University of Plymouth

PEARL **https://pearl.plymouth.ac.uk**

04 University of Plymouth Research Theses 01 Research Theses Main Collection

2017

The Impact of Robot Tutor Social Behaviour on Children

Kennedy, James R

http://hdl.handle.net/10026.1/8758

http://dx.doi.org/10.24382/384
University of Plymouth

All content in PEARL is protected by copyright law. Author manuscripts are made available in accordance with publisher policies. Please cite only the published version using the details provided on the item record or document. In the absence of an open licence (e.g. Creative Commons), permissions for further reuse of content should be sought from the publisher or author.

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's prior consent.

THE IMPACT OF ROBOT TUTOR SOCIAL BEHAVIOUR ON CHILDREN

by

JAMES R. KENNEDY

A thesis submitted to Plymouth University
in partial fulfilment for the degree of

DOCTOR OF PHILOSOPHY

School of Computing, Electronics and Mathematics
Faculty of Science and Engineering

March 22, 2017

Abstract

THE IMPACT OF ROBOT TUTOR SOCIAL BEHAVIOUR ON CHILDREN
JAMES KENNEDY

Robotic technologies possess great potential to enter our daily lives because they have the ability to interact with our world. But our world is inherently social. Whilst humans often have a natural understanding of this complex environment, it is much more challenging for robots. The field of social Human-Robot Interaction (HRI) seeks to endow robots with the characteristics and behaviours that would allow for intuitive multimodal interaction. Education is a social process and previous research has found strong links between the social behaviour of teachers and student learning. This therefore presents a promising application opportunity for social human-robot interaction.

The thesis presented here is that a robot with tailored social behaviour will positively influence the outcomes of tutoring interactions with children and consequently lead to an increase in child learning when compared to a robot without this social behaviour. It has long been established that one-to-one tutoring provides a more effective means of learning than the current typical school classroom model (one teacher to many students). Schools increasingly supplement their teaching with technology such as tablets and laptops to offer this personalised experience, but a growing body of evidence suggests that robots lead to greater learning than other media. It is posited that this is due to the increased social presence of a robot. This work adds to the evidence that robots hold a social advantage over other technological media, and that this indeed leads to increased learning.

In addition, the work here contributes to existing knowledge by seeking to expand our understanding of how to manipulate robot social behaviour in educational interactions such that the behaviour is tailored for this purpose. To achieve this, a means of characterising social behaviour is required, as is a means of measuring the success of the behaviour for the interaction. To characterise the social behaviour of the robot, the concept of *immediacy* is taken from the human-human literature and validated for use in HRI. Greater use of immediacy behaviours is also tied to increased cognitive learning gains in humans. This can be used to predict the same effect for the use of social behaviour by a robot, with learning providing an objective measure of success for the robot behaviour given the education application.

It is found here that when implemented on a robot in tutoring scenarios, greater use of immediacy behaviours generally does tend to lead to increased learning, but a complex picture emerges. Merely the addition of more social behaviour is insufficient to increase learning; it is found that a balance should be struck between the addition of social cues, and the congruency of these cues.

Contents

Abstract **i**

Table of Contents **iii**

List of Figures **ix**

List of Tables **xiii**

Acknowledgements **xv**

Author's declaration **xvii**

1 Introduction **1**

1.1 Scope . . . 2

1.1.1 Social Behaviour . . . 2

1.1.2 Dyadic Interactions . . . 3

1.1.3 Child-Robot Interaction . . . 4

1.2 The Thesis . . . 4

1.3 Approach and Experimentation . . . 6

1.4 Key Concepts . . . 8

1.5 Contributions . . . 9

1.6 Structure . . . 10

2 Background **13**

2.1 Human-Robot Interaction . . . 13

2.2 Current Social Robot Platforms . . . 14

2.3 The Motivation for Social Robot Tutors . . . 16

2.4 The Motivation for Manipulating Social Behaviour . . . 19

2.4.1 Gestures . . . 20

2.4.2 Gaze . . . 21

2.4.3 Touch . . . 22

2.4.4 Vocal Intonation . . . 23

2.4.5 Facial Expression . . . 24
2.4.6 Proximity and Body Orientation . . . 25
2.4.7 Verbal Content Cues . . . 27
2.5 Perspectives on Social Behaviour . . . 28
2.6 Approach to Teaching and Learning for this Thesis . . . 29
2.7 Robots as Educators . . . 32
2.7.1 Roles of Robots in Education . . . 32
2.7.2 Robot Behaviour in Educational Interactions . . . 34
2.8 Characterising Social Behaviour . . . 36
2.9 Nonverbal and Verbal Immediacy . . . 37
2.9.1 Application in Human-Human Interaction . . . 38
2.9.2 Application in Human-Robot Interaction . . . 40
2.9.3 Measuring Immediacy . . . 40
2.10 Summary . . . 42

3 Materials and Methods **43**
3.1 The NAO Robot . . . 43
3.2 The Sandtray Touchscreen . . . 45
3.3 The Microsoft Kinect . . . 46
3.4 Robot Behaviour Generation . . . 48
3.5 Video Coding . . . 49
3.6 Immediacy Questionnaires . . . 51
3.7 Crowdsourcing Immediacy Ratings . . . 53
3.7.1 Quality Assurance . . . 54
3.7.2 Video and Questionnaire Format . . . 56
3.8 Practical Procedures . . . 57
3.9 Summary . . . 58

4 Nonverbal Immediacy for use with Children and Robots **59**
4.1 Hypotheses . . . 60
4.2 Experimental Setup . . . 61
4.2.1 Participants . . . 61
4.2.2 Short Story . . . 62
4.2.3 Measures . . . 62
4.2.4 Conditions . . . 63
4.2.5 Robot Behaviour . . . 65
4.2.6 Procedure . . . 65
4.3 Results . . . 66

4.3.1 Nonverbal Immediacy Results . . . 66
4.3.2 Recall Results . . . 69
4.4 Discussion . . . 70
4.4.1 Students as Individuals . . . 71
4.4.2 The Novelty Aspect . . . 72
4.4.3 Nonverbal Immediacy and Interaction . . . 73
4.4.4 Using Nonverbal Immediacy in HRI . . . 74
4.5 Summary . . . 76

5 Robot Embodiment and Child Learning **77**
5.1 Hypotheses . . . 80
5.2 Experimental Setup . . . 81
5.2.1 Participants . . . 81
5.2.2 Conditions . . . 81
5.2.3 Interaction Scenario . . . 82
5.2.4 Alien Sorting Task . . . 83
5.2.5 Robot Behaviour . . . 84
5.2.6 The Learning Task . . . 86
5.2.7 Measuring Learning . . . 87
5.2.8 Interaction Scenario . . . 88
5.2.9 Video Data . . . 89
5.3 Learning Results . . . 90
5.3.1 Learning Bias . . . 92
5.3.2 Accounting for Bias . . . 94
5.4 Behaviour and Embodiment Results . . . 95
5.4.1 Compliance . . . 96
5.4.2 Gaze and Embodiment . . . 97
5.4.3 Gaze and Robot Behaviour . . . 99
5.5 Discussion . . . 100
5.5.1 Embodiment and Learning . . . 100
5.5.2 Social Behaviour and Embodiment . . . 101
5.5.3 Task Characteristics . . . 104
5.5.4 Nonverbal Immediacy . . . 106
5.5.5 Sample Size . . . 106
5.6 Summary . . . 107

6 Robot Social Behaviour and Child Learning **109**
6.1 Hypotheses . . . 111

6.2 Experimental Setup . . . 111
6.2.1 Participants . . . 112
6.2.2 Interaction Scenario . . . 112
6.2.3 Conditions . . . 116
6.2.4 Robot Behaviour . . . 117
6.2.5 Procedure . . . 119
6.3 Results . . . 119
6.3.1 Learning from Lessons . . . 120
6.3.2 Robot Presence . . . 120
6.3.3 Social Condition . . . 121
6.4 Discussion . . . 124
6.4.1 Child Perception and Ability . . . 126
6.4.2 Gender Differences . . . 127
6.4.3 Nonverbal Immediacy . . . 128
6.5 Summary . . . 128

7 Robot Nonverbal Immediacy and Child Learning 131
7.1 Hypotheses . . . 132
7.2 Experimental Setup . . . 133
7.2.1 Participants . . . 133
7.2.2 Interaction Protocol . . . 134
7.2.3 Robot Behaviour . . . 136
7.3 Results . . . 136
7.3.1 Learning Gains . . . 136
7.3.2 Questionnaire Data . . . 138
7.3.3 Gaze Analysis . . . 139
7.4 Discussion . . . 140
7.4.1 Adult Immediacy Ratings . . . 141
7.5 Summary . . . 142

8 Robot Verbal Immediacy and Child Learning 143
8.1 Language Learning and Social Robots . . . 144
8.2 Hypothesis . . . 147
8.3 Experimental Setup . . . 148
8.4 Evaluation . . . 151
8.4.1 Participants . . . 151
8.4.2 Measures . . . 151
8.4.3 Conditions and Robot Behaviour . . . 152

8.4.4 Procedure . . . 153
8.5 Results . . . 155
8.5.1 Perception of the Robot . . . 155
8.5.2 Learning Gains . . . 156
8.6 Discussion . . . 159
8.7 Summary . . . 162

9 Synthesis of a Framework for Learning and Social HRI 165
9.1 Human Prime Tutoring . . . 166
9.1.1 Methodology . . . 166
9.1.2 Results . . . 167
9.2 Nonverbal Immediacy and Learning . . . 168
9.2.1 Timing of Social Cues . . . 170
9.2.2 Adaptation of Social Cues . . . 171
9.2.3 Relative Importance of Social Cues . . . 172
9.2.4 Novelty of Character and Behaviour . . . 173
9.2.5 (In)congruency of Social Cues . . . 174
9.3 Summary . . . 179

10 Discussion 181
10.1 Experimental Limitations . . . 181
10.1.1 Ecological Validity and Generalisability . . . 182
10.1.2 Measures of Learning . . . 183
10.1.3 Robot Platform . . . 184
10.2 Ethical Questions . . . 186
10.3 Educator and Societal Acceptability . . . 188
10.4 Sociality and Learning . . . 191
10.5 Comparisons Between Children and Adults . . . 194
10.6 Future Work . . . 195
10.6.1 Building on the Social Cue and Congruency Model . . . 196
10.6.2 Accounting for Individual Differences . . . 197
10.6.3 Increasing Interactivity . . . 197
10.6.4 The Robot Role . . . 198
10.6.5 Sustained Use and Adoption . . . 198
10.7 Summary . . . 199

11 Contribution and Conclusion 201
11.1 Summary . . . 201
11.2 Contributions . . . 203

11.3 Conclusion 204

Glossary **207**

Acronyms **209**

Appendices **211**

A Example Video Coding Manual **212**

B Crowdsourced Immediacy Results **221**

C Chapter 4 Short Story Script **223**

D Chapter 4 Short Story Recall Questionnaire **225**

E Robot Nonverbal Immediacy Questionnaire **227**

F Child Nonverbal Immediacy Questionnaire **231**

G Chapter 5 Robot Script **235**

H Robot Immediacy Questionnaire **237**

I Robot Relationship Questionnaire **241**

J Chapter 8 French Language Test **243**

Bibliography **247**

List of Figures

2.1 A depiction of the role of social interaction for an individual, with two possible outcomes: social performance and learning performance - adapted from Kreijns et al. (2003) 19

2.2 The revised educational objectives 'Taxonomy Table' (adapted from Krathwohl, 2002). Crosses indicate the areas focused on in studies throughout the research here, with the red cross signifying the intersection at which performance is most often measured. 30

3.1 The Aldebaran NAO used in the majority of the evaluations throughout this thesis. 44

3.2 Version 1 of the Sandtray touchscreen used in experimental evaluations. A laptop is stored within the wooden housing to run the software, with the touchscreen used as a display. 45

3.3 Version 2 of the Sandtray touchscreen used in experimental evaluations. This version is much improved, with a sturdier yet smaller construction, larger screen, lighter weight, and built-in computer. 45

3.4 Version 2 of the Sandtray touchscreen used in experimental evaluations, with the Microsoft Kinect hardware extension. The Kinect is mounted on an additional lightweight metal frame, at a fixed distance from the robot. The robot stands with a raised rectangle on the base board. This board helps with robot stability (it allows the feet to slide more easily than the carpets often present in evaluation environments), and prevents the robot from turning away from the children when performing gestures (a common problem without fixing the feet in place). 47

3.5 Microsoft Kinect application used to process data from the Kinect sensor and to communicate with the Urbi server running on the robot. This figure shows the Kinect with the visualisation of the camera stream on, but this is turned off when evaluations are running as the application cannot be seen, and turning off the visual rendering substantially reduces CPU load. . . . 48

3.6 Screenshot of the ANVIL software package, showing *clockwise* from the *top-left*: (1) the menu, (2) the video player, (3) annotation panel, (4) coding/annotation time-line. Segments can be created on the time-line to indicate when behaviours occur. 51

3.7 Screenshot from the online crowdsourcing service used to gather adult nonverbal immediacy ratings. Radio boxes are used for answers to each question. The questionnaire is the same as that shown in Appendix E, but with the language switched to be an observer report, rather than self-report (i.e., "you" is changed to "the child"). 55

3.8 Screenshot from the online crowdsourcing service used to gather adult nonverbal immediacy ratings. The video is embedded in the page, without controls. Nonverbal immediacy questions are not revealed until the number from the end of the video is entered, at which point the video is removed from the page. 56

4.1 Still images from the conditions used in the evaluation; *left* to *right*: (1) low nonverbal immediacy robot, (2) high nonverbal immediacy robot, (3) human. Red backgrounds for the robot were not used in practice and are just used to ease visibility here; the video was shown in widescreen format, with a black background covering the unused space, as in the figure. . . . 62

4.2 Robot nonverbal immediacy scores as rated by children and adults, relating to hypothesis H2. Both children and adults perceive the difference in nonverbal immediacy between the two robot conditions. ** indicates significance at the $p<.01$ level, and *** indiciates significance at the $p<.001$ level. *Error bars* show the 95% Confidence Interval 68

4.3 Recall scores for high and low nonverbal immediacy robot conditions relating to hypothesis H1. Children recall significantly more information when the story is read by a robot with higher nonverbal immediacy. * indicates significance at the $p<.05$ level. *Error bars* show the 95% Confidence Interval 70

4.4 Representation of the role of social cues in dyadic HRI. Social cues are used as modulation behaviour within the interaction. 73

5.1 Side-by-side images of the virtual (*left*) and real (*right*) robots used for this study: the Aldebaran Nao. The images are stills taken from one of the cameras used for filming. Scaling has been kept consistent between the two images so that size comparisons can be made. 82

5.2 Schematic overview of the mediation device-centred interactions under investigation in this chapter. Two interactants (the child and the robot) face one another over the touchscreen. Two video cameras record the interactants during the studies. The experimenters are also in the room, but out of view of the child. Figure not to scale. 83

5.3 From *left* to *right*: (1) the 'orange planet' category image, (2) an outline of one of the aliens with shading to differentiate the body parts which can be manipulated (an actual image has not been used due to copyright restrictions), (3) the 'purple planet' category image 84

5.4 Calculated bias value for each of the images used in the experimental pre- and post-tests using Equation 5.1. 94

5.5 Pre-test and post-test percentage of correct answers for images, grouped by bias type and direction. *Error bars* show the 95% Confidence Interval. . . . 95

5.6 Child responses to robot suggestions to make a move. Children clearly comply with the robot's requests. 96

5.7 Seconds per minute that the child spends gazing towards the robot, split by learning phase third, comparing embodiment conditions. *Error bars* show 95% Confidence Interval. 98

5.8 Seconds per minute that the child spends gazing towards the real and virtual robots, split by interaction segment. *Error bars* show 95% Confidence Interval. 100

6.1 Schematic overview of the interactions under investigation in this paper. Two interactants (the child and the robot) face one another over the touchscreen. Two video cameras record the interactants during the studies. A Microsoft Kinect tracks the child's face. Two experimenters are in the room, but out of view of the child. Figure not to scale. 113

6.2 Example of the sorting task used. This is a screenshot of one of the tests used in the experiment. Children can touch a number, drag it over the 'prime' or 'not prime' label and release to make a categorisation. The number will then shrink and move into the boxes beside the category label. 114

6.3 Structure of the task used in the interactions, showing robot lesson positions.114

6.4 Snapshots taken from the video recordings of interactions. Both the social (*left*, looking at the child) and asocial robot (*right*, actively avoiding the gaze of the child) conditions are pictured to show the difference in gaze behaviour between them. 116

6.5 Improvement between division pass one and division pass two in percent for the division only and screen only conditions. Significantly greater improvement occurred in the 'screen only' condition (where division lessons were present) when compared to the 'division only' condition (where division lessons were not present), indicating that the lessons have a significant, positive impact on child division. *Error bars* show 95% Confidence Interval, ** indicates significance at the $p<.01$ level. 120

6.6 Pre-test and post-test scores for the asocial robot, social robot and screen only conditions. *Error bars* show 95% Confidence Interval, * indicates significance at the $p<.05$ level. 122

6.7 Child gaze towards the robot in seconds per minute, split by robot condition. *Error bars* show 95% Confidence Interval, * indicates significance at the $p<.05$ level. 123

6.8 Post-questionnaire responses of the children when asked what they thought the robot was like. Eight options, including 'teacher' were available. . . . 124

7.1 Structure of the task used in the interactions, showing robot lesson positions (identical to Chapter 6 Figure 6.3). 134

7.2 A snapshot from the 'high' nonverbal immediacy condition. 134

7.3 A snapshot from the 'low' nonverbal immediacy condition. 135

7.4 Pre- and post-test scores on recognising prime numbers for the low and high nonverbal immediacy (NVI) conditions (*left*); NVI scores for the designed low and high NVI conditions (*right*). Children improve more in recognising prime numbers when taught by a high immediacy robot. *Error bars* show 95% CI. 138

