook, C. and R. Sutherland (2017). Technology and Theories of Learning. Technology Enhanced Learning: Research Themes. E. Duval, M. Sharples and R. Sutherland. Cham, Springer International Publishing**:** 11-27.
Crooke, P., R. Hendrix and J. Rachman (2007). "Measuring the Effectiveness of Teaching Social Thinking to Children with Asperger Syndrome (AS) and High Functioning Autism (HFA)." Journal of Autism and Developmental Disorders **38**(3): 10.
Dalton, K. M., B. M. Nacewicz, T. Johnstone, H. S. Schaefer, M. A. Gernsbacher, H. H. Goldsmith, A. L. Alexander and R. J. Davidson (2005). "Gaze fixation and the neural circuitry of face processing in autism." Nat Neurosci **8**(4): 519-526.
Davis, M., B. Robins, K. Dautenhahn, C. Nehaniv and S. Powell (2005). A Comparison of Interactive and Robotic Systems in Therapy and Education for Children with Autism. Assistive Technology: From Virtuality to Reality. A. Pruski and H. Knops, IOS Press**:** 5.
Davis, N. O., A. S. Carter, F. R. Volkmar, R. Paul, S. J. Rogers and K. A. Pelphrey (2014). Social Development in Autism. Handbook of Autism and Pervasive Developmental Disorders, Fourth Edition, John Wiley & Sons, Inc.
de Klerk, S., T. J. H. M. Eggen and B. P. Veldkamp (2016). "A methodology for applying students' interactive task performance scores from a multimedia-based performance assessment in a Bayesian Network." Computers in Human Behavior **60**: 264-279.
Delano, M. and M. E. Snell (2006). "The Effects of Social Stories on the Social Engagement of Children with Autism." Journal of Positive Behavior Interventions **8**(1): 29-42.
Dowling, L. S. (2010). "The Efficacy of Skillstreaming Social Skills Program with Elementary School Children with High Functioning Autism."
Dowrick, P. W. (2012). "Self modeling: Expanding the theories of learning." Psychology in the Schools **49**(1): 30-41.
Dunn, M. A. (2005). SOS Social Skills in Our Schools: A social skills program for children with pervasive developmental disorders, including high-functioning autism and Asperger syndrome, and their typical peers, AAPC Publishing.
Elias, M. J. (1983). "Improving coping skills of emotionally disturbed boys through television-based social problem solving." American journal of Orthopsychiatry **53**(1): 61.
Elias, M. J. and L. B. Butler (2005). Social decision making/social problem solving for middle school students: Skills and activities for academic, social, and emotional success, Research PressPub.
Elias, M. J., M. Gara, M. Ubriaco, P. A. Rothbaum, J. F. Clabby and T. Schuyler (1986). "Impact of a preventive social problem solving intervention on children's coping with middle-school stressors." American Journal of Community Psychology **14**(3): 259-275.
Fancsali, S. E., S. Ritter, M. Yudelson, M. Sandbothe and S. R. Berman (2016). Implementation Factors and Outcomes for Intelligent Tutoring Systems: A Case Study of Time and Efficiency with Cognitive Tutor Algebra. FLAIRS Conference.
Finnegan, E. and A. L. Mazin (2016). "Strategies for Increasing Reading Comprehension Skills in Students with Autism Spectrum Disorder: A Review of the Literature." Education & Treatment of Children **39**(2): 187-219.
Fletcher-Watson, S., H. Pain, S. Hammond, A. Humphry and H. McConachie (2016). "Designing for young children with autism spectrum disorder: A case study of an iPad app." International Journal of Child-Computer Interaction **7**: 1-14.
Gay, V., P. Leijdekkers and A. Pooley (2016). Building Social Awareness for Teens and Young Adults with Autism via Gamification. Serious Games: Second Joint International Conference, JCSG

2016, Brisbane, QLD, Australia, September 26-27, 2016, Proceedings. T. Marsh, M. Ma, M. F. Oliveira, J. Baalsrud Hauge and S. Göbel. Cham, Springer International Publishing**:** 116-127.
Ge, Z. and L. Fan (2017). Social Development for Children with Autism Using Kinect Gesture Games: A Case Study in Suzhou Industrial Park Renai School. Simulation and Serious Games for Education. Y. Cai, S. L. Goei and W. Trooster. Singapore, Springer Singapore**:** 113-123.
Gillis, J., E. Callahan and R. Romanczyk (2010). "Assessment of social behavior in children with autism: The development of the Behavioral Assessment of Social Interactions in Young Children." Research in Autism Spectrum Disorders.
Graesser, A., P. Chipman, B. Haynes and A. Olney (2005). "AutoTutor: An intelligent tutoring system with mixed-initiative dialogue." IEEE Transactions on Education **48**(4).
Graesser, A., K. Wiemer-Hastings, P. Wiemer-Hastings and R. Kreuz (1999). "AutoTutor: A simulation of a human tutor." Cognitive Systems Research **1**(1): 35-51.
Grafsgaard, J. F., J. B. Wiggins, K. E. Boyer, E. N. Wiebe and J. C. Lester (2013). Automatically recognizing facial indicators of frustration: a learning-centric analysis. Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, IEEE.
Grandin, T. (1995). The Learning Style of People with Autism: An Autobiography Teaching children with autism: Strategies to enhance communication and socialization. K. Quill, Cengage Learning.
Gray, C. (2001). "How to Respond to a Bullying Attempt: What to Think, Say and Do." The Morning News **13**(2).
Grimm, P. (2010). "Social desirability bias." Wiley International Encyclopedia of Marketing.
Gunn, K. C. M. and J. T. Delafield-Butt (2016). "Teaching Children With Autism Spectrum Disorder With Restricted Interests." Review of Educational Research **86**(2): 408-430.
Hailpern, J. (2007). "Encouraging speech and vocalization in children with autistic spectrum disorder." SIGACCESS Access. Comput.(89): 47-52.
Hailpern, J., K. Karahalios and J. Halle (2009). Creating a spoken impact: encouraging vocalization through audio visual feedback in children with ASD. Proceedings of the 27th international conference on Human factors in computing systems. Boston, MA, USA, ACM.
Hall, L., S. Woods and M. Hall (2009). "Lessons Learned Using Theory of Mind Methods to Investigate User Social Awareness in Virtual Role-Play." Human technology **5**(1).
Hamari, J., J. Koivisto and H. Sarsa (2014). Does Gamification Work? -- A Literature Review of Empirical Studies on Gamification. 2014 47th Hawaii International Conference on System Sciences.
Hamer, J., H. Purchase, A. Luxton-Reilly and P. Denny (2015). "A comparison of peer and tutor feedback." Assessment & Evaluation in Higher Education **40**(1): 151-164.
Hattie, J. and H. Timperley (2007). "The power of feedback." Review of educational research **77**(1): 81.
Helal, A. S., M. Mokhtari and B. Abdulrazak, Eds. (2008). The Engineering Handbook of Smart Technology for Aging, Disability and Independence. New Jersey, John Wiley & Sons, Inc.
Herrera, G., F. Alcantud, R. Jordan, A. Blanquer, G. Labajo and C. De Pablo (2008). "Development of symbolic play through the use of virtual reality tools in children with autistic spectrum disorders: Two case studies." Autism **12**(2): 143-157.
Herring, P. J. (2015). Design and evaluation of a CAL system to support communication development in children with autism, The Open University.
Hirschberg, J., D. Litman and M. Swerts (2004). "Prosodic and other cues to speech recognition failures." Speech Communication **43**(1-2): 155-175.
Hmelo-Silver, C. E. and H. S. Barrows (2015). "Problem-based learning: goals for learning and strategies for facilitating." A. Walker, H. Leary, CE Hmelo-Silver, & P. A. Ertmer, PA (Eds.), Essential readings in problem-based learning: Exploring and extending the legacy of Howard S. Barrows: 69-84.
Ho, C.-C. and K. F. MacDorman (2017). "Measuring the Uncanny Valley Effect." International Journal of Social Robotics **9**(1): 129-139.

Holm, S. (1979). "A simple sequentially rejective multiple test procedure." Scandinavian journal of statistics: 65-70.
Hopkins, I. M., M. W. Gower, T. A. Perez, D. S. Smith, F. R. Amthor, F. Casey Wimsatt and F. J. Biasini (2011). "Avatar Assistant: Improving Social Skills in Students with an ASD Through a Computer-Based Intervention." Journal of Autism and Developmental Disorders **41**(11): 1543-1555.
Hoque, M. E. (2008). Analysis of speech properties of neurotypicals and individuals diagnosed with autism and down. Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility. Halifax, Nova Scotia, Canada, ACM.
Hourihan, F. and D. Hoban (2004). "Learning, Enjoying, Growing, Support model: an innovative collaborative approach to the prevention of conduct disorder in preschoolers in hard to reach rural families." Australian Journal of Rural Health **12**(6): 269-276.
Hu, X. and H. Xia (2010). Automated Assessment System for Subjective Questions Based on LSI, IEEE.
Huijnen, C. A. G. J., M. A. S. Lexis, R. Jansens and L. P. de Witte (2016). "Mapping Robots to Therapy and Educational Objectives for Children with Autism Spectrum Disorder." Journal of Autism and Developmental Disorders **46**(6): 2100-2114.
Huskens, B., A. Palmen, M. Van der Werff, T. Lourens and E. Barakova (2015). "Improving Collaborative Play Between Children with Autism Spectrum Disorders and Their Siblings: The Effectiveness of a Robot-Mediated Intervention Based on Lego® Therapy." Journal of Autism and Developmental Disorders **45**(11): 3746-3755.
Iacobelli, F. and J. Cassell (2007). Ethnic Identity and Engagement in Embodied Conversational Agents. Intelligent Virtual Agents. C. Pelachaud, J.-C. Martin, E. André et al., Springer Berlin Heidelberg. **4722:** 57-63.
Jacklin, A. and W. Farr (2005). "The computer in the classroom: a medium for enhancing social interaction with young people with autistic spectrum disorders?" British Journal of Special Education **32**(4): 202-210.
Jackson, G., R. Guess and D. McNamara (2010a). "Assessing cognitively complex strategy use in an untrained domain." Topics in Cognitive Science **2**(1): 127-137.
Jackson, G. T., K. B. Dempsey and D. S. McNamara (2010b). The Evolution of an Automated Reading Strategy Tutor: From the Classroom to a Game-Enhanced Automated System. New Science of Learning: Cognition, Computers and Collaboration in Education. M. S. Khine and I. M. Saleh. New York, NY, Springer New York**:** 283-306.
Jackson, S. L. J. and B. Dritschel (2016). "Modeling the impact of social problem-solving deficits on depressive vulnerability in the broader autism phenotype." Research in Autism Spectrum Disorders **21**: 128-138.
Jacovina, M. E., G. T. Jackson, E. L. Snow and D. S. McNamara (2016). Timing Game-Based Practice in a Reading Comprehension Strategy Tutor. International Conference on Intelligent Tutoring Systems, Springer.
James, S. and J. Mellor (2007). Evaluating the use of the Playing and Learning to Socialise (PALS) Programme, Citeseer.
Jokisch, O., H. Hain, R. Petrick and R. Hoffmann (2009). Robustness optimization of a speech interface for child-directed embedded language tutoring, ACM.
Jones, P., C. Wilcox and J. Simon (2016). Evidence-Based Instruction for Students with Autism Spectrum Disorder: TeachTown Basics. Technology and the Treatment of Children with Autism Spectrum Disorder, Springer**:** 113-129.
Jones, R. S. P., C. Quigney and J. C. Huws (2003). "First-hand accounts of sensory perceptual experiences in autism: a qualitative analysis." Journal of Intellectual & Developmental Disability **28**(2): 112-121.