8.1 A child answering a question on screen during the interaction. 148

8.2 Screenshot from the touchscreen showing a question. Children can touch a word, drag it to the blank space and release to answer. Here the correct answer being 'Portugal'. 149

8.3 Structure of the task. *R* refers to robot explanation sections and *C* refers to child question answering sections. The robot dictates the structure of the interaction through speech and by presenting questions on the touchscreen, informing the child of when it is their turn answer questions on the screen. The HIGH condition includes many manipulations in the verbal behaviour to make it more 'available'. 150

8.4 Schematic overview of the interactions being investigated in this paper. The child and the Aldebaran NAO robot sit across a touchscreen from one another. An experimenter sits behind and out of view of the child. Two video cameras record the interaction. Figure not to scale. 154

8.5 Verbal and nonverbal immediacy scores for the high immediacy (HIGH) and low immediacy robot (LOW) conditions. The HIGH condition is perceived to have significantly higher verbal immediacy while having the same nonverbal immediacy. *Error bars* show 95% CI. 155

8.6 Pre-test, post-test and retention test scores by condition (chance score=3; maximum score=12). Children learn a significant amount from the robot between pre- and post-tests; this gain is sustained to the retention test. *Error bars* show 95% CI. Darker dashed line indicates the 'chance' baseline. . . . 158

9.1 Pre-test and post-test scores for prime number learning of children when tutored by a human. The children improve significantly from the pre-test to the post-test. *Error bars* show 95% Confidence Interval. 168

9.2 Nonverbal immediacy scores as judged by adults and learning effect sizes for the prime number task. The dotted green line indicates a trend towards greater nonverbal immediacy of the tutor leading to increased learning. *Error bars* show 95% Confidence Interval. 169

9.3 Nonverbal immediacy scores as judged by the children in the interaction and learning effect sizes for the prime number task. The dotted green line indicates a trend towards greater perceived nonverbal immediacy of the tutor leading to increased learning. *Error bars* show 95% Confidence Interval.170

9.4 Guttman's Lambda 6 against learning effect size for each of the prime tutoring conditions. The dotted line indicates a trend towards greater internal consistency (measured through G6) leading to greater learning. . 176

9.5 Learning, congruency and social behaviour for each of the 5 conditions. Learning is measured in effect size between pre- and post-test for children. Congruency is indicated through Guttman's Lambda 6 of the adult nonverbal immediacy scores. Social behaviour is characterised through nonverbal immediacy ratings from adults. An interactive version of this figure is available online to provide different perspectives of the space: `https://goo.gl/ZNPxc8`. 177

10.1 The revised educational objectives 'Taxonomy Table' (adapted from Krathwohl, 2002). Crosses indicate the areas focused on in studies throughout the research here, with the red cross signifying the intersection at which performance is most often measured. Reproduced from Chapter 2. 192

List of Tables

1.1 Experiments conducted as part of this thesis, detailing location, chapter and findings summary. Subject numbers are post-exclusion figures. 7

4.1 Operationalisation of behavioural manipulations between robot immediacy conditions 64

4.2 Mean nonverbal immediacy scores by condition 68

4.3 Mean recall scores by condition 70

5.1 Inter-coder agreement by track coded. 90

5.2 Learning effect *t*-test results, comparing many different variables. 'Gain' refers to the increase in score between pre- and post-tests. The maximum score for all conditions is 12. * indicates a significant p value at the .05 level. All distributions do not significantly deviate from normality (Kolmogorov-Smirnov test; $p > .05$). For unpaired tests, observations are independent. All measures are continuous. 91

5.3 Learning effect Mann-Whitney U test results grouped by bias. All observations are independent, but some distributions deviate significantly from normality (Kolmogorov-Smirnov test; $p < .05$), making the U test more efficient than *t*-tests. Average percent correct is shown for both pre- and post-tests. * indicates a significant p value at the .05 level. 95

7.1 Robot behaviour for high and low nonverbal immediacy (NVI) conditions. 136

8.1 A summary of studies conducted in HRI to investigate different aspects of robots on child language learning. *L1* indicates where English was being taught to native speakers, whereas *L2* refers to cases where English was being taught as a foreign language. 145

8.2 Test score results for pre-test, post-test and retention tests, by condition. . 158

9.1 Guttman's G6 and learning effect size by condition. G6 is used as an indicator of social cue congruency, with a higher value indicating greater congruency between cues. 176

B.1 Adult crowdsourced nonverbal immediacy ratings for robot behaviours, including the chapter that the behaviour is evaluated in, and basic demographic details. 221

Acknowledgements

This work would not have been completed (or at least would not have been nearly as fun to complete) without the fantastic people around me. First and foremost, a great deal of thanks goes to my supervisory team: Tony Belpaeme, Paul Baxter and Angelo Cangelosi. To Angelo, for creating a vibrant group in Plymouth with whom I have shared the last 4 years. To Paul, for your unparalleled support and far too many other things to list, but dark corners and a passion for linguistic obfuscation are a highlight. And to Tony, for so many life changing opportunities, your wit and ineluctable wisdom.

Thank you to all of the ABC group in Plymouth past and present, of whom there are too many to name without forgetting someone, so I won't even try! In particular, thank you to those who have shared an office with me and made the arduous commute worth it: Paul Baxter, Robin Read, Joachim de Greeff, Emily Ashurst, Emmanuel Senft, Séverin Lemaignan, Fotios Papadopoulos, Bahar Irfan, and Thomas Colin.

The international HRI community has of course shaped my work in many ways. Thank you to anyone who I have discussed my work or ideas with at conferences, workshops, project meetings, summer schools, and so on. I have learnt so much not only academically, but about different places and cultures, which has been an indescribably positive and transformative experience. The wonderful people in the EU ALIZ-E, L2TOR and DREAM projects: you have been a pleasure to collaborate with and I look forward to the remaining time we shall have together in DREAM and L2TOR.

The support of my family has been unwavering, as always. I guess I should thank my dad specifically for finding the advert for the PhD studentship and suggesting that I might like to chance an application to do robotics despite not knowing anything about robots. Somehow that didn't seem to matter.

Dedicated to the memory of R.A.D. Kennedy.

I know how proud you were when I started this journey; I wish you could have seen the end.

Author's declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award. Work submitted for this research degree at Plymouth University has not formed part of any other degree either at Plymouth University or at another establishment.

This work has been carried out by James Kennedy under the supervision of Prof. Dr. Tony Belpaeme, Prof. Dr. Angelo Cangelosi, and Dr. Paul Baxter. The work was funded by a studentship from Plymouth University School of Computing, Electronics and Mathematics, and the author has also received funding from the European Union FP7 projects ALIZ-E (grant no.: 248116) and DREAM (grant no.: 611391), and the Horizon 2020 project L2TOR (grant no.: 688014).

Parts of this thesis have been published by the author:

Kennedy, J., Baxter, P., & Belpaeme, T. (2013). Constraining Content in Mediated Unstructured Social Interactions: Studies in the Wild. In *Proceedings of the 5th International Workshop on Affective Interaction in Naturalistic Environments (AFFINE'13), at ACII'13*, (pp. 728–733). IEEE Computer Society

Kennedy, J., Baxter, P., & Belpaeme, T. (2014a). Children Comply with a Robot's Indirect Requests. In *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 198–199)

Kennedy, J., Baxter, P., & Belpaeme, T. (2015a). Can Less be More? The Impact of Robot Social Behaviour on Human Learning. In *Proceedings of the 4th International Symposium on New Frontiers in HRI at AISB 2015*

Kennedy, J., Baxter, P., & Belpaeme, T. (2015b). Comparing Robot Embodiments in a Guided Discovery Learning Interaction with Children. *International Journal of Social Robotics*, *7*(2), 293–308

Kennedy, J., Baxter, P., & Belpaeme, T. (2015c). Head Pose Estimation is an Inadequate Replacement for Eye Gaze in Child-Robot Interaction. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, (pp. 35–36). ACM

Kennedy, J., Baxter, P., & Belpaeme, T. (2015d). The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 67–74). ACM

Kennedy, J., Baxter, P., Senft, E., & Belpaeme, T. (2015e). Higher Nonverbal Immediacy Leads to Greater Learning Gains in Child-Robot Tutoring Interactions. In *Proceedings of the International Conference on Social Robotics*, (pp. 327–336). Springer

Kennedy, J., Baxter, P., Senft, E., & Belpaeme, T. (2015f). Using Immediacy to Characterise Robot Social Behaviour in Child-Robot Interactions. In *Proceedings of the 1st Workshop on*

Evaluating Child-Robot Interaction at ICSR 2015

Kennedy, J., Baxter, P., Senft, E., & Belpaeme, T. (2016a). Heart vs Hard Drive: Children Learn More From a Human Tutor Than a Social Robot. In *Proceedings of the 11th ACM/IEEE Conference on Human-Robot Interaction*, (pp. 451–452). ACM

Kennedy, J., Baxter, P., Senft, E., & Belpaeme, T. (2016b). Social Robot Tutoring for Child Second Language Learning. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 67–74). ACM

Kennedy, J., Lemaignan, S., & Belpaeme, T. (2016c). The Cautious Attitude of Teachers Towards Social Robots in Schools. In *Proceedings of the Robots 4 Learning Workshop at RO-MAN 2016*

Kennedy, J., Baxter, P., & Belpaeme, T. (2017). Nonverbal Immediacy as a Characterisation of Social Behaviour for Human-Robot Interaction. *International Journal of Social Robotics*, *9*(1), 109–128

Kennedy, J., Baxter, P., & Belpaeme, T. (under review). The Impact of Robot Tutor Nonverbal Social Behaviour on Child Learning. *Frontiers in ICT: Human-Media Interaction*

Word count for the main body of this thesis: **62,911**

Signed: _____

Date: _____

Chapter 1

Introduction

Human-Robot Interaction (HRI) is concerned with 'understanding, designing and evaluating robotic systems for use by or with humans' (Goodrich & Schultz, 2007). As a consequence, the field of research combines several disciplines (Fong et al., 2003). Research directions are influenced by findings and methodologies from cognitive science, artificial intelligence, robotics, and social science (amongst many others; Baxter et al., 2016). The work in this thesis approaches HRI from a social perspective. This perspective considers the design and impact of robots in social settings; Breazeal (2002) provides a succinct description of such machines:

> *"A sociable robot is able to communicate and interact with, understand and even relate to us, in a personal way."*

This entails that the robot should be able to understand humans (and itself) in social terms. It is this distinction of a sociable robot, the understanding of humans in social terms, that this thesis aims to contribute to, as opposed to a social robot that is purely *evocative*, *receptive*, or an *interface* (Breazeal, 2003). Whilst there are many technical and theoretical challenges to this approach, the rewards have great potential value. The *Social Intelligence Hypothesis* (or *Social Brain Hypothesis*; Dunbar, 2002) posits that the complex social world has helped to shape how human intelligence has evolved (Holekamp, 2007). Indeed, human children have similar abilities for dealing with the physical world as chimpanzees, but have superior cognitive skills for dealing with the social world (Herrmann et al., 2007). By tapping into our *social brains*, social robots seek to elicit reactions and interactions which would not be possible without social behaviour; they are *socially evocative* (Fong et al., 2003). Such robots rely on the human tendency to anthropormorphise (Reeves &

Nass, 1996), which seems to happen without mindful thought (Kim & Sundar, 2012).

It could be argued that the domain of child education is inherently social (Bandura & McClelland, 1977; Vygotsky, 1980). From an early age, children use social signals to select who to learn from (Birch et al., 2008; Wu & Kirkham, 2010), and it has been suggested that social interaction is *essential* for certain aspects of child learning (Kuhl, 2007b). However, we still have relatively little understanding of how learning actually takes place (VanLehn et al., 2003), and what impact combinations of multimodal social cues have on learning in complex settings (Roth & Lawless, 2002). The work herein seeks to explore the social behaviour that a robot can use to influence interactions. The intention is to discover how children respond to robot social behaviour in real-world environments, and how the behaviour of a robot can positively affect child behaviour as measured, in part, through learning outcomes of such interactions.

1.1 Scope

As HRI lies at the intersection of many fields of research, and the application domain of child learning is large in itself, it is necessary to more tightly define the scope of the work undertaken as part of this thesis. The following subsections seek to specify the focus of the research conducted, accompanied by justifications for any restrictions in the scope where necessary.

1.1.1 Social Behaviour

The emphasis throughout the research conducted here is from the social HRI perspective. The aim is to provide an account for how to develop robot social behaviour for use with children in educational interactions, rather than to investigate effective teaching strategies when using robots. Child learning and social responses provide useful metrics for evaluating the impact of robot behaviour manipulations, and the effect of varying robot social behaviour will be reflected upon using these measures throughout. As such, in the majority of experiments, the teaching strategy is consistent, whilst robot social behaviour is varied (although there is occasionally overlap between the two and it is acknowledged that the teaching strategy must be explored sufficiently to ensure that learning is possible within the interactions).

Exploration of social behaviour within the field of HRI is useful as it can teach us not only about how we interact with machines, but also about human psychology. The field of Human-Computer Interaction (HCI) has been established for a longer period than that of Human-Robot Interaction, but has many similarities (indeed HRI could be seen as a sub-level of HCI research if a robot is considered a specific type of computer, or a parallel field if not). It is suggested that HCI relies on understanding and modelling humans, and in particular, social aspects of human behaviour, for example: 'attention', 'perception', and 'communication' (Hewett et al., 1992). This reliance translates to HRI as well, and makes study of social behaviour in these contexts worthwhile (human development is embedded within a social context; Rogoff, 1990), as it can further our understanding of human psychology (Morse et al., 2011). To provide a concrete example of where this has been achieved, developmental HRI has assisted in evaluating models of language acquisition in an interaction (Morse et al., 2015). Study of social behaviour in HRI is therefore a potentially interesting and valuable direction of research in itself, which is why it forms the primary basis of exploration in this thesis.

Given the application domain of child learning, measurements of learning are often used to form part of the evaluation of implemented robot behaviours, but the aim is not to develop optimal teaching strategies. Teaching strategies are commonly specific to particular tasks, however, social behaviour is common throughout all learning tasks. It is an understanding of this more generalisable social behaviour that the work here aims to make a contribution to. As such, the learning task itself can be varied between experimental scenarios depending on the requirements of the aspect(s) of social behaviour under consideration. It is however recognised that social behaviour is likely to be constrained by the specific context (Kennedy et al., 2013), and so the interaction context is often kept consistent, even if the educational content is not. The distinction between HRI research exploring social behaviour in the context of educational interactions (the approach taken here), and research aiming to develop the best teaching strategies for robots will be expanded upon in Chapter 2.

1.1.2 Dyadic Interactions

Dyadic interactions between one robot and one child will be the primary context for the work here. This decision was made for several reasons. Firstly, significant advantages have

been demonstrated in terms of child learning in one-to-one scenarios, making it a clearly relevant application domain (Bloom, 1984; VanLehn, 2011); this will be discussed further in Chapter 2. Secondly, many manipulations of social behaviour rely on some form of personalisation (Belpaeme et al., 2012; Syrdal et al., 2007a). Multi-party interactions would add a layer of complexity as decisions may need to be made about who to personalise to and when, which is not a trivial task in itself (Leite et al., 2013a). Additionally, many technical challenges arise which rely on robust perception (for example, to tell which child responds to a question). This technology is not yet reliable enough for use in naturalistic environments, particularly with children (Belpaeme et al., 2013). Finally, the dynamics of the interaction are also highly likely to change when an additional child is present (Leite et al., 2015). This dynamic may then rely more on the relationship between the children, diminishing the role that the robot can assume, which could hinder the exploration of particular effects, or the ability to attribute any findings to the behaviour of the robot.

1.1.3 Child-Robot Interaction

Children aged 6 to 9 years old will be the target group for evaluating robot behaviours with. Children of this age present a greater opportunity to take advantage of *suspension of disbelief*, where judgements of implausibility are suspended (Coleridge, 1817; Duffy & Zawieska, 2012). It has been shown that children will readily suspend disbelief in an interaction with a social robot, which can bring about advantages to researchers, such as children overlooking minor technical problems (e.g., slow speech response) or being more willing to treat the robot as a social character (Belpaeme et al., 2012, 2013). This decision also allows the work here to build on the expertise of the European projects (ALIZ-E and DREAM) which this work is conducted alongside, as these age ranges overlap. Of course, this also requires careful managing from an ethical perspective (Kahn Jr et al., 2004; Sharkey, 2015), which will be elaborated on in Chapter 3.

1.2 The Thesis

The main thesis that this document seeks to put forward is as below.

> A robot with tailored social behaviour will positively influence the outcomes of tutoring interactions with children and consequently lead to an increase in

child learning when compared to a robot without this social behaviour.

Additional research questions are also introduced here. These research questions are used to support and direct the experimental research conducted in pursuit of demonstrating the primary thesis.

- **What advantages (if any) do robots offer in terms of learning outcomes when compared to other technological media?**

  A growing body of evidence suggests that the socially evocative aspect of robots (as discussed at the start of this chapter and in Fong et al., 2003; further described in Section 2.3) leads to benefits in interactions when compared to other media. However, relatively few rigorous studies have confirmed these findings in relation to a positive impact on child learning outcomes. Whilst the main thesis is concerned with tailoring robot behaviour, the case for using a robot must first be justified.

- **How do children perceive and respond to robot social behaviour in educational interactions?**

  As established in the scope for this work (Section 1.1.3), children will suspend disbelief when interacting with robots, but it is unclear how they will perceive and interpret social behaviour of robotic characters, particularly in a context in which robots are not commonplace. The thesis assumes that interactions will form and that robot social behaviour will influence child behaviour; this research question aims to establish whether this is indeed the case.