Jones, V. (2010). An evaluation of PALS (Play and Learning to Socialise) in primary school settings in Ireland, Institute of Education, University of London.

Joshi, R. (2013) "The new Autism Spectrum Disorder (ASD) Diagnostic Criteria as defined by the DSM-V."
Kam, C.-M., M. T. Greenberg and C. A. Kusché (2004). "Sustained Effects of the PATHS Curriculum on the Social and Psychological Adjustment of Children in Special Education." Journal of Emotional and Behavioral Disorders **12**(2): 66-78.
Kamps, D. M., B. R. Leonard, S. Vernon, E. P. Dugan, J. C. Delquadri, B. Gershon, L. Wade and L. Folk (1992). "Teaching social skills to students with autism to increase peer interactions in an integrated first-grade classroom." Journal of Applied Behavior Analysis **25**(2): 281-288.
Kathy, Y., K. Luke, P. M. Cheung, S. T. T. Tank, L. Cheng and I. Wong (2009). "A case series on the social thinking training of mainstreamed secondary school students with high-functioning autism." Mental Health **35**: 10-17.
Ke, F. and T. Im (2013). "Virtual-Reality-Based Social Interaction Training for Children with High-Functioning Autism." The Journal of Educational Research **106**(6): 441-461.
Kerr, S. J. (2002). Scaffolding: design issues in single & collaborative virtual environments for social skills learning. Proceedings of the workshop on Virtual environments 2002. Barcelona, Spain, Eurographics Association.
KidsMatter. (2016a). "Australian Mental Health and Well-Being Initiative." Retrieved 5 December, 2016, from https://www.kidsmatter.edu.au/.
KidsMatter. (2016b). "Friendly Kids, Friendly Classrooms." Australian Primary Schools Health Initiative Retrieved 5 December, 2016, from https://www.kidsmatter.edu.au/primary/programs/friendly-kids-friendly-classrooms.
KidsMatter. (2016c). "I Can Problem Solve." Australian Primary Schools Health Initiative Retrieved 5 December, 2016, from http://www.kidsmatter.edu.au/primary/programs/i-can-problem-solve.
KidsMatter. (2016d). "PALS Social Skills Program: Playing and Learning to Socialise." Australian Early Childhood Mental Health Initiative Retrieved 5 December, 2016, from https://www.kidsmatter.edu.au/early-childhood/programs/pals-social-skills-program-playing-and-learning-socialise.
KidsMatter. (2016e). "PATHS Curriculum." Australian Primary Schools Health Initiative Retrieved 5 December, 2016, from https://www.kidsmatter.edu.au/primary/programs/paths-curriculum.
KidsMatter. (2016f). "Social Decision Making Problem Solving." Australian Primary Schools Health Initiative Retrieved 5 December, 2016, from https://www.kidsmatter.edu.au/primary/programs/social-decision-making-problem-solving.
Kinchin, I., D. Hay and A. Adams (2000). "How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development." Educational Research **42**(1): 43-57.
Kirby, A. V., B. A. Boyd, K. L. Williams, R. A. Faldowski and G. T. Baranek (2017). "Sensory and repetitive behaviors among children with autism spectrum disorder at home." Autism **21**(2): 142-154.
Knight, V., E. Sartini and A. D. Spriggs (2015). "Evaluating Visual Activity Schedules as Evidence-Based Practice for Individuals with Autism Spectrum Disorders." Journal of Autism and Developmental Disorders **45**(1): 157-178.
Kohls, G., B. Yerys and R. T. Schultz (2014). "Striatal development in autism: repetitive behaviors and the reward circuitry." Biological psychiatry **76**(5): 358-359.
Koning, C., J. Magill-Evans, J. Volden and B. Dick (2013). "Efficacy of cognitive behavior therapy-based social skills intervention for school-aged boys with autism spectrum disorders." Research in Autism Spectrum Disorders **7**(10): 1282-1290.
Kort, B., R. Reilly and R. Picard (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion, Citeseer.
Kozima, H., M. P. Michalowski and C. Nakagawa (2009). "Keepon." International Journal of Social Robotics **1**(1): 3-18.

Krämer, N. (2006). Theory of mind as a theoretical prerequisite to model communication with virtual humans, Springer-Verlag.
Krämer, N. and G. Bente (2010). "Personalizing e-Learning. The Social Effects of Pedagogical Agents." Educational Psychology Review **22**(1): 71-87.
Kroncke, A. P., M. Willard and H. Huckabee (2016). What Is Autism? History and Foundations. Assessment of Autism Spectrum Disorder: Critical Issues in Clinical, Forensic and School Settings. Cham, Springer International Publishing**:** 3-9.
Kusché, C. A. and M. T. Greenberg (1994). The PATHS curriculum, Seattle, WA: Developmental Research and Programs.
Lakens, D. (2013). "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs." Frontiers in Psychology **4**(863).
Laugeson, E. A., F. Frankel, A. Gantman, A. R. Dillon and C. Mogil (2012). "Evidence-Based Social Skills Training for Adolescents with Autism Spectrum Disorders: The UCLA PEERS Program." Journal of Autism and Developmental Disorders **42**(6): 1025-1036.
Laushey, K., L. Heflin, M. Shippen, P. Alberto and L. Fredrick (2009). "Concept Mastery Routines to Teach Social Skills to Elementary Children with High Functioning Autism." Journal of Autism and Developmental Disorders **39**(10): 1435-1448.
Lehman, J. F. (1998). Toward the use of speech and natural language technology in intervention for a language-disordered population. Proceedings of the third international ACM conference on Assistive technologies. Marina del Rey, California, United States, ACM.
Leslie, A. M. (1987). "Pretense and Representation: The Origins of "Theory of Mind"." Psychological Review **94**(4): 15.
Limoges, E., L. Mottron, C. Bolduc, C. Berthiaume and R. Godbout (2005). "Atypical sleep architecture and the autism phenotype." Brain **128**(5): 1049-1061.
Locke, J., C. Kasari and J. J. Wood (2013). "Assessing Social Skills in Early Elementary-Aged Children With Autism Spectrum Disorders." Journal of Psychoeducational Assessment **32**(1): 62-76.
Loomes, R., L. Hull and W. P. L. Mandy (2017). "What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis." Journal of the American Academy of Child & Adolescent Psychiatry **56**(6): 466-474.
Lopata, C., M. L. Thomeer, M. A. Volker and R. E. Nida (2006). "Effectiveness of a cognitive-behavioral treatment on the social behaviors of children with Asperger disorder." Focus on Autism and Other Developmental Disabilities **21**(4): 237-244.
Lorimer, P. A., R. L. Simpson, B. Smith Myles and J. B. Ganz (2002). "The Use of Social Stories as a Preventative Behavioral Intervention in a Home Setting with a Child with Autism." Journal of Positive Behavior Interventions **4**(1): 53-60.
Lovaas, O. I. (1987). "Behavioral treatment and normal educational and intellectual functioning in young autistic children." Journal of consulting and clinical psychology **55**(1): 3.
Luerssen, M. and T. Lewis (2009). Head X: Tailorable Audiovisual Synthesis for ECAs. HCSNet Summerfest 2009, Sydney.
MacMullin, J. A., Y. Lunsky and J. A. Weiss (2016). "Plugged in: Electronics use in youth and young adults with autism spectrum disorder." Autism **20**(1): 45-54.
Madsen, M., R. e. Kaliouby, M. Goodwin and R. Picard (2008). Technology for just-in-time in-situ learning of facial affect for persons diagnosed with an autism spectrum disorder. Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility. Halifax, Nova Scotia, Canada, ACM.
Magiati, I., X. W. Tay and P. Howlin (2014). "Cognitive, language, social and behavioural outcomes in adults with autism spectrum disorders: A systematic review of longitudinal follow-up studies in adulthood." Clinical Psychology Review **34**(1): 73-86.
Marcus, A. and D. Wilder (2009). "A comparison of peer video modeling and self video modeling to teach textual responses in children with autism." Journal of applied behavior analysis **42**(2): 335.

Massaro, D. W. (2004). Symbiotic Value of an Embodied Agent in Language Learning. Hawaii International Conference on System Sciences. **5:** 50132c-50132c.
Matson, J. L., M. Horovitz, S. Mahan and J. Fodstad (2013). "Reliability of the Matson Evaluation of Social Skills with Youngsters (MESSY) for children with Autism Spectrum Disorders." Research in Autism Spectrum Disorders **7**(2): 405-410.
McCleery, J. P. (2015). "Comment on Technology-Based Intervention Research for Individuals on the Autism Spectrum." Journal of Autism and Developmental Disorders **45(**12): 3832-3835.
McGinnis, E. and A. P. Goldstein (2012). Skillstreaming the elementary school child: A Guide for Teaching Prosocial Skills, Research Press.
McGrath, H. and S. Francey (1991). Friendly kids, friendly classrooms: teaching social skills and confidence in the classroom, Longman Cheshire.
McNamara, D., I. Levinstein and C. Boonthum (2004). "iSTART: Interactive strategy training for active reading and thinking." Behavior Research Methods, Instruments, & Computers **36**(2): 222.
Meder, A. M. and J. R. Wegner (2015). "iPads, Mobile Technologies, and Communication Applications: A Survey of Family Wants, Needs, and Preferences." Augmentative and Alternative Communication **31**(1): 27-36.
Mei, C., L. Mason and J. Quarles (2015). How 3D Virtual Humans Built by Adolescents with ASD Affect Their 3D Interactions. Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility. Lisbon, Portugal, ACM**:** 155-162.
Merrill, D. C., B. J. Reiser, M. Ranney and J. G. Trafton (1992). "Effective tutoring techniques: A comparison of human tutors and intelligent tutoring systems." The Journal of the Learning Sciences **2**(3): 277-305.
Meyer, J. and R. Land (2010). "Threshold concepts and troublesome knowledge (5): Dynamics of assessment." Threshold concepts and transformational learning: 61-79.
Michael, D. R. and S. L. Chen (2005). Serious games: Games that educate, train, and inform, Muska & Lipman/Premier-Trade.
Milne, M., R. Leibbrandt, P. Raghavendra, M. Luerssen, T. Lewis and D. Powers (2013). Lesson authoring system for creating interactive activities involving virtual humans the thinking head whiteboard. Intelligent Agent (IA), 2013 IEEE Symposium on.
Milne, M., M. Luerssen, T. Lewis, R. Leibbrandt and D. Powers (2011). Designing and Evaluating Interactive Agents as Social Skills Tutors for Children with Autism Spectrum Disorder. Conversational Agents and Natural Language Interaction: Techniques and Effective Practices. D. Perez-Marin and I. Pascual-Nieto. United States of America, IGI Global**:** 23-48.
Milne, M., D. Powers and R. Leibbrandt (2009). Development of a software-based social tutor for children with autism spectrum disorders, ACM.
Mitchel, K., K. Regehr, J. Reaume and M. Feldman (2010). "Group social skills training for adolescents with Asperger Syndrome or high functioning autism." Journal on Developmental Disabilities.
Mitrovic, A. (2001). "Investigating students' self-assessment skills." User Modeling: 247-250.
Mondragon, A. L., R. Nkambou and P. Poirier (2016). Evaluating the Effectiveness of an Affective Tutoring Agent in Specialized Education. Adaptive and Adaptable Learning: 11th European Conference on Technology Enhanced Learning, EC-TEL 2016, Lyon, France, September 13-16, 2016, Proceedings. K. Verbert, M. Sharples and T. Klobučar. Cham, Springer International Publishing**:** 446-452.
Monkaresi, H., N. Bosch, R. A. Calvo and S. K. D'Mello (2017). "Automated detection of engagement using video-based estimation of facial expressions and heart rate." IEEE Transactions on Affective Computing **8**(1): 15-28.
Mori, M. (1970). "The uncanny valley." Energy **7**(4): 33-35.