- **How do the verbal and nonverbal behaviours of a social robot influence learning outcomes?**

  Robots can employ social behaviours across a multitude of modalities. Section 1.1.2 highlighted personalisation of social behaviour as a common means of behavioural manipulation. Such personalisation could be performed through verbal or nonverbal channels, through one or more social cues. Building up a picture of how social cues combine to influence human behaviour is of fundamental interest to not only the field of HRI, but also psychology (Section 1.1.1).

- **How can the perception of robot social behaviour be characterised?**

  Some form of characterisation of robot social behaviour is required in order to readily compare between different behaviours, or sets of behaviours, thus enabling

a more explicit link between social behaviour characterisations and outcomes, such as learning. Moving beyond descriptions of specific behavioural implementations, and towards how behaviours are actually perceived by humans may provide greater insight into the effects of those behaviours (Section 1.1.1).

1.3 Approach and Experimentation

The thesis and the research questions are explored in this document through a series of experimental evaluations. First, previous work is discussed with the aim of establishing what is already known in the field of HRI, specifically in terms of the principles of designing social behaviour for learning interactions (mainly through Human-Human Interaction HHI literature). This review of the literature also helps in designing methodologies for assessing child learning and robot social behaviour. The literature review (Chapter 2) reveals that there is no established means of characterising robot social behaviour, nor an agreed upon technique or principle by which social behaviour should be designed and implemented. This led to the formulation of the research questions above.

Child perception and responses to robot social behaviour are considered throughout the experimental evaluations performed in Chapters 4 through 8 (summarised in Table 1.1). This is done through a variety of means, including video coding of both robot and child social behaviour, child subject surveys and adult observer surveys. These methods are described in Chapter 3. The advantages of robots and effects of embodiment are explored through consideration of prior literature, and furthered here through experiments that compare a physically present robot to a variety of different control conditions, including a virtual form of the same robot, and having the same material presented on screen. This research question is addressed in Chapters 5 and 6.

Learning outcomes of children in response to different robot social behaviours are explored experimentally throughout, with many findings presented in chapters with evaluations. These findings are all brought together and contextualised in accordance with the robot social behaviour through a common means of characterisation in Chapter 9. This chapter discusses all of the findings and provides some insight into why the findings may have been found, whilst also relating the observations here to prior and contemporary literature.

Date	Site	Chapter	Subjects	Summary
18/12/2012	Braunton CAEN primary		27	Study for sandtray interaction style; published in Kennedy et al. (2013) and used to inform subsequent experiments.
12/04/2013	Plymouth University	5	2	Human behaviour acquisition for use as robot model.
10/07/2013-11/07/2013	Salisbury Road primary	5	28	Evidence that a physically present robot makes a difference to child behaviour in educational interactions.
08/05/2014	Braunton CAEN primary	6	6	Human behaviour acquisition/pilot.
17/06/2014-23/06/2014	Okehampton primary	6	45	Evidence that a physically present robot leads to more learning in interactions, but that care must be taken with social behaviour implementation.
01/12/2014	Braunton CAEN primary	4	83	Child data. First application of nonverbal immediacy on a robot in a lecture-based interaction. Children recall more information from a robot with high nonverbal immediacy.
08/12/2014	Plymouth University/online	4	31	Adult data. Children and adults perceive robots with differing levels of nonverbal immediacy as intended, rating the behaviour in a similar manner.
11/05/2015-13/05/2015	Stuart Road primary	7	23	Higher robot nonverbal immediacy leads to more child learning in dyadic interactions.
06/07/2015-10/07/2015	Widey Court primary	8	67	Children can learn aspects of a second language from a robot. The verbal immediacy of the robot does not lead to learning differences.
13/07/2015	Widey Court primary	8	67	Retention test follow-up. Children retain their learning from the robot.
22/07/2015	Braunton CAEN primary	9	11	Human primes data collection.
>31/07/2015	Online		229	Immediacy ratings for all conditions. Used to support each experimental chapter and to produce a model of social behaviour for robots in Chapter 9.

Table 1.1: Experiments conducted as part of this thesis, detailing location, chapter and findings summary. Subject numbers are post-exclusion figures.

1.4 Key Concepts

This section seeks to provide working definitions, or pointers to such definitions, for key concepts used throughout this thesis.

- **Immediacy** - is a concept from human psychology and communication literature that was introduced in the 1960's by Mehrabian (1968) and is defined as the 'psychological availability' of an interaction partner. Immediacy is further introduced as being a measure that indicates "the attitude of a communicator toward his addressee", and in a general form "the extent to which communication behaviours enhance closeness to and nonverbal interaction with another" Mehrabian (1968). A number of specific social behaviours are listed (touching, distance, forward lean, eye contact, and body orientation) to form part of this measure, which were later utilised by researchers that sought to create and validate measuring instruments for immediacy. Immediacy can be broken down into verbal and nonverbal aspects. An exploration of the immediacy literature can be seen in Section 2.9.

- **Congruency** - when used in the context of social behaviour, congruency refers to the extent to which social cues are aligned to one another. This is not just in terms of cue timing, but also the quantity of social cues being used. For example, if two social cues are both used regularly and appropriately, then they would be *congruent*, whereas they would be *incongruent* if one was used regularly and another not at all.

- **Tailored** - used in reference to the thesis of this work, 'tailored' describes the intentional design of social behaviours such that they are adapted specifically for the application domain of educational interactions.

- **Learning** - for a full discussion of definitions of learning, please see Section 2.6. Throughout the thesis, learning will often refer specifically to cognitive learning gains that require the application of newly acquired knowledge to a novel problem, i.e., more than mere recall of information.

1.5 Contributions

The original contributions of this thesis in the context of the research questions are outlined below. Where appropriate, chapters and published papers in which these contributions are made will be referred to. Table 1.1 shows the experiments conducted as part of this work, along with the chapters which they are related to.

- Adaptation and development of the Child Nonverbal Immediacy Questionnaire (CNIQ), the Robot Nonverbal Immediacy Questionnaire (RNIQ), and the Robot Immediacy Questionnaire (RIQ): scales for use with children for characterising verbal and nonverbal social behaviour of humans and robots (Chapters 3, 4, 8; published in Kennedy et al., 2016b, and Kennedy et al., 2017).

- Validation of the CNIQ and RNIQ with children and adults for humans and robots (Chapter 4 and Kennedy et al., 2017).

- Evidence (in addition to prior work by other researchers) for the advantage of robots on child social responses, particularly in gaze towards the robot, when compared to virtual agents (Chapter 5 and Kennedy et al., 2015b).

- Evidence for the advantage of robots on learning outcomes with findings showing that children learn more when a robot is present than when only a touchscreen is used, regardless of having the same lesson content (Chapter 6 and Kennedy et al., 2015d).

- Findings to show that a robot with tailored nonverbal social behaviour leads to greater child learning (Chapters 7 and 9, and Kennedy et al., 2015e).

- In short-term interactions, verbal social behaviour of a robot does not seem to improve the learning of children, nor the retention of this learning (Chapter 8 and Kennedy et al., 2016b).

- A model of the relationship between robot social behaviour and child learning. This model incorporates the quantity and congruence of social cues, where a combination of a high number of social cues with high congruency will lead to maximal possible learning (Chapter 9 and Kennedy et al., under review).

The work presented in this thesis has been conducted alongside two European Union FP7 projects: ALIZ-E and DREAM. The work has contributed to these projects and some of the initial code resources were developed by researchers of the ALIZ-E project (not by the author). Instances where this was the case will be made clear in the text throughout. Many of the findings from the work presented in this thesis were also used extensively in the successful proposal for the L2TOR Horizon 2020 project.

1.6 Structure

The structure of this thesis is outlined below to provide an overview of the content and context for each chapter. A summary of key experimental findings are included at the start of each relevant chapter for ease of reference.

- This chapter provided an introduction to the general field of this research (robot tutors for children), the research questions including the central *thesis*, scope, and contributions of the work presented in later chapters.
- Chapter 2 provides a background for the research, including the motivation for the scenario and application to HRI. Various work from the literature considering robot tutors, social behaviour, and learning is discussed, identifying gaps in knowledge which this thesis aims to address.
- Chapter 3 describes the hardware tools used to conduct the research, and the development of measurement instruments based on *immediacy* for characterising social behaviour. These resources are used throughout the experimental studies presented in subsequent chapters. The procedure used to gather immediacy ratings from adults for all of the experimental conditions used throughout the various studies is also described.
- Chapter 4 presents an experiment conducted to validate the use of a modified nonverbal immediacy questionnaire for use with children and adults, for observing humans and robots. This allows a stronger connection between the immediacy literature which is mainly concerned with adults, and the scenario with robots and children used here.
- Chapter 5 presents the findings of an experiment in which children are guided

through a two-category sorting task by either a virtual, or real (physical) robot. Significant learning differences are not found, but there is evidence to suggest that children respond differently in their social behaviour depending on which of the robots they see.

- Chapter 6 examines the impact of multimodal social behaviour and personalisation on child learning through the use of two different 'sociality' conditions that are based on human tutor behaviour. Two further conditions which omit the use of a robot and lessons about the learning material are used to validate the study and to draw comparisons with other claims from the literature.

- Chapter 7 extends the experiment conducted in Chapter 6 by using the same methodology, but basing behavioural manipulations explicitly on aspects of behaviour measured through the nonverbal immediacy scale. It is found that a robot with higher nonverbal immediacy leads to significantly improved child learning.

- Chapter 8 continues the experimental exploration of the impact of social behaviour on learning by comparing two different robot conditions in a language learning task. The verbal content of the robot's speech is varied to produce high and low verbal immediacy conditions. It is found that the manipulations did not significantly affect the learning of the children.

- Chapter 9 describes an additional experiment to collect data and brings together findings on the impact of multimodal robot social cues on child learning into one framework provided by nonverbal immediacy.

- Chapter 10 draws on the experimental work from previous chapters, alongside the context supplied by related work, to form a discussion about the broader context and findings of the thesis. Limitations of the work conducted here are outlined, leading to suggestions for future directions of research.

- Chapter 11 concludes the thesis with a summary of the main contributions.

Chapter 2

Background

This chapter provides a theoretical background for the thesis and the experimental work conducted in later chapters. A broad overview of HRI is provided before the more specific topics pertaining to this thesis are introduced. The application scenario of using a robot to tutor children is motivated through a consideration of social behaviour, its influence in human-human tutoring, and comparisons to tutoring with technology. Learning is defined for the purpose of this thesis, clarifying its usage and connecting the work here to research in pedagogy. Prior work exploring the use of robots as tutors and in educational environments is examined, revealing areas where potential contributions to current knowledge can be made.

Whilst the effectiveness of social robots in educational environments has been demonstrated, it remains unclear how social behaviour influences learning in HRI. A complex picture emerges from the literature of robots used for tutoring when social behaviour is varied, and comparisons between studies are challenging to draw. It is apparent that a characterisation of the robot social behaviour would help to clarify the differences between studies and provide a means by which certain factors could be accounted for in analysis. To this end, an overview of characterising social behaviour is also provided here.

2.1 Human-Robot Interaction

Whilst there are a large number of robots in use throughout the world, the majority are not in the public domain where they are required to be used by the general population. This is due in part to the tasks that robots currently execute; a robot that assists in car manufacture

does not need to interact directly with humans. However, certain applications will require robots to become *social* in order to comprehend social signals used by humans, to express intentions to humans, and to communicate effectively. This perspective gives rise to the field of HRI, or more specifically in this case, social HRI.

As part of this field, an increasing amount of research is being conducted into making robots social for the purpose of allowing access to new application domains, such as in schools (Baxter et al., 2015), care homes (Broekens et al., 2009), and hospitals (Coninx et al., 2016). A growing body of evidence shows that when a social robot is used in such applications, the outcomes tend to improve (Li, 2015). For example, a robot tutor seems to lead to faster puzzle completion times when compared to a screen presenting the same information (Leyzberg et al., 2012). This may in part be due to the Social Intelligence Hypothesis as introduced in Chapter 1, or it may be exploiting another human inclination, such as the tendency to treat inanimate objects (like computers or robots) in a social manner (Reeves & Nass, 1996). Either way, such effects are commonly observed, and part of the motivation for HRI research is to explore social responses and interaction outcomes further in order to provide a greater understanding of the underlying processes involved in social interaction. As HRI research commonly builds upon findings or approaches from other fields (Baxter et al., 2016), the following section will consider the motivation for robot tutors through the exploration of HRI and human-human interaction (HHI) literature.

2.2 Current Social Robot Platforms

When using robots in experiments, particularly with children, platform selection is an important topic given issues such as the Uncanny Valley effect where a robot could be perceived as 'eerie' if it is not designed appropriately (Moore, 2012; Mori et al., 2012). Many social robot platforms are currently humanoid in form, or at least have aspects of human features. These platforms will be explored here, with a focus on the range of available options for this thesis, and where such platforms sit in the wider robotics context.

An increasing number of humanoid platforms are being developed, particularly in light of the recent DARPA challenges (Destephe et al., 2015). These robots are highly sophisticated from a mechanical perspective, and often look explicitly like machines with motors and components exposed. One prominent example of such a platform is Atlas from Boston

Dynamics[1]. However, such a robot is clearly not an appropriate tool for interacting with children due to its size and appearance, which children would likely find intimidating. For this purpose, it is preferable to move to a part of the robot design space with smaller robots, that are less physically advanced, but are more focussed on a pleasant physical appearance (achieved by, for example, hiding mechanical components with plastic coating). For social HRI, physical capability is also not as strong of a requirement as it is for the manual tasks involved in domains like the DARPA challenges. However, the ability to generate multimodal social cues is desirable so that social cue use can be inspired from the human-human literature, and to provide a range of possible social cue manipulations.

A limited number of small, social-oriented robots are currently commercially available (this number was even fewer when the program of research described here commenced). Currently one of the most used platforms in social HRI research environments is the Aldebaran NAO (e.g., Ramachandran et al. 2016; Tanaka et al. 2007; Zaga et al. 2015). The NAO is a small (58cm tall) humanoid robot with relatively limited physical manipulation abilities, and a fixed face, but a simple appearance, unlikely to fall into the Uncanny Valley. The humanoid form is somewhat similar to the Sony QRIO that had previously been used in research environments (Tanaka & Matsuzoe, 2012), but was never commercially released. The advantage of this small form is that it is approachable to children, but when sufficiently social, it appears that it is not treated as a toy by children (Tanaka et al., 2007).

One limiting factor for social behaviour research is the lack of a face that can move and create facial expressions. One solution to this problem is to use a mobile device in place of a physical face and to animate the screen, as employed by DragonBot (Setapen, 2012) and its successor Tega (Park et al., 2017). This provides a nice solution for facial expressions, however neither of these robots are commercially available, and nor are they humanoid in appearance. This potentially inhibits the applicability of findings in human-human studies to these forms; for example, gesturing using human-form arms may be perceived differently when compared to a wing of a zoomorphic robot. This is a similar problem for the KeepOn robot that has been used in tutoring scenarios, e.g., Leyzberg et al. (2012).

An alternative to using a mobile device to generate facial expressions is to use a retro-projected head, such as in LightHead (Delaunay et al., 2009) or Furhat (Al Moubayed et al., 2012). Whilst this solution is convincing in terms of generating facial expressions for a social robot, the necessary face size means that these robots are already fairly large.

[1] `http://www.bostondynamics.com/robot_Atlas.html`

To provide a body to scale, a much larger machine is required, which again prevents comfortable use with children.

The NAO robot provides a balance between the humanoid form, and the size necessary for use with children. Whilst there are limitations in facial expression production (a point that will be returned to in Section 10.1.3), there are no platforms currently fulfilling all of the criteria for multimodal social behaviour generation in a form factor appropriate to use with children. This may change in the near future given the increasing number of robots of this size currently under design, incorporating screens as faces, e.g., the Buddy robot[2] or the QT robot (Ziafati et al., 2017).

2.3 The Motivation for Social Robot Tutors

The '2-Sigma Problem' (Bloom, 1984) asked how we might make classroom teaching as effective as one-to-one tutoring. However, tutoring is consistently found to be more effective than group education (VanLehn, 2011). As such, a more appropriate question may be to ask how we can increase the provision of one-to-one tutoring. It would be ideal if humans could be used to provide this tutoring, but this is not a practical solution (due to a lack of trained instructors and financial resources). This presents an opportunity for technological solutions such as Intelligent Tutoring Systems (ITS) which nearly perform as well as humans in tutoring scenarios (VanLehn, 2011). The field of ITS has aimed to fulfil the tutoring need, but research has revealed that the social presence of physical robots may lead to improved learning outcomes. This section will introduce HRI studies that compare the use of a social robot with other media, with a particular focus on educational contexts and task performance.

Some compelling results show that the physical presence of a robot can have a positive impact on task performance. Adult participants completed a logic puzzle in a significantly faster time when given advice by a physical robot as opposed to when they received no advice, the advice without the robot present, or the advice from the robot on a screen (Leyzberg et al., 2012). Similarly, adults earnt a higher score in a negotiation game when interacting with a robot compared to a simulation of the same robot (Bartneck, 2003). This suggests that merely the physical presence of a robot can have a large impact on human behaviour, possibly due to the social presence theory (Biocca et al., 2003) and a

[2] http://www.bluefrogrobotics.com/en/buddy/

social facilitation effect (Triplett, 1898; Zajonc, 1965). This effect posits that people will perform better, or at least differently, in a task when other social actors are observing them (Uziel, 2007) and also relates to the Hawthorne effect (Landsberger, 1958) where people modify their behaviour when they are aware they are being observed. Social facilitation effects have been demonstrated in HRI (Riether et al., 2012). Other social presence effects with robots have been found where people are significantly more likely to comply with peculiar requests from a colocated, real robot than a virtual robot, and also to prefer the interaction (Bainbridge et al., 2008, 2011). Evidence also suggests that people are more likely to comply with requests from a robot for longer than requests from other media (Kidd, 2008).