Mostow, J. (2005). Project LISTEN: A Reading Tutor that Listens. Pittsburgh, Carnegie Mellon University, School of Computer Science. **2010**.

Myles, B. S., K. Hagen, J. Holverstott, A. Hubbard, D. Adreon and M. Trautman (2005). Life Journey Through Autism: An Educator's Guide to Asperger Syndrome. Arlington, VA, Organization for Autism Research, Inc.
Navarathna, R., D. Dean, P. Lucey and S. Sridharan (2010). "Audio visual automatic speech recognition in vehicles."
Neely, L. C., J. B. Ganz, J. L. Davis, M. B. Boles, E. R. Hong, J. Ninci and W. D. Gilliland (2016). "Generalization and Maintenance of Functional Living Skills for Individuals with Autism Spectrum Disorder: a Review and Meta-Analysis." Review Journal of Autism and Developmental Disorders **3**(1): 37-47.
Ng, A. H. S., K. Schulze, E. Rudrud and J. B. Leaf (2016). "Using the Teaching Interactions Procedure to Teach Social Skills to Children With Autism and Intellectual Disability." American Journal on Intellectual and Developmental Disabilities **121**(6): 501-519.
Nicholas, M., P. Van Bergen and D. Richards (2015). "Enhancing learning in a virtual world using highly elaborative reminiscing as a reflective tool." Learning and Instruction **36**(Supplement C): 66-75.
Noterdaeme, M., E. Wriedt and C. Höhne (2010). "Asperger's syndrome and high-functioning autism: language, motor and cognitive profiles." European Child & Adolescent Psychiatry **19**(6): 475-481.
O'Regan, K. (2003). "Emotion and e-learning." Journal of Asynchronous learning networks **7**(3): 78-92.
Owens, G., Y. Granader, A. Humphrey and S. Baron-Cohen (2008). "LEGO® Therapy and the Social Use of Language Programme: An Evaluation of Two Social Skills Interventions for Children with High Functioning Autism and Asperger Syndrome." Journal of Autism and Developmental Disorders **38**(10): 1944-1957.
Ozsivadjian, A., C. Hibberd and M. J. Hollocks (2014). "Brief Report: The Use of Self-Report Measures in Young People with Autism Spectrum Disorder to Access Symptoms of Anxiety, Depression and Negative Thoughts." Journal of Autism and Developmental Disorders **44**(4): 969-974.
Panerai, S., L. Ferrante and M. Zingale (2002). "Benefits of the Treatment and Education of Autistic and Communication Handicapped Children (TEACCH) programme as compared with a non-specific approach." Journal of Intellectual Disability Research **46**(4): 318-327.
Park, U. and R. Calvo (2008). Automatic concept map scoring framework using the semantic web technologies.
Parsons, S., L. Beardon, H. Neale, G. Reynard, R. Eastgate, J. Wilson, S. Cobb, S. Benford, P. Mitchell and E. Hopkins (2000). "Development of social skills amongst adults with Asperger's Syndrome using virtual environments: the 'AS Interactive'project." 3rd ICDVRAT, Sardinia Italy; University of Reading: 163-170.
Parsons, S. and P. Mitchell (2002). "The potential of virtual reality in social skills training for people with autistic spectrum disorders." Journal of Intellectual Disability Research **46**(5): 430-443.
Parsons, S., P. Mitchell and A. Leonard (2005). "Do adolescents with autistic spectrum disorders adhere to social conventions in virtual environments?" Autism **9**(1): 95-117.
Parsons, S., N. Yuill, M. Brosnan and J. Good (2017). "Interdisciplinary perspectives on designing, understanding and evaluating digital technologies for autism." Journal of Enabling Technologies **11**(1): 13-18.
Parsons, S., N. Yuill, J. Good, M. Brosnan, L. Austin, C. Singleton, B. Bossavit and Barnabear (2016). What Technology for Autism Needs to be Invented? Idea Generation from the Autism Community via the ASCmeI.T. App. Computers Helping People with Special Needs: 15th International Conference, ICCHP 2016, Linz, Austria, July 13-15, 2016, Proceedings, Part II. K. Miesenberger, C. Bühler and P. Penaz. Cham, Springer International Publishing**:** 343-350.
Peck, T. C., S. Seinfeld, S. M. Aglioti and M. Slater (2013). "Putting yourself in the skin of a black avatar reduces implicit racial bias." Consciousness and cognition **22**(3): 779-787.

Perry, A., H. E. Flanagan, J. Dunn Geier and N. L. Freeman (2009). "Brief Report: The Vineland Adaptive Behavior Scales in Young Children with Autism Spectrum Disorders at Different Cognitive Levels." Journal of Autism and Developmental Disorders **39**(7): 1066-1078.
Pickles, A., D. K. Anderson and C. Lord (2014). "Heterogeneity and plasticity in the development of language: a 17-year follow-up of children referred early for possible autism." Journal of Child Psychology and Psychiatry **55**(12): 1354-1362.
Pierce, J. M., A. D. Spriggs, D. L. Gast and D. Luscre (2013). "Effects of Visual Activity Schedules on Independent Classroom Transitions for Students with Autism." International Journal of Disability, Development and Education **60**(3): 253-269.
Piper, A. M., E. O'Brien, M. R. Morris and T. Winograd (2006). SIDES: a cooperative tabletop computer game for social skills development. Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work. Banff, Alberta, Canada, ACM.
Ploog, B. O., A. Scharf, D. Nelson and P. J. Brooks (2013). "Use of Computer-Assisted Technologies (CAT) to Enhance Social, Communicative, and Language Development in Children with Autism Spectrum Disorders." Journal of Autism and Developmental Disorders **43**(2): 301-322.
Posserud, M., M. Hysing, W. Helland, C. Gillberg and A. Lundervold (2016). "Autism traits: the importance of "co-morbid" problems for impairment and contact with services. Data from the Bergen Child Study." Research in developmental disabilities.