Positive effects of robots have also been found in child-robot interaction. Although there was not a comparison to an equivalent agent, Alemi et al. (2014) demonstrate that adding a robot to language education (as well as the regular human teacher) can lead to significant improvement in child learning over the human teacher alone. This work was conducted over 5 weeks, with a total of 10 sessions, thus providing some evidence for the ability of positive learning effects persisting outside of short-term interactions. The same experiment[3] found that the addition of robots also reduced the anxiety of students about speaking a foreign language and improved their attitude towards the learning material (Alemi et al., 2015).

Several other studies have found that children learn more from robots than from paper-based or screen-based instruction. Han et al. (2005) suggest that children learn more, concentrate more, and show more interest when content is delivered by a robot, although the measurement of learning is unclear. An expanded version of this study (Han et al., 2008) finds the same results, but reveals that the concentration analysis is quite coarse (a single data point per 10 minutes), and that pre-existing knowledge was not explicitly controlled for. Hyun et al. (2008) find that children improve significantly more in their English language skills when taught by a robot compared to screen-based instruction, however it should be noted that the experiment was not time restricted. This resulted in children spending approximately 5 times as long interacting with the robot, so the learning results are likely a product of exposure, but clearly there was a difference in the motivation of the children to interact. Such differences are often attributed to a novelty

[3]It is not explicitly stated in the paper that the experiment is the same one, but given identical methodologies, participant numbers and authors, it can reasonably be assumed

effect, e.g., Kanda et al. (2004).

Kose-Bagci et al. (2009) find that children perform better in a drumming game when interacting with a robot present and visible, when compared to the robot not being visible, or a virtual robot. Children also prefer the interaction with the physical robot present. The authors suggest that although there are challenges in creating complex autonomous behaviour for robots (presumably in comparison to virtual characters), there is a "need for physically embodied interaction in suitable scenarios" (Kose-Bagci et al., 2009).

However, not all results show a positive effect from using robots. There are also suggestions that the social presence of a real robot could act as a distraction and lead to reduced recall in a conversation (Powers et al., 2007). Wainer et al. (2007) found no significant difference between the performance of participants completing a logic puzzle whether they were coached by a physical robot, a robot on screen, or a robot simulation, although the robot in this case was not anthropomorphic, so social effects may have been minimised.

Other work involving children in educational contexts has not revealed learning effects due to the presence of a robot. Kory Westlund et al. (2015) observed a preference of children to be taught new words by a robot, but did not observe significant learning differences whether children were taught by a human, a tablet, or a robot. This is likely due to the small quantity of data available for comparison due to the methodology used (2 learned words compared between each condition). Looije et al. (2012) used an Aldebaran NAO robot in virtual and real form with children to study the effect of physical presence. While behavioural differences in increased gaze towards the real robot were found, no significant learning differences were seen, but this is likely due to the number of subjects (8) being too low for such statistical analysis.

In summary, studies with adults and children both show potential for the advantage of social robots as a means of improving learning or task performance. However, many of the studies with children do not tightly control the learning aspects and measurements, or do not find conclusive results. There is sufficient evidence to motivate exploring a social robot for education (e.g., Alemi et al., 2014; Han et al., 2008), but the utility of a physical robot is highlighted as a potential area where further research would be beneficial. This thesis makes a contribution to this line of research in Chapters 5 and 6.

2.4 The Motivation for Manipulating Social Behaviour

The previous section demonstrated that using robots can confer greater advantages in interactions when compared to virtual characters or other types of media. Being physically present in the world can lead robots to having greater perceived social presence, which has the potential to lead to improved social interaction between robots and humans (Jung & Lee, 2004; Wainer et al., 2007). In HHI it has been posited that greater learning gains occur in *more social* interactions because social behaviours are thought to increase a learners' interest (Atkinson et al., 2005).

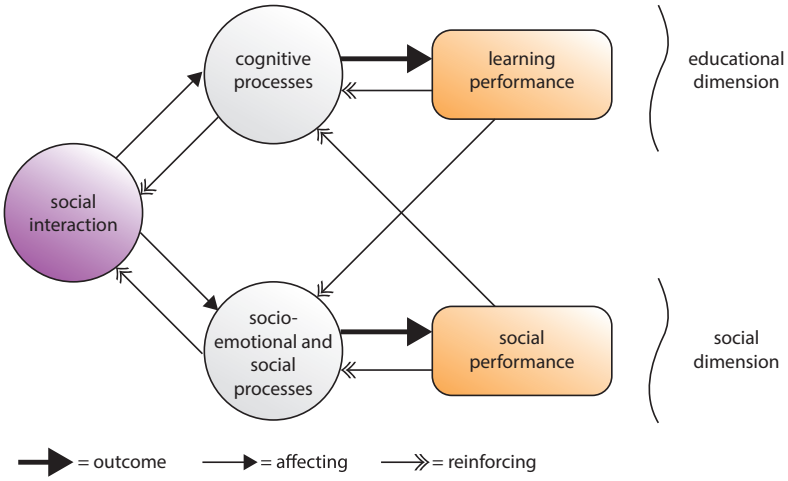

Figure 2.1: A depiction of the role of social interaction for an individual, with two possible outcomes: social performance and learning performance - adapted from Kreijns et al. (2003)

Social interaction can be considered as the bond between cognitive processes and socio-emotional processes (Kreijns et al., 2003). The outcome of such interactions can be measured through social performance or learning performance, either of which can in turn reinforce the cognitive or socio-emotional processes taking place in an individual (Figure 2.1). This concept is supported through definitions of learning, which can be broken down into 'affective' and 'cognitive' learning (Bloom et al., 1956; this is further expanded on in Section 2.6). Social interaction has the ability to influence both of these learning elements, and indeed HRI researchers have sought to do just this. Some researchers have focussed on the social behaviour of the robot with the aim of influencing cognitive processes (Kennedy et al., 2015e; Szafir & Mutlu, 2012), whereas others have sought to influence the socio-emotional processes to a greater extent (Castellano et al., 2013).

Social interaction will rely in large part on the social cues used. Between humans, social

cues used by teachers are often found to influence learning and a robot's ability to utilise similar cues could be an advantage over virtual technologies. If the cues are appropriate, then there is more likelihood of a fluid interaction and for the robot to have greater social presence. As such, the impact of social cues in HHI will now be explored to provide a background for the motivation of manipulating robot social behaviour in the context of education and learning tasks. Where effects seen in the HHI literature have been shown in HRI contexts, this will also be indicated.

2.4.1 Gestures

Gestures play an important role in teaching and learning (Kelly et al., 2008; Macedonia & von Kriegstein, 2012). Children are more likely to repeat the speech of a teacher if a matching (non-symbolic) gesture accompanies the speech when compared to the same speech without a gesture, but less likely with a mismatched gesture compared to no gesture (Goldin-Meadow & Wagner, 2005; Goldin-Meadow et al., 1999). This basic recall is a first step towards learning. Furthermore, these studies show that children can use gestures in understanding problem-solving strategies, giving them the potential to learn both through problem solving and how to approach solving problems.

For young children, it has been suggested that gesture use (specifically symbolic gestures) can facilitate cognition (Goodwyn & Acredolo, 1998); possibly because gestures can lighten cognitive load, lending more resources to memory tasks (Goldin-Meadow et al., 2001). Indeed when children are slightly older (aged 8-10) gestures can help learning to 'last' for longer, with correct answers in an algebra follow-up test four weeks after a learning session staying higher in a gesture and speech condition than in a speech only condition (Cook et al., 2008). Equally, gestures made by children can be used to assess their learning (Goldin-Meadow et al., 1992), with adults able to be more certain of their judgements of children's learning when their gestures matched their verbal explanation.

Such findings are reinforced in studies concerning instructional communication for learning, with children's performance improving more when given instructions with gestures as opposed to without in a symmetry recognition test (Valenzeno et al., 2003). These findings seem to have been partially replicated in HRI, with a robot utilising contingent gesturing leading to increased recall of material from a presentation (Szafir & Mutlu, 2012). However, precisely how to use gestures to influence learning in HHI is an open field with

many questions still necessitating futher exploration (Roth, 2001); this is even more true for HRI where less work examining the use of gesture and learning has been conducted.

The use of hands seems to be particularly important. It is not just the orienting of attention, such as with a laser pointer, but the fact that the gesture is done with a hand that leads to an improvement in learning (Rumme et al., 2008). It has been shown that humans can accurately interpret pointing by a humanoid robot (an Aldebaran NAO), but that for best results, the arm on the side which the object to be pointed at should be used (Wang et al., 2014). However, whether the hand of robot has the same attentional and learning impact as that of a human is not known. It has also been established that being present (as opposed to on video) does not affect how much attention gestures draw between humans (Gullberg & Holmqvist, 2002), but no such study comparing humans and robots could be found.

2.4.2 Gaze

From an early age, children use social cues such as eye gaze to help direct their learning. Despite social cues distracting briefly from the material to be learnt, infants learn more with gaze cues present than when their learning is not directed by such cues (Wu et al., 2010). These positive effects have also been successfully implemented in computational models (Yu & Ballard, 2007). Even at 15 months old, children have a tendency to use the gaze of a social interaction partner, instead of distracting and erroneous saliency cues for word learning associations (Houston-Price et al., 2006). The power of gaze, or even just the eyes, in influencing behaviour is still observed in adults, with surprisingly strong results. For example, just an image of eyes near a donation point can increase charitable donations by almost 50% (Powell et al., 2012).

Selective processing of social cues for learning has far-reaching implications for HRI. Head movement alongside eye gaze can assist humans in responding to robot cues (Boucher et al., 2010); use of this social cue could have advantages in learning. However, this has not been found in infants learning from robots, where they follow the gaze direction of both a robot and a human, but only the human gaze facilitated the learning of an object's appearance (Okumura et al., 2013). It was suggested that this could signify a disposition of infants to consider humans a superior source for learning. It remains to be seen whether this holds true for slightly older children, or with children more familiar with the concept

of robots. Equally, this result could be a demonstration that humans process robot gaze in a cognitively different manner, as argued in Admoni et al. (2011).

College students who receive gaze at the start of each sentence when receiving verbal information can recall significantly more than those who receive no gaze (Sherwood, 1987). This holds true for both simple and difficult material, for both genders. It is hypothesised that this is because the interaction feels more 'intimate' and prevents mind-wandering whilst receiving the information. These findings have also been shown to occur with younger children, aged between 6 and 7 (Otteson & Otteson, 1979). Greater gaze from a storyteller led to increased recall from children when subsequently asked questions, compared to those in a lesser (but still some) gaze condition. This study reveals a trend towards possible interaction effects between the information content, gender and gaze, speculating that females are less affected by gaze than males when the material is more difficult.

Logically, it follows that using appropriate robot gaze towards a child might be beneficial for recall and learning. Work done in virtual environments demonstrates that caution must be used, as simply staring at a human interactant actually reduces their willingness to engage in mutual gaze, despite increased opportunity (Dalzel-Job et al., 2011). It should be noted that this difference in mutual gaze did not actually translate to a difference in task performance, but this was hypothesised as being due to the relative simplicity of the task. A similar effect has been observed in human-robot interaction studies as well (Kennedy et al., 2015b; Looije et al., 2012), where a tutoring robot received more gaze, which could theoretically be beneficial for the learning, but no learning differences were found.

Nevertheless, gaze can clearly have positive effects on learning (Otteson & Otteson, 1979; Sherwood, 1987; Okumura et al., 2013; Yu & Ballard, 2007), but if it is not meaningful, or is too abundant then it can discourage mutual gaze, thereby limiting potentially positive effects (Dalzel-Job et al., 2011). This remains a challenge, as it is not trivial to decide how much gaze is 'just right', or precisely when a gaze fixation should be made by a robot.

2.4.3 Touch

Touch has been shown to lead to a positive affective state in HHI, even with very short touches and when subjects were unaware of the touch (Fisher et al., 1976). This positive response to touch has also been shown in HRI. When a robot offered an 'unfair proposal'

to participants with touch, their EEG response showed less negativity towards the robot than when the robot did not touch as they made the proposal (Fukuda et al., 2012). To place this work in the context of educational interactions under consideration here, liking does not necessarily result in better learning, but there are indications that if students like an instructor more they will achieve more highly (Gurung & Vespia, 2007).

Touch has also been linked with compliance (Guéguen, 2002), a useful tool for teachers when they need to influence students in order to get them to engage with lessons. The potential for utilising touch in HRI and educational contexts has previously been highlighted (Salter et al., 2006) but, as yet, remains underexplored. More generally, it has been demonstrated in several experiments that touch can influence perceptions of robots by humans. A robot initiating touch can encourage humans to see them more positively and to work for longer at a monotonous task (Shiomi et al., 2016). However, Cramer et al. (2009) emphasise that care is needed when incorporating behaviours such as touch in interactions as combined effects can lead to negative outcomes. Social cue combinations have also been found by Chen et al. (2011) between touch and verbal warnings; these types of effects will be discussed further in Section 2.5.

2.4.4 Vocal Intonation

The voice that an agent uses can dictate how much they are liked and how hard humans try to understand the material they are presented with (Atkinson et al., 2005). Those who interacted with an agent who had a human voice preferred the agent and also did better in learning transfer tests when compared to those who interacted with the same agent with a machine-synthesised voice. The sound of a voice can have a significant impact on retention and transfer of a novel subject when presented through narration (Mayer et al., 2003). Retention is better when a voice has a 'standard' (as opposed to foreign) accent and is human rather than machine-like, as well as being more likeable in both cases.

However, this result was found with college students and virtual agents. It has not been established whether this effect is also observed outside of this restricted demographic, nor whether specific embodiments of robots create expectations that violate these rules. For example, it may be less appropriate to have a deep male human voice when using a robot such as the Aldebaran Nao[4] than a RoboThespian[5]. It is suggested that a possible uncanny

[4]https://www.aldebaran.com/en/humanoid-robot/nao-robot
[5]https://www.engineeredarts.co.uk/robothespian/

valley effect (Moore, 2012; Mori et al., 2012) may occur, where participant expectations are violated when a human voice is played alongside a not-convincing-enough animated agent. An indication in this direction has been found with virtual agents, where participants preferred an animated agent with a machine-like voice and a non-animated agent with a human voice (Baylor et al., 2003).

Vocal intensity can also be used to influence learning. Compliance, a factor in learning, can be increased through raising vocal intensity, as in Remland & Jones (1994). This HHI study was conducted in a public space where compliance was greatest when using a medium level of vocal intensity; around 70dB. It is likely that this level would need adjusting depending on the ambient noise in the space a robot tutor would be acting in, and how far from a student it would be. Vocal intensity has successfully been combined with gestures in a model which is based on nonverbal immediacy to improve attention and recall of a human in an HRI presentation scenario (Szafir & Mutlu, 2012). Whilst not confirming all of the results discussed in this section relating to vocal prosody, it certainly demonstrates that there is great potential for many of the same principles from HHI being applied to HRI.

Interestingly, speech rate does not appear to have a significant impact on recall (Simonds et al., 2006). This could potentially be explained by the capacity of humans for speech. The average human speech rate is 125-150 words per minute, but adults have twice as much cognitive capacity, being able to process speech at 250-300 words per minute (Fulford, 1992). This gives a broad scope for increasing speech rate without any great change in terms of the listener's cognitive processing. There is also a distinction between recall of spoken information and understanding, which may have a different outcome when speech rate is varied.

2.4.5 Facial Expression

Human studies have shown that recognition of emotion from facial expression is used from a very young age to help understand social events and regulate social interactions (Tronick, 1989; Zarbatany & Lamb, 1985). In an HHI study examining the relationship between the social cues and cognitive learning across a number of different cultures it was found that alongside gaze and vocal prosody, smiling from the teacher was one of the more strongly correlated cues to student learning (McCroskey et al., 1996). This result has

also been replicated more recently (Velez & Cano, 2008), additionally showing the positive relationship between social cues and motivation (with facial expressions having a large effect size).

Experimental data from human-computer interaction (HCI) with an embodied conversational agent revealed no significant difference in recall of subjects when interacting with an agent which was either neutral, or able to express joy and anger (Becker-Asano et al., 2013). Several reasons are put forward as to why this may have been the case, including a ceiling effect within the task, the amount each emotion was displayed, or that the facial expressions were simply ignored in favour of focussing on the task. As such, it is unclear whether the benefits of facial expression seen in HHI will translate to HCI and HRI.

Despite the suggested impact of facial expressions on learning or motivation in HHI, no data could be found regarding the impact of learning and facial expressions of robots. A possible explanation is that much of the research to-date regarding learning in HRI is performed with robots such as the Aldebaran NAO, Keepon, and Wakamaru which have largely non-manipulable faces. Due to the movement required in expressing facial emotion, the uncanny valley (Mori et al., 2012) could also be a current limitation for robots.

2.4.6 Proximity and Body Orientation

The proximity between interactants is correlated to compliance effects (Peters, 2007). It is suggested that a distance of 1-2 feet (30-60cm) is optimally conducive to compliance between humans (from studies conducted in Western cultures; Segrin, 1993), however whether this is the same for HRI has not been established. This is possibly because judging the physical proximity at which a robot should be from a student would not necessarily be as simple as a strict 1-2 feet rule. In human interactions, verbal feedback can modulate (positively and negatively) the proxemic impact on compliance (Greene, 1977). In HRI, comfortable distances are dictated through the complex interplay of factors such as the size of the robot (Hiroi & Ito, 2011), how much the robot gazes towards a human and how likeable they previously perceive the robot to be (Kim & Mutlu, 2014).

Only about 60% of people conform to the same proxemic social norms with robots as they do with people (Walters et al., 2005a). That being said, compliance effects have been seen in educational interactions between children and robots at a distance of about 2 feet (60cm), although this hasn't been compared against a control with closer or further

distances (Kennedy et al., 2014a). Additionally, it would appear that younger children have a smaller personal space, presumably due to their smaller size, so further work would need to be done for people of different sizes (Aiello & Aiello, 1974).

Research conducted with a robot in a variety of task contexts show humans generally prefer the robot to be 0.46-1.22 metres away (Huettenrauch et al., 2006). However, it is warned that the dynamic nature of interaction with a robot should not necessarily be reduced to a simplistic rule. Indeed, the previous paragraph suggested the impact of variable robot appearance and behaviour, but there are also environmental and task factors to consider. For instance, if it is important to hear speech in a noisy environment, then it might be that a closer distance between interaction partners is more comfortable, when outside of these parameters it would usually not be.

Several design guidelines for robotic proximity are presented in Takayama & Pantofaru (2009). It is suggested that people who are familiar to the robot can be approached more closely, to direct gaze away from the face of a human as an approach is made, and to factor in the human's attitude towards robots when maintaining distance. The impact of human attitude towards robots is further supported experimentally in Mumm & Mutlu (2011) where the necessity of building rapport before increasing closeness is emphasised. This could be an important factor in tutoring in order to gain compliance.

Studies directly examining the impact of body orientation on learning could not be found; this is possibly due to the entanglement of body orientation with many other social cues. If not orientated to an interaction partner only limited eye gaze will be possible, gestures may be occluded and it may be more difficult to hear any speech. Nor could any studies be found studying the specific impact of co-located physical proximity on learning; most work considers co-located learning against distance learning (not co-located), but this then becomes about social presence rather than proxemics. Logically, it would seem reasonable that a middle-ground should be sought. The robot should not be too far away as then the student may struggle to perceive verbal instructions and non-verbal signals. If more compliance is required, then a closer distance should be sought. Further research is required to decide what is to be considered 'too close' in specific scenarios, with humans of certain ages and certain robot sizes/designs; work such as Rae et al. (2013); Walters et al. (2005a) provides a strong starting point in this direction.

2.4.7 Verbal Content Cues

Social aspects of the verbal content that a teacher uses can influence student learning (Gorham, 1988; Witt et al., 2004). These social aspects can include: personalisation, discussion outside of lesson content, asking questions, revealing personal information, and the type of language used (possessive or not, e.g., 'our' or 'the'). Research has been done in HRI with a view to improving the bond between children and robots through some of these means (Belpaeme et al., 2012), but often this is not in the context of educational interactions. The exploration of verbal cues has been explored from various angles in HRI with regards to human perception (for example, Andrist et al., 2013), but not so much in respect to learning outcomes. With children, it has been found that 'off-activity talk' - dialogue with a robot which does not concern the task being completed - encourages compliance in a therapeutic setting (Kruijff-Korbayova et al., 2014). Personalisation in therapeutic contexts has also been considered. Children were asked a number of questions about their preferences and the robot then mentioned these in an interaction, the children who interacted with a personalised robot enjoyed the interaction more, but subject numbers were too low for statistical comparisons (Blanson Henkemans et al., 2013).

When broadening to consider the impact of verbal content on learning in HHI and other technology based fields (such as HCI), a much greater quantity of research exists. Mayer et al. (2004) found differences in learning when an animated character used language in an explanation to students that was conversational, compared to when it was more formal. This was a very subtle change, of just changing the word 'the' to 'your' throughout descriptions. Students retained equivalent information in both conditions, but could transfer their knowledge significantly better when the explanation had used the conversational ('your') style.

A summary of tens of studies which consider learning in response to variations in the cues described at the start of this section can be seen in Witt et al. (2004). All of these studies deal with HHI, but have a mix of measures and subject matters. It is found that there are large correlations between greater use of the verbal cues and perceived student learning (r=.49), however the cases where cognitive learning is actually measured reveals a much smaller correlation (r=.06). Differences between these measurements will be discussed in greater detail in Section 2.6. These findings suggest that verbal cues play a much smaller role in learning than nonverbal cues (which have almost 3 times as large a correlation in

the same meta-analysis).

2.5 Perspectives on Social Behaviour

Social cues do not occur in isolation, neither from other cues, nor from the environment and the interaction they are being used in. Behaviour is multimodal, and cues used must be congruent with other social cues being utilised in order to be interpreted correctly and efficiently. Social cues could be perceived as a single percept, which requires that cues be considered as an integrated whole (Zaki, 2013).

These concepts are exemplified experimentally by Byrd et al. (2014) who further explored the conclusions drawn from studies such as those done by Cook et al. (2008) regarding gestures and learning (discussed previously in Section 2.4.1). They found that when children did not copy eye movements accompanying gestures the lasting learning effect disappears. Similar results have been found elsewhere. In Langton (2000); Langton & Bruce (2000), head gaze, gestures and spoken words were all used to direct attention. When any of the cues were incongruent (e.g., responses had to be made to head-gazes, whilst a pointing gesture was made in a different direction), interference effects were found, slowing down responses. If social cues are not synchronous and congruent then interactions will likely be impeded by this additional processing time.

In HRI there is often a tendency to manipulate only one behavioural cue at a time in order to tightly control experiments. This approach allows the findings to be directly attributed to the cue that is manipulated, but conclusions are often made which advise how this cue should be used without consideration for the larger context that this cue was used in. For example, if gestures are found to improve recall (as in Szafir & Mutlu, 2012) when mutual gaze is often made, it is unclear if the same effect on recall would be observed if mutual gaze is no longer made. Zaki (2013) posits that perceptions (and subsequently outcomes) are a product of the combined gaze and gesture cues used, so the findings for gestures would depend on the gaze model used. Indeed, such interaction effects between behavioural cues have been found in HRI (previously discussed in Section 2.4.3; Chen et al., 2011; Cramer et al., 2009).

Where larger scale manipulations have been made, some studies have sought to tease out the influence of particular cues in order to make more specific recommendations.

Huang & Mutlu (2013) use statistical techniques to derive the gestural predictors for interaction outcomes such as information recall. However, it is unclear how this could be re-implemented as a behaviour and whether the findings would remain should other social cues change. Not just the cues being used, but also their contingency can influence interactions. A robot which displays more contingent social cues, such as appropriate gaze and pointing gestures in response to a human, can elicit greater participation in an interaction (Lohan et al., 2012). When applied to an educational context, it is reasonable to suggest that greater participation will lead to an increase in learning (Anderson, 1975).

Based on the evidence put forward in this section, there is a need for further research in evaluating robot behaviour as a unified set of cues, i.e., as the product of a combination of social cues, rather than as separable cue elements additively forming a behaviour. The work in this thesis contributes to such research in Chapters 6, 7, and 8. As a consequence of this approach, conclusions will often not be drawn about specific social cues, but about sets of cues forming higher-level behaviours. This also means that analysis of robot behaviour needs to consider social cues in the context of one another; this challenge will be addressed in Section 2.8.

Furthermore, evidence suggests that social behaviour will vary depending on the environment that interactions take place in (Ros et al., 2011; Salter et al., 2008). As such, if a study applied to a domain is to have ecological validity, then it ought to be conducted in the environment in which the interaction would actually take place in, or at a minimum with 'experimental realism' (Berkowitz & Donnerstein, 1982). In the case of educational interactions, this might mean a school, or a home, depending on the specific application. Conducting experiments in naturalistic environments, or 'the Wild', introduces a number of methodological challenges (Ros et al., 2011), but they are worth the additional effort if the results have more ecological validity. Given the application of dyadic tutoring interactions being studied throughout this work, all experiments take place in children's schools.

2.6 Approach to Teaching and Learning for this Thesis

It is important to note that throughout the work presented here, the focus is on the social behaviour of the robot, rather than on the higher level teaching strategy. This is necessitated through the focus on the research questions and scope laid out in Chapter 1.

Nonetheless, an understanding of definitions and processes of learning are required to use learning as a metric in studies, so these will be explored in this section with the goal of describing the position taken in the subsequent research presented. When considering the tutoring literature it is important to distinguish between different intended meanings of the term 'learning'. Learning can be broken down into several different domains and stages, with various researchers attempting to provide a taxonomy of learning to formalise these concepts (Bloom et al., 1956; Krathwohl et al., 1964). Learning can also be approached in many different ways, with a variety of adopted perspectives determining how learning, and teaching, are carried out.

Krathwohl (2002) developed a revised taxonomy of educational objectives based on the original version by Bloom et al. (1956). The aim of the original taxonomy was to provide not only a measurement tool for education, but also to provide a common language for communicating about learning and this is maintained in the revised version Krathwohl (2002). This feature makes it ideally placed for use in defining learning for this thesis. In the revised taxonomy there are two dimensions: 'cognitive process' and 'knowledge'. Each dimension has several stages, which form a loose hierarchy (Figure 2.2).

		Cognitive Process dimension					
		Remember	Understand	Apply	Analyse	Evaluate	Create
Knowledge dimension	Factual	X	X	X			
	Conceptual	X	X	X			
	Procedural	X	X	X			
	Meta-cognitive						

Figure 2.2: The revised educational objectives 'Taxonomy Table' (adapted from Krathwohl, 2002). Crosses indicate the areas focused on in studies throughout the research here, with the red cross signifying the intersection at which performance is most often measured.

The knowledge dimension consists of *factual*, *conceptual*, *procedural* and *meta-cognitive* knowledge types. Briefly, these can be defined as follows (Krathwohl, 2002):

- *Factual* - terminology or details required to work in a discipline
- *Conceptual* - category, principle and generalisation knowledge
- *Procedural* - skill, algorithm, technique and method knowledge for subject-specific tasks
- *Meta-cognitive* - knowledge of cognitive processes and one's own cognition

For the work in this thesis, the goal is for children to learn a new skill and to generalise this skill to novel stimuli. As such, this addresses the following cognitive educational objectives: *remember*, *understand*, and *apply*. *Apply* is at the highest level of the hierarchy of these, and so this will constitute the primary evaluation of the learning: to apply requires the children to have remembered and understood. Therefore, to assess the *apply* stage also implicitly assesses the *remember* and *understand* stages. Mayer (2002) discusses the importance of learning going beyond the *remember* stage and to be able to transfer any new knowledge to new problems.

Of course, when measuring the application of skills that the children have acquired, the actual measurement is not necessarily of *learning*, but of *performance* in the task (Mikulas, 1977). Task performance may rely not just on learning but also on several other factors such as motivation, fatigue, and so on. Due to complexities in separating learning from other factors in task performance (Mikulas, 1977), for the purposes of the work here, it will be assumed that changes in task performance reflect learning to at least some degree if experimental factors remain constant. Throughout, the aim is to influence learning, rather than task performance as task performance can be improved through mere repetition, whereas learning involves some aspect of knowledge transfer to solve previously unseen problems (which has the potential to be more useful).

An important distinction lies in the difference between 'affective learning', 'cognitive learning' and 'perceived learning', which are commonly presented in the immediacy literature (Witt et al., 2004). *Affective* learning considers constructs such as attitudes, values and motivation (Krathwohl et al., 1964). *Cognitive* learning consists of the elements discussed in Krathwohl (2002) and shown in Figure 2.2; these are typically topic specific knowledge and skills. *Perceived* learning is a measure of how much students believe they have learnt, or how confident they are in what they have learnt, such as in Gorham (1988).

With perceived learning, students are asked how much they believe they have learnt, often on a Likert scale, and this is used as the learning measure. This approach is commonly justified through precedence, for example in Butland & Beebe (1992), but does have some basis in evidence (Perry & Smart, 2007). This measure has the advantage of being applicable across multiple domains and learning objectives. However, there are challenges in using self-reporting measures of this nature with children (Borgers et al., 2000), and some journals question the correlation between perceived and actual learning, refusing

data based on self-assessments of learning (DiPiro, 2010). Due to the self-reporting nature of affective learning, the same challenges as perceived learning are faced when surveying children (Borgers et al., 2000). However, cognitive learning often has clear and quantitative means of measurement, lending itself well to a measurable outcome of interactions with robots. For this reason, the studies in this thesis focus on measuring cognitive learning.

2.7 Robots as Educators

This section will build upon the literature exploring embodiment and social presence effects of robots on learning and task performance introduced in Section 2.3 by considering the literature studying the effects of robot behaviour on learning and social responses in HRI in educational contexts. It should be noted that the focus is on robots used to deliver some aspect of teaching content (robots as educators), rather than robots used as a teaching platform (educational robotics). Section 2.4 showed that if the social behaviour of an agent can be improved then the social presence will increase and interaction outcomes should improve further, but it is unclear how social behaviour should be implemented to achieve such aims. This has resulted in researchers exploring various aspects of robot social behaviour and attempting to measure the outcomes of interactions in educational contexts, but a complex picture is emerging. This section will introduce literature from HRI with a view to establishing the state of the field in using robots to educate children. Particular attention will be paid to work involving children, measuring cognitive learning gains, as this closely relates to the scenario in this thesis.

2.7.1 Roles of Robots in Education

Robots can adopt a variety of roles when interacting with children in learning environments. These roles have been the subject of various research endeavours, which will be summarised here. Robots have been used as peers for children. This approach has the advantage that the robot can be programmed with less knowledge than the child, so can convincingly play this role in the way that their human class teachers cannot (and nor can their peers if they are lower ability within their cohort; Kennedy et al., 2016a). A robot was employed as a less-able peer to help improve children's handwriting, with promising preliminary results (Hood et al., 2015). In a similar methodology, Tanaka & Matsuzoe (2012) use a robot which requires care (a care-receiving robot) from children.

The aim is for the children to learn by teaching, and their results show that children's vocabulary acquisition is indeed higher in sessions with the robot than without the robot. The children also retain this vocabulary a number of days later. This work was conducted with the Aldebaran NAO, and more recent efforts have applied the same principle to the Aldebaran Pepper, with results forthcoming (Tanaka et al., 2015). It should be noted that this work does not compare a care-receiving robot to a non-care-receiving robot, so no conclusion can currently be drawn as to whether this technique holds advantages over other methods of tutoring. Other work with forthcoming results also aim to employ robots in a peer role Kory & Breazeal (2014).

Other researchers have employed robots in roles closer to teaching assistants. Kanda et al. (2004) found learning of a language was related to time interacting with a robot, however, this was only true for the second week of a 2 week study. Novelty effects may account for the lack of significance between interaction and learning in the first week, although the study was not designed to attempt to teach in an optimal manner, but to investigate the possibility of maintaining relationships with robots on a daily basis. A more controlled study by Alemi et al. (2014) found that using a a robot to supplement teaching over a 5 week period led to significant learning increases when compared to the same material being covered with a human teacher without a robot. This is strong evidence for the positive impact that robots can have in education.

Robots have also been used in a tutoring role. As previously discussed in Section 2.4, a Keepon robot was used as a tutor in a logical puzzle solving task with adults to explore embodiment effects (Leyzberg et al., 2012). This work was extended further to study the teaching strategy that the tutor adopted by personalising when hints were delivered. It was found that when lessons were personalised based on an assessment of human skill, significant improvement was found in puzzle solving times when compared to a no lesson or randomized lesson control group (Leyzberg et al., 2014). More in-depth exploration of such personalisation behaviours for robot tutors is also planned by other researchers (Charisi et al., 2015).

Based on these different approaches, some researchers have attempted to establish the most effective role for robots to assume in an interaction. Blancas et al. (2015) found no significant differences in performance of humans whether taught history by a teacher or a peer robot. Their conditions were operationalised through changes in verbal content and

nonverbal behaviour such as gestures. However, their study was conducted with adults, and the teaching element only lasted for 2.5 minutes with unidirectional interaction (the robot read a script to the human). Whether these findings would persist when transferred to interactions with children, over longer periods of time, and with a greater degree of interaction, remains to be seen. Zaga et al. (2015) also explore this question, comparing a robot as a peer with a robot as a tutor, when supervising pairs of children completing a logical puzzle. It was found that children performed better with the peer robot and looked at the peer robot more, but prior ability was not controlled for and subject numbers were relatively low (5 pairs per condition). Diyas et al. (2016) does not observe conclusive differences in response to different robot roles ('peer' vs 'teacher') when manipulating verbal behaviour. This therefore seems to remain an open question, but the best role for the robot may well depend on its morphology as this may set particular expectations (a large, complex robot may not make a convincing peer in the same way a NAO might, for example).

2.7.2 Robot Behaviour in Educational Interactions

Aspects of a robot's non-verbal behaviour have been investigated in one-on-one tutoring scenarios with mixed results. Herberg et al. (2015) found that the HHI literature would predict an increase in learning performance with increased gaze of a robot towards a pupil, but the opposite was observed. An Aldebaran NAO would either look towards or away from a child while they completed a worksheet based on material they had learnt from the robot. Additional gaze towards the child was predicted to increase performance. This was not found to be the case, but a potential confound of robot movement *vs* non-movement may have played a role in the behaviour of the children. However, Saerbeck et al. (2010) varied socially supportive behaviours of a robot in a novel second language learning scenario. These behaviours included gestures, verbal utterances and emotional expressions. Children learnt significantly more when the robot displayed these socially supportive behaviours.

The impact on child learning of verbal aspects of robot behaviour have also been investigated. Gordon et al. (2015) developed robot behaviours to promote curiosity in children with the ultimate aim of increased learning. Whilst the children were reciprocal in their curiosity, their learning did not increase as the HHI literature would predict. Kanda et al.

(2012) compared a 'social' robot to a 'non-social' robot (operationalised through verbal utterances to children when they are completing a task). Children showed a preference for the social robot, but no learning differences were found.