Pritschet, L., D. Powell and Z. Horne (2016). "Marginally Significant Effects as Evidence for Hypotheses." Psychological Science **27**(7): 1036-1042.
Promising Practices Network. (2016). "Social Decision Making / Problem Solving." What Works for Children and Families Retrieved 5 December, 2016, from http://www.promisingpractices.net/program.asp?programid=154.
Putnam, C. and L. Chong (2008). Software and technologies designed for people with autism: what do users want? Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility. Halifax, Nova Scotia, Canada, ACM.
Putnam, R. T. (1987). "Structuring and Adjusting Content for Students: A Study of Live and Simulated Tutoring of Addition." American Educational Research Journal **24**(1): 13-48.
Quill, K. A. (1997). "Instructional Considerations for Young Children with Autism: The Rationale for Visually Cued Instruction." Journal of Autism and Developmental Disorders **27**(6): 697-714.
Quirmbach, L., A. Lincoln, M. Feinberg-Gizzo, B. Ingersoll and S. Andrews (2008). "Social Stories: Mechanisms of Effectiveness in Increasing Game Play Skills in Children Diagnosed with Autism Spectrum Disorder Using a Pretest Posttest Repeated Measures Randomized Control Group Design." Journal of Autism and Developmental Disorders **39**(2): 299-321.
Radley, K. C., W. B. Ford, M. B. McHugh, K. Dadakhodjaeva, R. D. O'Handley, A. A. Battaglia and J. D. K. Lum (2015). "Brief Report: Use of Superheroes Social Skills to Promote Accurate Social Skill Use in Children with Autism Spectrum Disorder." Journal of Autism and Developmental Disorders **45**(9): 3048-3054.
Rai, D., J. Beck and I. Arroyo (2013). Causal Modeling to Understand the Relationship between Student Attitudes, Affect and Outcomes. Educational Data Mining 2013.
Ramdoss, S., A. Mulloy, R. Lang, M. O'Reilly, J. Sigafoos, G. Lancioni, R. Didden and F. El Zein (2011). "Use of computer-based interventions to improve literacy skills in students with autism spectrum disorders: A systematic review." Research in Autism Spectrum Disorders **5**(4): 1306-1318.
Rao, P. A., D. C. Beidel and M. J. Murray (2008). "Social Skills Interventions for Children with Asperger's Syndrome or High-Functioning Autism: A Review and Recommendations." Journal of Autism and Developmental Disorders **38**: 353–361.
Rapin, I. and R. F. Tuchman (2008). "Autism: Definition, Neurobiology, Screening, Diagnosis." Pediatric Clinics of North America **55**(5): 1129-1146.
Reichow, B. and F. Volkmar (2010). "Social Skills Interventions for Individuals with Autism: Evaluation for Evidence-Based Practices within a Best Evidence Synthesis Framework." Journal of Autism and Developmental Disorders **40**(2): 149-166.

Reynhout, G. and M. Carter (2006). "Social Stories™ for Children with Disabilities." Journal of Autism and Developmental Disorders **36**(4): 445-469.
Rice, C. E., L. B. Adamson, E. Winner and G. G. McGee (2016). "A Cross-sectional Study of Shared Attention by Children With Autism and Typically Developing Children in an Inclusive Preschool Setting." Topics in Language Disorders **36**(3): 245-265.
Ritter, S., J. Anderson, K. Koedinger and A. Corbett (2007). "Cognitive Tutor: Applied research in mathematics education." Psychonomic bulletin & review **14**(2): 249.
Roberts, V. and R. Joiner (2007). "Investigating the efficacy of concept mapping with pupils with autistic spectrum disorder." British Journal of Special Education **34**(3): 127-135.
Robertson, A. E. and D. R. Simmons (2013). "The Relationship between Sensory Sensitivity and Autistic Traits in the General Population." Journal of Autism and Developmental Disorders **43**(4): 775-784.
Robins, B., K. Dautenhahn, R. te Boekhorst and A. Billard (2005). "Robotic Assistants in Therapy and Education of Children with Autism: Can a Small Humanoid Robot Help Encourage Social Interaction Skills?" Universal Access in the Information Society **4**(2): 20.
Robison, J., S. McQuiggan and J. Lester (2009). Modeling Task-Based vs. Affect-based Feedback Behavior in Pedagogical Agents: An Inductive Approach.
Rosenberg, N., M. Congdon, I. Schwartz and D. Kamps (2015). "Use of Say-Do Correspondence Training to Increase Generalization of Social Interaction Skills at Recess for Children with Autism Spectrum Disorder." Education and Training in Autism and Developmental Disabilities **50**(2): 213.
Rotaru, M. and D. Litman (2006). Dependencies between student state and speech recognition problems in spoken tutoring dialogues, Association for Computational Linguistics.
Rubin, E. (2007) "A Unique Mind: Learning Style Differences in Asperger's Syndrome and High-Functioning Autism." The ASHA Leader
Saadatzi, M. N., R. C. Pennington, K. C. Welch, J. H. Graham and R. E. Scott (2017). "The Use of an Autonomous Pedagogical Agent and Automatic Speech Recognition for Teaching Sight Words to Students With Autism Spectrum Disorder." Journal of Special Education Technology **0**(0): 0162643417715751.
Sallows, G. O. and T. D. Graupner (2005). "Intensive Behavioral Treatment for Children With Autism: Four-Year Outcome and Predictors." American Journal on Mental Retardation **110**(6): 417-438.
Sansosti, F. (2010). "Teaching social skills to children with autism spectrum disorders using tiers of support: A guide for school-based professionals." Psychology in the Schools **47**(3): 257-281.
Sansosti, F. J. and K. A. Powell-Smith (2008). "Using Computer-Presented Social Stories and Video Models to Increase the Social Communication Skills of Children With High-Functioning Autism Spectrum Disorders." Journal of Positive Behavior Interventions **10**(3): 162-178.
Scassellati, B. (2005). Quantitative metrics of social response for autism diagnosis.