Methodological complications meant that in a study by Short et al. (2014) it is unclear whether children learnt about nutrition from a robot. The study attempted a longer-term interaction protocol, with 6 sessions. Response time for choosing food was used a metric, but the questions also became progressively more difficult, which acted as a confound to the chosen indication for learning. As part of the same project, but with a different educational focus (mathematics tutoring), Ramachandran et al. (2016) used a robot to shape requests that children made for help. They found that a robot which shaped sub-optimal help strategies from children led to the children modifying their behaviour and learning more when compared to a control condition, where the robot did not shape sub-optimal strategies. It has also been suggested that children are more likely to ask for help from robots in educational interactions as there is less of a social stigma involved with doing so when compared to asking other humans for help (Howley et al., 2014).

Ultimately, it is a difficult task to present a coherent overview of the current state of research in the child-robot interaction (cHRI) educational domain with many results appearing to contradict one another, or not being comparable due to the difference in learning task or behavioural context. More researchers are now using the same robotic platforms and peripheral hardware than before (quite commonly the Aldebaran NAO or DragonBot with a large touchscreen, e.g., Baxter et al., 2012), but there remain few other similarities between studies. Behaviour of various elements of the system are reported alongside learning outcomes, but it is difficult to translate from these descriptions to something which can be compared between studies. As such, it becomes almost impossible to determine if differing results between studies (and discrepancies with HHI predictions) are due to differences in robot behaviour, the study population, other contextual factors, or indeed a combination of all three. It is apparent that a characterisation of the robot social behaviour would help to clarify the differences between studies and provide a means by which certain factors could be accounted for in analysis; this will be explored in the following section, and demonstrated throughout the experiment-focussed chapters of this thesis (Chapters 5-8).

Prior work reveals two main avenues of exploration in HRI for education: *(1)* study of

different teaching strategies (e.g., Leyzberg et al., 2014; Ramachandran et al., 2016), and *(2)* study of the effects of social behaviour/presence (e.g., Kanda et al., 2004; Saerbeck et al., 2010; Zaga et al., 2015). One of the main distinctions between the two approaches is in the nature of the specificity of findings. Manipulations to the teaching strategies are specific to the particular educational topic of focus, whereas the behavioural findings could theoretically be applied to other topics. It is not clear from HRI findings that this transfer still produces the same outcomes, but HHI literature suggests that it would (as discussed in Section 2.4). However, there are few clear guidelines on how to operationalise robot behaviour to optimise child learning; this is an area that this thesis will contribute to.

2.8 Characterising Social Behaviour

To allow researchers to make clearer comparisons between studies and across contexts, a metric (or set thereof) to characterise the social behaviour of a robot is desirable. Various metrics have been used before in HRI and some of these will be introduced here. For the purposes of the research conducted as part of this thesis it is desirable to have a metric which is appropriate for use with children.

Retrospective video coding has been used in several HRI studies as a means of measuring differences in human behavioural responses to robots, for example (Kennedy et al., 2015b; Moshkina et al., 2014; Tanaka & Matsuzoe, 2012). However, for cross-context and cross-study comparisons it is important for the metric to be holistic in its characterisation of the social behaviour so that all social cues are considered in the context of one another. For example, when studying gaze, it is important to know whether gestures are occurring at the same time, and so on, as this could influence human perception and experimental findings (Zaki, 2013). This should be accounted for within the metric for characterisation, and so manual video coding is not well suited for this kind of approach due to the time intensive requirements of such analysis.

The Godspeed questionnaire series developed by Bartneck et al. (2009b) has been used in many HRI studies to measure users' perception of robots (Bartneck et al., 2009a; Ham et al., 2011). The animacy and anthropomorphism elements of the scale in particular consider the social behaviour and perception of the robot. However, it is not particularly suited to use with children due to the language level (i.e., use of words such as 'stagnant', 'organic', 'apathetic'). It may also be that the questionnaire would measure aspects of the robot not

directly related to social behaviour as it is asking about more general perceptions. Whilst this could be of use in many studies, for the aim of characterising social behaviour in the case here, these aspects prevent suitable application.

Immediacy was introduced by Mehrabian (1968) and is defined as the 'psychological availability' of an interaction partner. Immediacy can be broken into nonverbal and verbal aspects. Several versions of surveys have been developed and validated for measuring the nonverbal immediacy of adults (Richmond et al., 2003). Surveys have also been developed for verbal immediacy (Gorham, 1988), but their ability to measure precisely the concept of verbal immediacy remains the subject of debate (Robinson & Richmond, 1995). Both verbal and nonverbal measures consider observed overt behaviour more than, but not excluding, perceptions. Immediacy has recently been used in HRI as a means of motivating robot behaviour manipulations (Szafir & Mutlu, 2012) and characterising social behaviour (Kennedy et al., 2015e). Whilst the immediacy surveys are not designed for use with children, the consideration of only overt behaviour allows a researcher to simplify the language to an appropriate level, which is not necessary possible with more abstract concepts (such as 'organic' seen in the Godspeed series).

BEHAVE (Joosse et al., 2011) and BEHAVE-II (Joosse et al., 2013) are measures for assessing users' attitudinal and behavioural responses to a social robot. These measures in part utilise concepts of nonverbal immediacy, but combine them with more extensive questioning of personal space/proxemics, and observations of emotion. Whilst these measures are designed for measuring user behaviour rather than robot behaviour, it could be reasonable to manipulate the questions to make them robot-centric. However, the perception and interpretation of a robot's emotion (if existent) by children would not necessarily be trivial and the proxemic behaviour questions have less relevance when the interactants have fairly static locations as in the interactions considered in this work.

2.9 Nonverbal and Verbal Immediacy

As stated previously, immediacy was introduced by Mehrabian (1968) and is defined as the 'psychological availability' of an interaction partner. More generally, this is "the extent to which communication behaviours enhance closeness to and nonverbal interaction with another" (Mehrabian, 1968). A number of specific social cues are used (touching, distance, forward lean, eye contact, and body orientation) to form part of the measure for

this concept, which were later utilised by researchers that sought to create and validate measuring instruments for immediacy, e.g., Richmond et al. (2003).

Immediacy was selected as the most appropriate metric for characterising robot social behaviour throughout the work presented here. There is a consensus on the instruments used to measure nonverbal immediacy (whereas this is less clear for verbal immediacy) and it is also transparent in terms of how participants are judging the robot. The Godspeed questionnaire is a useful tool for gathering perceptions, but nonverbal immediacy focusses on accounting for overt social behaviour and so it is ideal given the aim of trying to characterise social behaviour (often with children). Immediacy has undergone extensive evaluation and validation in human-human studies. This brings with it the advantages that it can be considered to be reliable, and that it can tie findings to other literature exploring nonverbal immediacy and learning. The immediacy metric as adapted and applied in this thesis will be discussed in greater detail in Chapter 3, while a more general introduction and background to the concept will be provided here.

2.9.1 Application in Human-Human Interaction

A reasonable volume of data already exists for studies considering immediacy, with over 80 studies (and N nearly 25,000) from its inception to 2001 (Witt et al., 2004), and more since. This provides a context for NVI findings in HRI scenarios and a firm grounding in the human-human literature from which roboticists can draw. It has found extensive application in educational research, most often in university lecture scenarios (Witt et al., 2004). This application to education is particularly relevant given the domain of this thesis, although the target population here is young children rather than university students.

When used in studies between humans, the correlation with measured cognitive learning gains is only moderate, however relatively few studies have used experimental measures; most have used perceived learning, which has a particularly strong correlation with teacher immediacy (Witt et al., 2004). It has been experimentally found that perceived learning and actual recall are moderately correlated in such contexts (Chesebro & McCroskey, 2000), so whilst perceived learning is not as strong as measuring actual learning, it can at least be used as an indication of the nature of relationships. A positive correlation between nonverbal immediacy and perceived cognitive learning has been validated across several cultures, including the United States, Puerto Rico, Finland and Australia (McCroskey

et al., 1996). From this, McCroskey et al. (1996) postulate that expectation of immediacy plays a key role in how cues are interpreted, presenting opportunities for high immediacy teaching to have a strong positive impact in generally low immediacy cultures, but a negative impact for low immediacy teaching in high immediacy cultures. A similar suggestion relating to the use of robot social cues in teaching contexts has also been raised in HRI (Kennedy et al., 2015a).

Immediacy can be broken down into verbal and nonverbal aspects. Nonverbal immediacy measures consider specific social cues: gaze, gesture, body posture, proximity, etc. However, verbal immediacy is less clearly defined as a measure (Robinson & Richmond, 1995). The verbal immediacy metric proposed by Gorham (1988) includes various personalisation aspects such as using an interacting partners' name, along with aspects of familiarity, such as revealing personal information. Whilst the nonverbal immediacy scale is a general characterisation of an individuals' social behaviour, this proposed verbal immediacy scale is specifically designed for educational environments, with items such as "the instructor gets into conversations with individual students before or after class" (Gorham, 1988). The implications of this in the context of this thesis are discussed in Section 3.6.

Both verbal and nonverbal immediacy behaviours have been shown to lead to an increase in motivation and, in turn, student learning (Christensen & Menzel, 1998; Christophel, 1990). In some cases, such as in a task to recall contents of a lecture (Chesebro, 2003), cognitive learning gains are not found, but affect for the instructor and material increases when the instructor is judged to have higher nonverbal immediacy. However, there are other examples demonstrating a link between more nonverbal immediacy and increased recall (Goodboy et al., 2009; Witt & Wheeless, 2001). An extensive review of the potential benefits of immediacy (both verbal and nonverbal) can be seen in Chesebro & McCroskey (1998). There are suggestions that the relationship may not be wholly linear in nature (Comstock et al., 1995; Witt & Schrodt, 2006), but instead an inverted U shape. This would suggest that if social behaviour is too high in nonverbal immediacy then it could start to become detrimental to learning. As such, it remains unclear how immediacy should be implemented for social robots, which this thesis aims, in part, to make a contribution towards. This lack of clarity in findings between humans also creates a limitation in that it makes it more challenging to identify whether findings are a property of using robots, or a property of nonverbal immediacy itself.

2.9.2 Application in Human-Robot Interaction

Nonverbal immediacy has been studied only briefly in HRI contexts before, and besides the work later in this thesis, verbal immediacy does not seem to have been applied in an HRI context. Szafir & Mutlu (2012) use nonverbal immediacy as a means of motivating and evaluating robot behaviour during a recall task with adults. In line with literature studying nonverbal immediacy with humans, they find that as immediacy increases, so does recall. The adults were also able to notice when the nonverbal immediacy of the robot had increased, confirming that adults are sensitive to such cues in robots. Nonverbal immediacy concepts have also been used by the same research group to motivate behavioural manipulations for persuasive robots (Chidambaram et al., 2012). However, it should be noted that it doesn't appear that a complete nonverbal immediacy questionnaire was used in either of the studies. Finally, nonverbal immediacy has recently been proposed for use in HRI studies to motivate exploring the perception of a robot when posture and nodding behaviour is varied (Jeong et al., 2015).

2.9.3 Measuring Immediacy

Several versions of surveys have been developed and validated for measuring the nonverbal immediacy of adults (Richmond et al., 2003). Surveys have also been developed for verbal immediacy (Gorham, 1988), but their ability to measure precisely the concept of verbal immediacy remains the subject of debate. Robinson & Richmond (1995) highlight the lack of negatively worded items and suggest that this inhibits the face validity of the survey. The construct validity is further called into question given the low correlation between verbal immediacy, nonverbal immediacy and affect towards an instructor using data from an experiment (Robinson & Richmond, 1995).

The nonverbal immediacy measure proposed by Richmond et al. (2003) has generally excellent face validity due to the fact that it contains items that explicitly explore the nonverbal behaviours included in the definition of the concept provided by Mehrabian (1968). However, there are also two items that are less clearly directly attributable to specific social behaviour cues, and could instead could be considered a more subjective rating, for example "Looks bland or neutral when talking to people". It is unclear how looking 'bland' should be interpreted. This problem is exacerbated in the version of the questionnaire developed as part of this thesis, where the language was modified for

children.

Wilson & Locker Jr (2007) have sought to validate the use of the immediacy measure proposed by Gorham (1988). This measure was proposed as a verbal immediacy measure, but also includes 5 items from the nonverbal immediacy measure (all positively worded, avoiding the more subjective items referred to in the previous paragraph). It was found that the immediacy ratings (both verbal and nonverbal) were correlated to perceptions of instructor effectiveness and that a factor analysis revealed both aspects are valid measures of immediacy (in both face and construct validity). The study was conducted with 44 instructors and 1572 participants, so this provides some reassurance that the surveys do indeed measure the concept of immediacy, although these findings are contrary to those of Robinson & Richmond (1995). Additionally, not all nonverbal immediacy items were included. This unclear validity is a potential limitation of the measure, and more so, the adaptation used here where items with questionable face validity become difficult to interpret when modifying the language for children (described in Section 3.6).

Perceptions of immediacy have been validated through physical manipulation of some of the social cues, specifically eye gaze and proximity, in order to ensure that the phenomenon works in practice and is not a product of affect or bias in survey responses (Kelley & Gorham, 1988). It was indeed found that the physical manipulations that were made which would lead to a higher immediacy score (standing closer and providing more eye gaze) did lead to increased short-term recall of information. While there is clearly a difference between recall and learning, recall of information is a promising first step to acquiring new understanding and skills. These results were hypothesised to exist in the other immediacy behaviours (such as gestures) as well.

Overall, the link between teacher immediacy and student learning is hypothesised to be a positive one, as reflected in the meta review by Witt et al. (2004) and many studies, e.g., Christensen & Menzel (1998); Comstock et al. (1995); McCroskey et al. (1996). Despite some debate about the validity of immediacy, it is particularly appealing for HRI as the measures can be used as lists of overt social behaviours that robot behaviour designers can manipulate. To the authors' knowledge, no other metric for social behaviour currently provides this relatively low-level guide for behaviour implementation, while simultaneously offering a means of characterising social behaviour and verifying behavioural manipulations.

2.10 Summary

This chapter has described the motivation for creating robot tutors, and provided a definition of learning for this thesis for clarity and to focus the scope of the work. Literature from both human-human studies and human-robot studies which led to the research questions formulated in Chapter 1 was discussed. Human-human studies reveal that social behaviour can positively influence learning, but human-robot studies present a more complex picture. Differences in methodologies and behaviours mean that it is unclear precisely how a robot should behave in educational interactions, and makes a measure for robot social behaviour desirable. A review of social behaviour metrics revealed that *immediacy* could be applied to HRI to characterise robot social behaviour, and possibly also to tie findings to the broader immediacy literature examining human learning in response to different social behaviours.

Chapter 3

Materials and Methods

This chapter describes the tools and measurement instruments used throughout the empirical studies described in later chapters. Specifically, the NAO robot and Sandtray touchscreen are introduced, with an explanation of how they operate in tandem to deliver the learning content in interactions. Standard immediacy metrics and the 'child-friendly' versions developed for use with robots in this research are also discussed. Study specific tests and materials will be described in the relevant later chapters.

3.1 The NAO Robot

The Aldebaran NAO robot is a 58cm tall humanoid robot which is used as the sole robotic platform throughout this research (Figure 3.1). The robot is available to researchers, developers and educational institutions. The robot (or NAO) has 25 degrees of freedom and a variety of sensors, including microphones, touch sensors and cameras. Additionally, two speakers are mounted in the head of the robot to play sound. The eyes of the robot contain an array of Light Emitting Diodes (LEDs) which can change colour.

Two different hardware versions of the NAO robot were used during this research: a v3 H25 body with a v3.3 and a v4 head. Both versions of the head incorporate a custom Linux distribution with the *NaoQi* operating system acting as an interface to the programmer. The versions of NaoQi used here (1.12.x and 1.22.x) include Nuance speech recognition and a Text-To-Speech engine provided by Acapela.

The NAO robot is particularly suited for interaction with children due to to its small size and approachable appearance (Shamsuddin et al., 2012). Humanoid robots also benefit

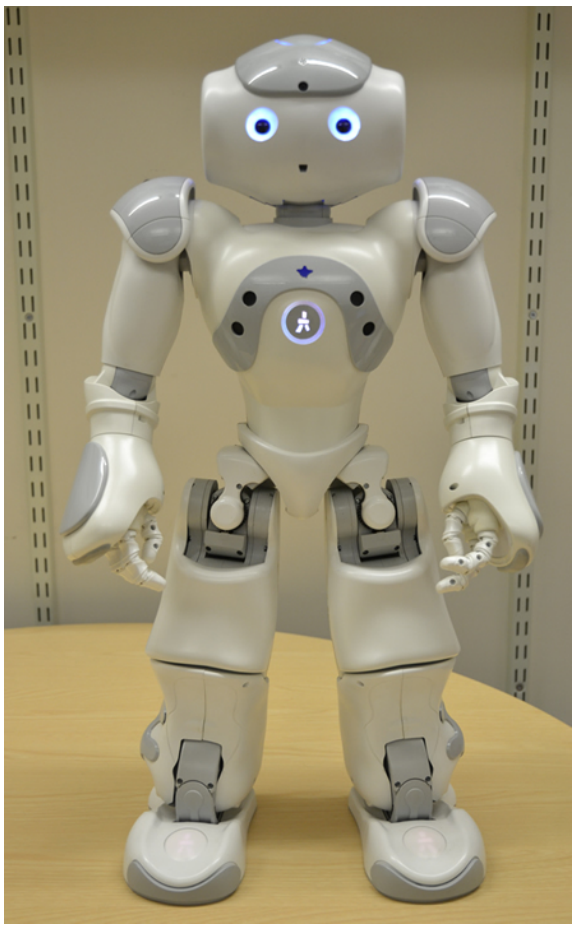

Figure 3.1: The Aldebaran NAO used in the majority of the evaluations throughout this thesis.

from having modalities that map to the human body, thus enabling implementation of social behaviour based on human models. Principles from human-human research can also be more readily transferred, providing a possible theoretical starting point for investigation or development of research questions.

The *Universal Real-Time Behavior* (Urbi) middleware (Baillie et al., 2008) was installed on the robot in addition to the standard packages. Urbi is specialised for robot control, providing many high-level orchestration functionalities, easing real-time multimodal behaviour control. A convenient scripting language (UrbiScript) is used for programming the robot. UrbiScript closely resembles C++, but has additional operators to determine whether lines, or blocks, of code should be executed in serial or parallel, synchronously or asynchronously. Urbi offers the ability to perform offboard processing, but all robot control and communication to external devices in this research was performed through UrbiScript running directly on the robot. Urbi versions 2.7.5 and 3.0 were used during this project; they are tied to NaoQi versions 1.12.x and 1.22.x respectively.

3.2 The Sandtray Touchscreen

The *Sandtray* touchscreen, developed by Baxter et al. (2012) to constrain HRI and provide a shared workspace for both the human and robot, is used in the majority of the studies for this research. Two different versions of the Sandtray have been used throughout this research (Figures 3.2 and 3.3). Both versions consist of a large touchscreen mounted horizontally. The first version had a 26 inch screen at a height of 30 centimetres above the floor, and the second version had a 27 inch all-in-one touchscreen computer mounted 26 centimetres above the floor. This is an ideal height for the NAO robot to convincingly appear to manipulate items on the screen, and is also easy for children to use when seated on the floor. When using the first Sandtray version, the robot would stand on a small platform to reduce the height difference to the desired 26 centimetres. The first version also required a separate computer to display on the touchscreen; to this end, a laptop was used and hidden in the housing below the screen.

Figure 3.2: Version 1 of the Sandtray touchscreen used in experimental evaluations. A laptop is stored within the wooden housing to run the software, with the touchscreen used as a display.

Figure 3.3: Version 2 of the Sandtray touchscreen used in experimental evaluations. This version is much improved, with a sturdier yet smaller construction, larger screen, lighter weight, and built-in computer.

The Sandtray runs a Windows operating system (versions 7, 8, and 8.1 at varying times),

with a custom software application. This application has two-way communication with an Urbi server running on the robot. Communication between the touchscreen software and the robot is performed through socket connections over a wireless network, where encoded strings are passed with payloads, such as commands to move images on screen. An early version of the software was written in C++ with DirectX and allowed two-category sorting games to played on screen. This application was later re-written with extended functionality (including the removal of limits on the number of categories that could be used) using the Qt C++ libraries (Qt SDK version 4.8.5). The software allows simple loading of different sets of images, so the same control architecture can be used across different learning contexts such as mathematics and language learning.

Using a touchscreen as a mediator in the interaction between the child and the robot brings about many benefits. In physical realms, there are objects that are too big for the robot to grip, and gripping can be a computationally heavy (and time-expensive) process; the touchscreen acts as a shared workspace where the child and the robot have equal ability at manipulation (Baxter et al., 2012). The touchscreen also acts as a focal point in the interaction, typically not only constraining the child's verbal behaviour (Kennedy et al., 2013), but their nonverbal behaviour as well. The use of the screen reduces the possible space where a child might be (as they will be within touching distance of the screen, on the opposite side to the robot). This significantly reduces complexity for tasks such as gaze detection and estimation (Lemaignan et al., 2016). For the child, the touchscreen interface is intuitive for children to use, and means that they can interact on screen without requiring proficiency in the use of traditional computer input methods like keyboards and mice (Park & Howard, 2013).

3.3 The Microsoft Kinect

In some experiments, the Sandtray hardware setup was extended with a Microsoft Kinect for Windows (Kinect v1). The Kinect was added to the hardware setup with a purpose built mount, that was a measured distance from the robot (Figure 3.4). As a result, the Kinect is not likely to move during interactions and as distances to the robot are known, the matrix transformations to translate and rotate the camera view to the frame of the robot can be executed reliably. The Kinect data stream is activated and managed by an application (developed by the author) that runs on the touchscreen (underneath the game

software) and communicates with the Urbi server running on the robot.

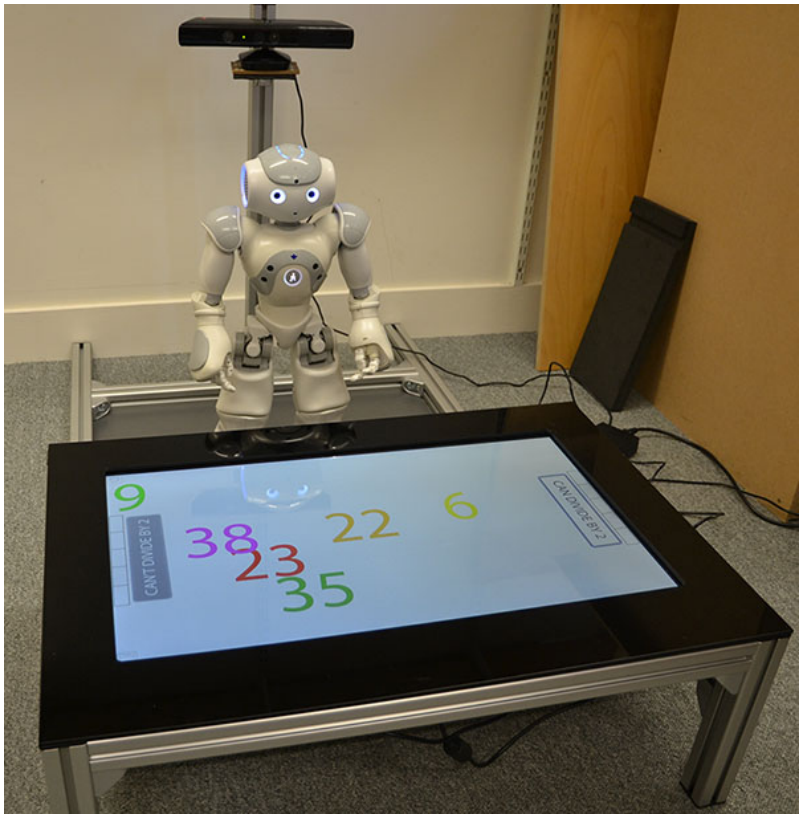

Figure 3.4: Version 2 of the Sandtray touchscreen used in experimental evaluations, with the Microsoft Kinect hardware extension. The Kinect is mounted on an additional lightweight metal frame, at a fixed distance from the robot. The robot stands with a raised rectangle on the base board. This board helps with robot stability (it allows the feet to slide more easily than the carpets often present in evaluation environments), and prevents the robot from turning away from the children when performing gestures (a common problem without fixing the feet in place).

The Kinect application was written in Microsoft Visual Studio 2010, using the Microsoft Kinect SDK v1.7, in C#.Net (Figure 3.5). The skeleton stream is used to select the nearest skeleton and track the head position of this skeleton. This position is rotated and translated to the robot frame of reference. Using the depth information and the head direction vector, it is calculated as to where this vector would intersect the plane of the robot. If this intersection is within a 12cm area around the robot's head, then the skeleton can be considered to be looking at the robot. The application provides events to the robot when particular actions occur (such as a new skeleton being tracked), and can be queried by the robot if particular information is required at a given time. The code running on the robot decides how to act upon this information, for instance, it may be appropriate to actively return the gaze. Simultaneously, the Kinect application logs all activity to the local Sandtray machine with a 1 second resolution, such that further post-hoc analysis can be performed when necessary. All major elements of the application are executed in separate

threads so that communication and logging can occur asynchronously, maximising the tracking frame rate.

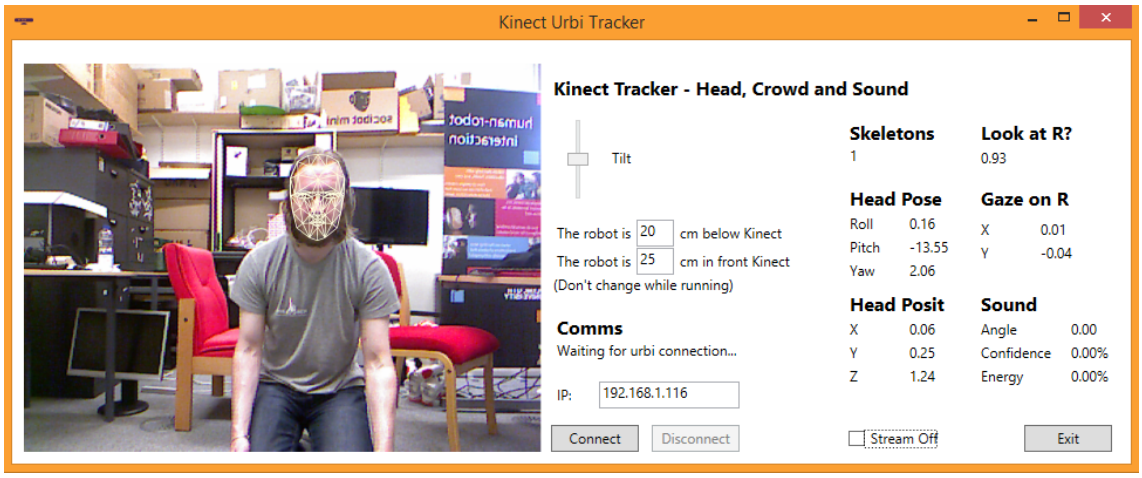

Figure 3.5: Microsoft Kinect application used to process data from the Kinect sensor and to communicate with the Urbi server running on the robot. This figure shows the Kinect with the visualisation of the camera stream on, but this is turned off when evaluations are running as the application cannot be seen, and turning off the visual rendering substantially reduces CPU load.

3.4 Robot Behaviour Generation

Despite evidence for the benefits of certain social cues (explored in detail in Section 2.4), it is still unclear precisely what can trigger learning in humans (VanLehn et al., 2003). Additionally, descriptions of social behaviours used when teaching are often not at the resolution required by social roboticists for development and implementation. As such, there is a need to explore what social behaviour a robot should employ when tutoring in order to elicit maximal learning gains from a human interactant.

Different approaches have been adopted within the HRI research community, and two different approaches have been applied in this research. The first approach taken is to use a human as a model for social behaviour, for example Sharma et al. (2013). It is assumed that a human is naturally social, and can provide a target behaviour to implement on a social robot. By observing a human conducting the task that the robot must complete, annotations can be made (with more accurate video coding if the interaction is filmed) and a behaviour can be derived. This approach is used in Chapters 5 and 6 of this thesis.

An alternative approach is to derive social behaviour from the HHI literature, as in Szafir & Mutlu (2012). While there are many task and context specific aspects of behaviour that

cannot be derived purely from the literature, broad guidelines can be created, as shown in Section 2.4. In this case, nonverbal and verbal immediacy guidelines are used to generate robot social behaviour in Chapters 4, 7 and 8.

One field of research that could also be considered as a source of information for robot behaviour is that of ITS and Artificial Intelligence in Education AIED (Freedman et al., 2000). This field has been active for much longer than HRI and has considered challenges in instructional design in detail (Vanlehn, 2006). However, these solutions often focus on the educational aspects of tutoring, with less regard for social behaviour. This becomes a greater consideration when a social robot is used due to the increased social presence of the interacting character. Even within this literature, there are still warnings that tasks need to be developed based on a model of the intended audience (Murray, 1999). A recent call within the AIED community has been to switch away from traditional computers and to adopt robots as an educational platform, but the challenges (and importance) of social behaviour are seemingly not yet recognised (Timms, 2016). As such, the research conducted as part of this thesis cannot draw as much as would be desired from the ITS field due to the greater focus on social behaviour here. Tasks have followed the principle described in the ITS literature (designing for the intended audience), but have not been re-used from the ITS literature – in part because appropriate tasks could not be established from ITS for the audience and topics under consideration here.

Section 2.5 indicated the importance of considering social cues in context. This means not just in the learning scenario, but also with respect to one another. Considering social cognition to be the simultaneous processing of a range of social cues as a single percept means that varying individual cues and trying to make conclusions about their use when combined with other cues and contexts is an inappropriate method (Zaki, 2013). It follows that a measure which considers the combination of social cues and their interactions with one another, such as immediacy (Mehrabian, 1968), is best suited for evaluating correlations in response to social behaviour.

3.5 Video Coding

In order to characterise both robot and child social behaviour, one technique often used in the HRI literature is to manually video code specific behavioural instances (Kahn Jr et al., 2003; Moshkina et al., 2014; Zaga et al., 2015). The aim is to be able to quantify particular

behaviours of interest by specifying when they occur, and for how long. For instance, a common social behaviour to code would be human gaze. A number of pertinent categories could be selected for this purpose, such as 'towards robot', 'towards touchscreen', and 'other'. Following the coding process, the number of instances and amount of time that each of these behaviours occurred can then be calculated. Additionally, if multiple behaviours are coded, then correlations and patterns between the behaviours can also be investigated. For example, it may be the case that the human looks at the robot each time the robot moves. This type of analysis can be used to reveal such effects.

The coding process is often performed manually for high accuracy and reliability. To provide further confidence in the coding objectivity, multiple coders are often used, with their coding cross-referenced to verify inter-coder agreement. This is usually done through a statistical measure such as Cohen's Kappa (Baxter et al., 2013). To ensure that the coding process is consistent and accurate, a coding manual is commonly used and provided to all coders. An example of an extensive coding manual can be seen in (Kahn Jr et al., 2003), with a further example of the manual provided to coders for studies in this work in Appendix A[1].

For the video coding performed as part of this thesis, a free, cross-platform video annotation software called ANVIL was used (Kipp, 2001), available online: `http://www.anvil-software.org/`. Many alternatives are available, but access to expertise within the ALIZ-E project made this a convenient choice. The software consists of a video player, menu bar, annotation panel, and coding time-line (Figure 3.6). Video coding specification files can be created within ANVIL, or in plain XML and used across several projects. These files dictate the 'tracks' which are to be coded; each track corresponds to a behaviour, such as 'child gaze'. Annotations can easily be created within the time-line, and the files can be exported in a variety of formats. To run statistical analysis on all of the files produced from video coding, the author created a piece of software to total, average and correlate behaviours, as well as automate inter-coder agreement calculations. This can be found on GitHub[2]. The software was created to add functionality to the built-in options of the software, and to allow integration with Microsoft Excel (used for further analysis).

All primary video coding performed in the experiments throughout this thesis was completed by the author. Video coding is an incredibly time consuming process, so second

[1] Please note that this manual is based on the work of Paul Baxter as part of the ALIZ-E project and was just extended by the author to include further details specific to the study under investigation

[2] `https://github.com/james-kennedy/xml-video-stats.git`

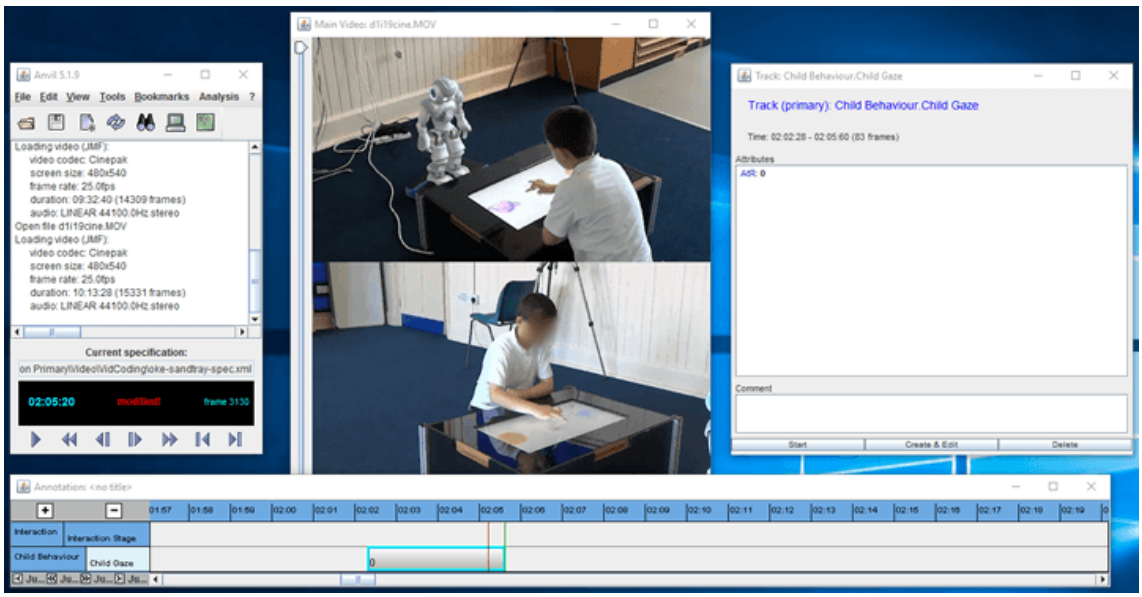

Figure 3.6: Screenshot of the ANVIL software package, showing *clockwise* from the *top-left*: (1) the menu, (2) the video player, (3) annotation panel, (4) coding/annotation time-line. Segments can be created on the time-line to indicate when behaviours occur.

coding is done on a portion of the videos to provide some indication of reliability. Various second coders were used throughout the experiments, the majority of which were not aware of the experimental hypotheses. They were provided with a coding manual and a short training period before coding the videos intended for reliability checking. This methodology has been used elsewhere in HRI (Moshkina et al., 2014). The videos for second coding were pseudo-randomly selected; it was first ensured that an equal split between experimental conditions would be used, and then selected randomly to provide this balance. The quantity of videos for each experiment and the inter-rater agreement is specified as part of the methodological description within respective chapters where this analysis is performed.

3.6 Immediacy Questionnaires

Chapter 2 introduced *immediacy* as a possible means of characterising robot social behaviour in HRI. Immediacy can be broken down into verbal and nonverbal aspects, each considering a different set of social cues. Questionnaires and scales for measuring both verbal and nonverbal immediacy exist from the HHI literature. Immediacy is used throughout the research in this thesis as a means of characterising social behaviour and to make comparisons between different experimental conditions. This section will discuss the immediacy scales offered by the HHI literature and how these were adapted for use with

children, and for HRI.