Schilbach, L., A. Wohlschlaeger, N. Kraemer, A. Newen, N. Shah, G. Fink and K. Vogeley (2006). "Being with virtual others: Neural correlates of social interaction." Neuropsychologia **44**(5): 718-730.
Schohl, K. A., A. V. Van Hecke, A. M. Carson, B. Dolan, J. Karst and S. Stevens (2014). "A Replication and Extension of the PEERS Intervention: Examining Effects on Social Skills and Social Anxiety in Adolescents with Autism Spectrum Disorders." Journal of Autism and Developmental Disorders **44**(3): 532-545.
Schreibman, L. (2000). "Intensive Behavioral/Psychoeducational Treatments for Autism: Research Needs and Future Directions." Journal of Autism and Developmental Disorders **30**(5): 373-378.
Schuller, B., M. Wöllmer, T. Moosmayr and G. Rigoll (2009). "Recognition of noisy speech: a comparative survey of robust model architecture and feature enhancement." EURASIP Journal on Audio, Speech, and Music Processing **2009**: 1-17.
Seligman, M. E. P., R. M. Ernst, J. Gillham, K. Reivich and M. Linkins (2009). "Positive education: positive psychology and classroom interventions." Oxford Review of Education **35**(3): 293-311.

Shane, H. C., M. O'Brien and J. Sorce (2009). "Use of a Visual Graphic Language System to Support Communication for Persons on the Autism Spectrum." Perspectives on Augmentative and Alternative Communication **18**(4): 130-136.
Shaughnessy, N. and M. Trimingham (2016). Autism in the Wild: Bridging the Gap Between Experiment and Experience. The Cognitive Humanities: Embodied Mind in Literature and Culture. P. Garratt. London, Palgrave Macmillan UK**:** 191-211.
Sherer, M., K. Pierce, S. Paredes, K. Kisacky, B. Ingersoll and L. Schreibman (2001). "Enhancing Conversation Skills in Children with Autism Via Video Technology: Which Is Better," Self" or" Other" as a Model?" Behavior Modification **25**(1): 140.
Shure, M. B. (1993). "I can problem solve (ICPS): Interpersonal cognitive problem solving for young children." Early Child Development and Care **96**(1): 49-64.
Shute, V. and B. Towle (2003). "Adaptive e-learning." Educational Psychologist **38**(2): 105-114.
Silver, M. and P. Oakes (2001). "Evaluation of a New Computer Intervention to Teach People with Autism or Asperger Syndrome to Recognize and Predict Emotions in Others." Autism **5**(3): 299-316.
Singular Inversions (2017). FaceGen Modeller. Toronto, ON. Canada., Singular Inversions,. **3.5**.
Skarbez, R., G. F. Welch, F. P. Brooks and M. C. Whitton (2017). Coherence changes gaze behavior in virtual human interactions. 2017 IEEE Virtual Reality (VR).
Sklar, E. and D. Richards (2010). "Agent-based systems for human learners." The Knowledge Engineering Review **25**(02): 111-135.
Sng, C. Y., M. Carter and J. Stephenson (2014). "A Review of Video Modelling and Scripts in Teaching Conversational Skills to Individuals with Autism Spectrum Disorders." Review Journal of Autism and Developmental Disorders **1**(2): 110-123.
Snow, E., M. Jacovina, G. T. Jackson and D. McNamara (2016). iSTART-2: A Reading Comprehension and Strategy Instruction Tutor. Adaptive educational technologies for literacy instruction. S. A. Crossley and D. S. McNamara, Routledge.
South, M. and J. Rodgers (2017). "Sensory, Emotional and Cognitive Contributions to Anxiety in Autism Spectrum Disorders." Frontiers in Human Neuroscience **11**: 20.
Sparrow, S. S. (2011). Vineland Adaptive Behavior Scales. Encyclopedia of Clinical Neuropsychology. J. S. Kreutzer, J. DeLuca and B. Caplan. New York, NY, Springer New York**:** 2618-2621.
Sparrow, S. S., D. V. Cicchetti and D. A. Balla (2005a). Vineland-II: Survey Interview Form. United States of America, NCS Pearson.
Sparrow, S. S., D. V. Cicchetti and D. A. Balla (2005b). Vineland-II: Vineland Adaptive Behaviour Scales - Survey Forms Manual. United States of America, NCS Pearson.
Sparrow, S. S., D. V. Cicchetti and C. A. Saulnier (2016). Vineland Adaptive Behavior Scales, Third Edition (Vineland-3). Bloomington, MN, Pearson.
Stokes, T. F. and D. M. Baer (1977). "An implicit technology of generalization." Journal of Applied Behavior Analysis **10**(2): 349-367.
Stokes, T. F. and P. G. Osnes (1989). "An operant pursuit of generalization." Behavior Therapy **20**(3): 337-355.
Storbeck, J., N. A. Davidson, C. F. Dahl, S. Blass and E. Yung (2015). "Emotion, working memory task demands and individual differences predict behavior, cognitive effort and negative affect." Cognition and Emotion **29**(1): 95-117.
Strickland, D. (1998). Virtual reality for the treatment of autism. Virtual reality in neuro-psyco-physiology. G. Riva. Amsterdam, IOS Press**:** 209.
Stuart, I. (2004). "The impact of immediate feedback on student performance: An exploratory study in Singapore." Global perspectives on accounting education **1**(1): 1.
Sturm, D., E. Peppe and B. Ploog (2016). eMot-iCan: Design of an assessment game for emotion recognition in players with Autism. Serious Games and Applications for Health (SeGAH), 2016 IEEE International Conference on, IEEE.

Tartaro, A. and J. Cassell (2006). Authorable Virtual Peers for Autism Spectrum Disorders. Proceedings of the Workshop on Language-Enabled Educational Technology at the 17th European Conference on Artificial Intelligence (ECAI06). Riva del Garda, Italy**:** 8.
Tartaro, A. and J. Cassell (2008). Playing with Virtual Peers: Bootstrapping Contingent Discourse in Children with Autism. Proceedings of International Conference of the Learning Sciences (ICLS). Utrecht, Netherlands**:** 8.
Teague, R. J. P. (2005). Social functioning in preschool children: Can social information processing and self-regulation skills explain sex differences and play a role in preventing ongoing problems? , Griffith University.
Ten Eycke, K. D. and U. Müller (2016). "Drawing links between the autism cognitive profile and imagination: Executive function and processing bias in imaginative drawings by children with and without autism." Autism: 1362361316668293.
Truong, H. M. (2016). "Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities." Computers in Human Behavior **55**: 1185-1193.
Tsiourti, C., M. Ben-Moussa, J. Quintas, B. Loke, I. Jochem and J. Albuquerque Lopes (2016). A virtual assistive companion for older adults: design implications for a real-world application. SAI Intelligent Systems Conference.
University of Cambridge. (2017). "Inclusive Design Toolkit." Retrieved 7 July, 2017, from http://www.inclusivedesigntoolkit.com/.
van de Pol, J., M. Volman, F. Oort and J. Beishuizen (2015). "The effects of scaffolding in the classroom: support contingency and student independent working time in relation to student achievement, task effort and appreciation of support." Instructional Science **43**(5): 615-641.
VanLehn, K. (2011). "The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems." Educational Psychologist **46**(4): 197-221.
Vygotsky, L. (1978). "Interaction between learning and development." Readings on the development of children **23**(3): 34-41.
Warren, Z., Z. Zheng, S. Das, E. M. Young, A. Swanson, A. Weitlauf and N. Sarkar (2015). "Brief Report: Development of a Robotic Intervention Platform for Young Children with ASD." Journal of Autism and Developmental Disorders **45**(12): 3870-3876.
Washington, P., C. Voss, N. Haber, S. Tanaka, J. Daniels, C. Feinstein, T. Winograd and D. Wall (2016). A Wearable Social Interaction Aid for Children with Autism. Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. Santa Clara, California, USA, ACM**:** 2348-2354.
Watson, M. K., J. Pelkey, C. R. Noyes and M. O. Rodgers (2016). "Assessing Conceptual Knowledge Using Three Concept Map Scoring Methods." Journal of Engineering Education **105**(1): 118-146.
Wellman, H. M., S. Baron-Cohen, R. Caswell, J. C. Gomez, J. Swettenham, E. Toye and K. Lagattuta (2002). "Thought-Bubbles Help Children with Autism Acquire an Alternative to a Theory of Mind." Autism **6**(4): 343-363.
Werry, I., K. Dautenhahn, B. Ogden and W. Harwin (2001). Can Social Interaction Skills Be Taught by a Social Agent? The Role of a Robotic Mediator in Autism Therapy. Proceedings of the 4th International Conference on Cognitive Technology: Instruments of Mind, Springer-Verlag.
Whalen, C., D. Moss, A. B. Ilan, M. Vaupel, P. Fielding, K. Macdonald, S. Cernich and J. Symon (2010). "Efficacy of TeachTown: Basics computer-assisted intervention for the Intensive Comprehensive Autism Program in Los Angeles Unified School District." Autism **14**(3): 179-197.
White, S. W., K. Keonig and L. Scahill (2007). "Social skills development in children with autism spectrum disorders: A review of the intervention research." Journal of autism and developmental disorders **37**(10): 1858-1868.
Whitehill, J., Z. Serpell, Y. C. Lin, A. Foster and J. R. Movellan (2014). "The Faces of Engagement: Automatic Recognition of Student Engagementfrom Facial Expressions." IEEE Transactions on Affective Computing **5**(1): 86-98.

Whyte, E. M., J. M. Smyth and K. S. Scherf (2015). "Designing Serious Game Interventions for Individuals with Autism." Journal of Autism and Developmental Disorders **45**(12): 3820-3831.
Wilkins, J. (2010). The Relationship Between Social Skills and Challenging Behaviors in Children with Autism Spectrum Disorders.
Williams, J. H., D. W. Massaro, N. J. Peel, A. Bosseler and T. Suddendorf (2004). "Visual-auditory integration during speech imitation in autism." Res Dev Disabil **25**(6): 559-575.
Winner, M. G. (2005). Think Social!: A Social Thinking Curriculum for School-age Students: for Teaching Social Thinking and Related Social Skills to Students with High Funtioning Autism, Asperger Syndrome, PDD-NOS, ADHD, Nonverbal Learning Disability and for All Others in the Murky Gray Area of Social Thinking, Michelle Garcia Winner.
Wittwer, J., M. Nückles and A. Renkl (2010). "Using a Diagnosis-Based Approach to Individualize Instructional Explanations in Computer-Mediated Communication." Educational Psychology Review **22**(1): 9-23.
Wolfe, P. (2006). "The Role of Meaning and Emotion in Learning." New Directions for Adult and Continuing Education **110**: 7.
Wong, C., S. L. Odom, K. A. Hume, A. W. Cox, A. Fettig, S. Kucharczyk, M. E. Brock, J. B. Plavnick, V. P. Fleury and T. R. Schultz (2015). "Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review." Journal of autism and developmental disorders **45**(7): 1951.
Woolf, B., I. Arroyo, K. Muldner, W. Burleson, D. Cooper, R. Dolan and R. Christopherson (2010). The Effect of Motivational Learning Companions on Low Achieving Students and Students with Disabilities. Intelligent Tutoring Systems, Springer Berlin Heidlberg.
Yee, N. and J. N. Bailenson (2006). "Walk a mile in digital shoes: The impact of embodied perspective-taking on the reduction of negative stereotyping in immersive virtual environments." Proceedings of PRESENCE **24**: 26.
Zeng, Z., M. Pantic, G. I. Roisman and T. S. Huang (2009). "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions." IEEE Transactions on Pattern Analysis and Machine Intelligence **31**: 39-58.
Zhao, R., T. Sinha, A. W. Black and J. Cassell (2016). Socially-Aware Virtual Agents: Automatically Assessing Dyadic Rapport from Temporal Patterns of Behavior. IVA.