The Nonverbal Immediacy Scale developed by Richmond et al. (2003) is the one typically adopted in most HHI studies using immediacy as a measure. However, there are multiple version of this scale available: observer report, self report, and short form (available online[3]). The longer form versions contain 26 questions, whereas the short form version has 16 questions. These questions are all answered on a 5-point Likert scale (never, rarely, occasionally, often, very often) and contain both positively and negatively worded questions. The social cues considered include: touch, gaze, proximity, gesture, facial expression, and vocal prosody.

The verbal immediacy scale is less widely agreed upon. Gorham (1988) proposed and used a verbal immediacy scale, which considers a series of verbal immediacy behaviours specifically applied to teaching. This scale has been contested by Robinson & Richmond (1995) who suggest that many of the items lack face validity with immediacy as defined by Mehrabian (1968). Factor analysis of the whole immediacy scale (verbal and nonverbal components combined) reveals that some of the verbal questions may not relate strongly to the immediacy concept of psychological availability and should therefore be removed (Wilson & Locker Jr, 2007). However, the immediacy measure is of interest here as a characterisation of social behaviour and because of the correlations found with learning. These correlations are found with the original version of the scale (Gorham, 1988). As such, the original version was used as a basis for the scale developed here. The verbal immediacy items consist of a wide range of aspects of availability through spoken communication, such as the use of people's names, soliciting opinions, and discussing topics unrelated to any tasks being conducted.

Immediacy scales are typically used with adults in lecture scenarios (e.g., McCroskey et al., 1996) and the scales were developed in this context. As such, the linguistic abilities of children are not considered in the question wording. So that the scales can be used with children, the wording subsequently needs to be adapted. This wording adaptation was performed with the assistance of both a parent and a teacher of children of the age under consideration in this program of research. The full scales used can be seen in Appendices E, F and H.

Where deemed necessary, the language used was simplified, and any abstractions were

[3] http://jamescmccroskey.com/measures/

made more direct with the intention of making it apparent to the children that the questions were about their experience. So for example, "he/she looks away from people while talking to them" becomes "the robot looks away from you while talking to you". The layout of the scale was also modified; in the original versions statements are listed and the observer must place a rating from 1 to 5 before each statement. Due to concerns over interpreting children's handwriting and them remembering the number associated with each option, the options were placed after each question and simply required children to circle the answer that they wished to choose. Despite research which suggests an optimal number of options for children on a scale being four (Borgers et al., 2004), the five used in the original immediacy scales were kept such that the overall immediacy calculations and comparisons with prior literature could be maintained.

For the verbal immediacy questionnaire, some questions were removed due to their irrelevance to the interaction context used in the research in later chapters. For example, "invites students to meet or telephone after class..." (Gorham, 1988), would not be appropriate given the context of the single interaction studies considered here.

3.7 Crowdsourcing Immediacy Ratings

The work presented in subsequent chapters explores child learning and social behaviour in the context of various dyadic learning tasks, with different characters (be it physical embodiment, or behavioural differences). To create a common ground for comparison throughout these experiments, nonverbal immediacy can be used to provide a characterisation of the social behaviour of the character that the children interact with.

In the case of experiments in subsequent chapters which explicitly manipulate immediacy, immediacy ratings from the child participants are taken as part of the experimental protocol. In studies where immediacy itself is not used to motivate behavioural comparisons, immediacy ratings are not taken from the children, but it would nevertheless be desirable to have some measure of the nonverbal immediacy in relation to behavioural manipulations for these studies. These ratings would allow all of the work throughout this thesis to be compared on the same scale, as well as providing greater context for the results on a per-study basis. These ratings were acquired post-hoc from adults due to the convenience of acquiring adult participants. Of course, this relies on adults perceiving nonverbal immediacy in a similar way to children, as the child perception is the influencing factor

for learning. This is investigated in Chapter 4 (performed chronologically before the data collection described in this section to confirm validity of the approach before it was selected).

The nonverbal immediacy ratings themselves are not presented or discussed here, but instead are incorporated as part of the discussion for each study as it appears in this document, as well as forming part of the larger framework in Chapter 9. As such, the procedure is outlined here, and the data can be found in Appendix B. To provide sufficient subject numbers for all of the conditions, an online crowdsourcing service[4] was used. Adults were shown short video clips from interactions with children and completed a nonverbal immediacy questionnaire (Figure 3.7). Details for the crowdsourcing procedure are outlined below.

3.7.1 Quality Assurance

To ensure the quality of the data collected from the crowdsourcing platform, a number of steps were taken in the creation of the crowdworker tasks, and in the verification of the data once it had been received. Some of these steps were automated through options within the crowdsourcing platform, whilst other checks were performed manually.

Using the crowdsourcing platform options, the participants were restricted to the USA and to English speakers in an effort to prevent too much variance due to cultural differences. The IP addresses of workers are monitored by the crowdsourcing platform for this purpose (so a determined worker could potentially still use a virtual private network to gain access to the job). Additionally, workers could only take part if they had a reliable record within the crowdsourcing platform. Specifically, only workers who had completed over 10 jobs and had no major flags, and up to 1 minor flag (flags refer to attempts to 'cheat' jobs as judged by those creating the jobs) could participate. Due to these strict level requirements, and the consequence of workers losing access to other jobs, there is little incentive to attempt to cheat.

A test question was put in place in the job design whereby participants had to enter a 4 digit number into a text box (Figure 3.8). This number was shown at the end of the video for 8 seconds (the video controls were disabled and the number would disappear after the video had finished). A different number was used for each video. If the participants did

[4] http://www.crowdflower.com/

not enter this number correctly then their response was discarded. This was a means of identifying any workers attempting to answer the questions without having watched the whole video. This was to ensure the experimental protocol was being adhered to.

The crowdsourcing platform did not allow the prevention of users completing multiple conditions, so any duplicates were removed, i.e., only those seeing a video for the first time were kept as valid responses. This was a semi-automated step employed by the author after the data had been collected; a programming script was used to filter out all but the first job completed by any worker. A total of 496 responses were collected, but 266 were discarded as they did not answer the test question correctly, the user had completed another condition[5], or the response was clearly spam (for example, all answers were '1'; this is a manual check by the author). One further response was excluded from analysis as it was an outlier (Grubbs' test). This left 229 responses across 7 conditions (Appendix B).

Figure 3.7: Screenshot from the online crowdsourcing service used to gather adult non-verbal immediacy ratings. Radio boxes are used for answers to each question. The questionnaire is the same as that shown in Appendix E, but with the language switched to be an observer report, rather than self-report (i.e., "you" is changed to "the child").

[5]The vast majority of exclusions were due to users having completed another condition.

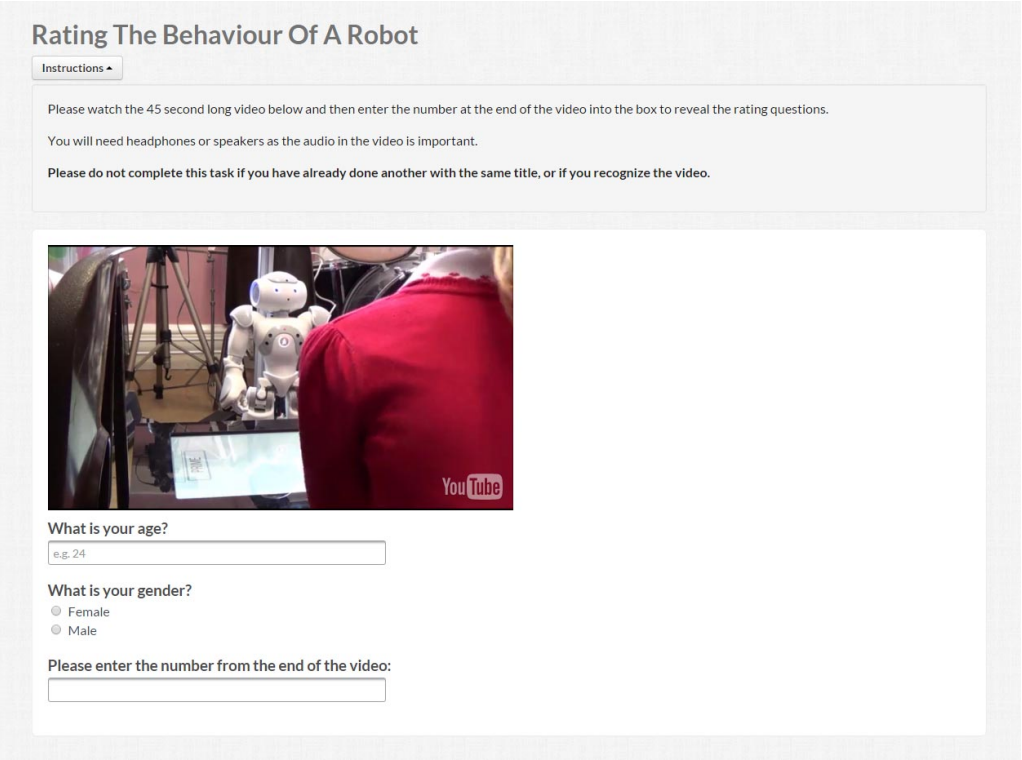

Figure 3.8: Screenshot from the online crowdsourcing service used to gather adult nonverbal immediacy ratings. The video is embedded in the page, without controls. Nonverbal immediacy questions are not revealed until the number from the end of the video is entered, at which point the video is removed from the page.

3.7.2 Video and Questionnaire Format

Videos shown to adults to acquire nonverbal immediacy scores were each 47 seconds long. The videos contained both the interaction video (42 seconds) and a verification code (5 seconds; details in the following paragraph). The length of video was selected to be 42 seconds as the literature suggests that at least around 6-30 seconds are required to form a judgement of social behaviour (Ambady & Rosenthal, 1993) and there was a natural pause at 42 seconds in the speech in all conditions so that it wouldn't cut part-way through a sentence. The interaction clips were all from the start of an interaction, so the same information was being provided by the tutor to the child in the clip.

Only the nonverbal immediacy scale was used as for verbal immediacy responses to be accurate, the participants would need to have seen a much greater length of the video (possibly the full interaction) due to how often certain verbal behaviours occur and how long they last for. This would have placed a much greater strain on data collection; collecting over 200 opinions on a video of up to 15 minutes long is a much greater challenge than the same procedure when the video is 47 seconds long (which is still not

straightforward). Additionally, in the majority of studies, verbal aspects of behaviour are constant between conditions, so a measure of the verbal behaviour is not as useful as a measure of the nonverbal behaviour. Where verbal behaviour is manipulated, child ratings for verbal immediacy were collected (Chapter 8).

3.8 Practical Procedures

Certain ethical considerations arise when conducting research with robots and children. These can be both general considerations that would be present when conducting research with children in any domain, and also domain specific. Many general considerations centre around the safety of the child and issues of consent. All of the experimental work conducted in this thesis was cleared by the university ethics board and adopts recommended best practices. This includes gaining consent from parents or guardians of children to conduct the research, but also informing the child of the nature of the research and providing them with the option to consent (or not) as well. To ensure child safety, they were always supervised when interacting with the robot, and no physical interaction between the child and the robot was present in any of the studies.

Domain specific ethical considerations consider those that relate either to children interacting with robots, or to the application of the research to child learning. Of course, there is an ethical responsibility to teach appropriate and correct material to the children. To this end, all of the work undertaken here was checked and approved by teachers. Some researchers have raised concerns over whether robots should be used to teach children at all (these broader aspects are explored further in Chapter 10). Sharkey (2016) presents several concerns related to the use of robots in classroom environments, including: privacy, deception, and reduced human contact. It is concluded that robots could offer new educational opportunities, but great care needs to be taken to prevent the robot from having a negative impact on child social relationships through loss of contact with other humans. The work conducted here primarily focusses on one-to-one tutoring for children, offering an addition to the current teaching environment, and certainly not a replacement for human teaching. Deception is kept to a minimum, partially through pursuit of autonomous, rather than 'Wizarded' robot behaviours, and also through presenting the robot to all of the children after each study. In these presentations, the robot is described such that the children can see that the robot is a piece of technology, rather than an independent

character, and they may ask any questions they have to the researcher. All data is kept anonymously and in accordance with university policy to protect the children's privacy.

When conducting experiments, effort was made to ensure that the children felt at ease not only interacting with the robot, but also with the experimenter(s). The class teacher would introduce the experimenters to the children on arrival at the school by their first name (as opposed to typical use of title and surname for adults in U.K. schools). The experimenters would inform the children about why they were visiting the school, and briefly set expectations for the children (for example, "You will get to play a game with a robot that will teach you maths. All of you will have the chance to see the robot, so do not worry if you are not picked on the first day we are here."). Further setting of child expectations was done by the experimenter immediately before the interaction; the experimenter would introduce the child to the robot and inform the child that the robot would tell them what they needed to do, whilst also making it known that they would be available if any problems arose, following advice from Ros et al. (2011).

3.9 Summary

This chapter has introduced the materials used in the research in the later studies, specifically, the practical tools: the NAO robot, the Sandtray touchscreen, and the Microsoft Kinect. The various immediacy questionnaires used have also been introduced, along with an explanation of the modifications made to make these questionnaires suitable for use with children and robots when compared to the originals used with adult humans. The processes involved in both video coding and adult crowdsourced nonverbal immediacy ratings were also described.

Chapter 4

Nonverbal Immediacy for use with Children and Robots

Key points:

- Nonverbal immediacy has been used extensively in adult lecture scenarios, where higher nonverbal immediacy is correlated with increased learning.
- Three different characters were used to read a story to both children and adults: an intended high nonverbal immediacy robot, low nonverbal immediacy robot, and a human.
- Both adults and children perceived the robot conditions as intended with respect to nonverbal immediacy. Adults and children also rate the behaviour of all three characters in a similar manner.
- Children recall more of the story from the robot with high nonverbal immediacy.
- These findings confirm hypotheses generated from the nonverbal immediacy literature, and provide a link between child-robot interaction and this literature.

Part of the work presented in this chapter has been published in Kennedy et al. (2017). The final publication is available from Springer via:
http://dx.doi.org/10.1007/s12369-016-0378-3

Nonverbal immediacy has been used extensively in adult human studies, often in lecture scenarios (Christophel, 1990; Gorham, 1988; Thweatt & McCroskey, 1998). It has also seen limited application in HRI evaluations (Chapter 2), and where this has been done, the immediacy scores have not been explicitly stated. As such, it is desirable to validate that behavioural manipulations of nonverbal immediacy when applied to a robot are perceived and reported as intended. Additional validation with children to check whether they interpret the behaviour in the same manner as adults would provide a solid basis for proceeding with using nonverbal immediacy in child-robot tutoring scenarios. This also provides a further link between the adult human immediacy literature and child-robot interactions, which can be useful in hypothesis generation and discussion of results. This chapter presents a study which aims to bridge the gap between the adult human literature and the child-robot interactions under consideration in this thesis. In particular, the study seeks to explore whether the nonverbal immediacy measure can be understood by children, whether children perceive manipulations in nonverbal immediacy, and whether the positive correlation between nonverbal immediacy and recall observed with adults also applies to children. If successful, this could then be used as a means of characterising robot social behaviour, and also assist in designing the robot behaviour (given the explicit list of cues that form the NVI metric).

4.1 Hypotheses

Based on the literature explored in Chapter 2, three hypotheses for the study were devised, in addition to a manipulation check. Chapter 2 explored the nonverbal immediacy literature, finding that several researchers have found a link between increased perceptions of nonverbal immediacy and learning or recall (Comstock et al., 1995; McCroskey et al., 1996; Witt et al., 2004). This has also been found for some robot behaviours when evaluated with adults (Szafir & Mutlu, 2012). These findings lead to a consideration of whether this effect will also be observed with children. To provide a greater link between the adult literature and the child-focussed group under consideration in this thesis, both the effect on perceptions of nonverbal immediacy and recall will be considered, leading to hypotheses H1 and H2. The effect at the group level will be explored through these hypotheses, with a further hypothesis devised to consider the impact of nonverbal immediacy perception on recall at the individual level in H3. The outcome here will demonstrate whether changes in

nonverbal immediacy behaviours are perceived by children, and whether the link between NVI and learning from the nonverbal immediacy literature applies when interacting with robots.

H1: Recall of a story will be greater when read by a character with higher nonverbal immediacy.

H2: Children and adults will perceive nonverbal immediacy in the same manner for both robots and humans (i.e., children and adults ranking of immediacy will agree).

H3: As nonverbal immediacy of the character reading a story is perceived to increase by an individual, their recall of the story will also increase.

As the robot behaviour is intentionally manipulated along the nonverbal immediacy measure (further details in Section 4.2.4), a manipulation check was conducted to verify that robot behaviour designed to be more or less immediate is perceived as such, as measured through the nonverbal immediacy scale.

4.2 Experimental Setup

A 2 (adults vs. children) x 3 (high NVI robot, low NVI robot, human) between-subject study was devised to explore how nonverbal immediacy would impact recall and to evaluate whether children and adults interpret the behaviour of a robot and a human in the same way. In order to achieve this, a scenario which could be understood by both groups was required. As such, the study design started from the perspective of the children (who are presumed to have a shorter attention span and more limited knowledge in some areas such as vocabulary) and was then applied to adults. Recall of a presented short story was decided to be an appropriate task for this purpose as this matched the lecture-style scenarios of immediacy studies (Gorham, 1988; Thweatt & McCroskey, 1998).

4.2.1 Participants

A total of 117 participants took part in the study, but one child had to be excluded due to an incomplete questionnaire and two adults were excluded due to inconsistent online video timestamps; this will be expanded on later in this section. 83 children (age M=7.8 years, SD=0.7; 47 F, 36 M) and 31 adults (age M=23.5 years, SD=3.9; 7 F, 24 M) remained

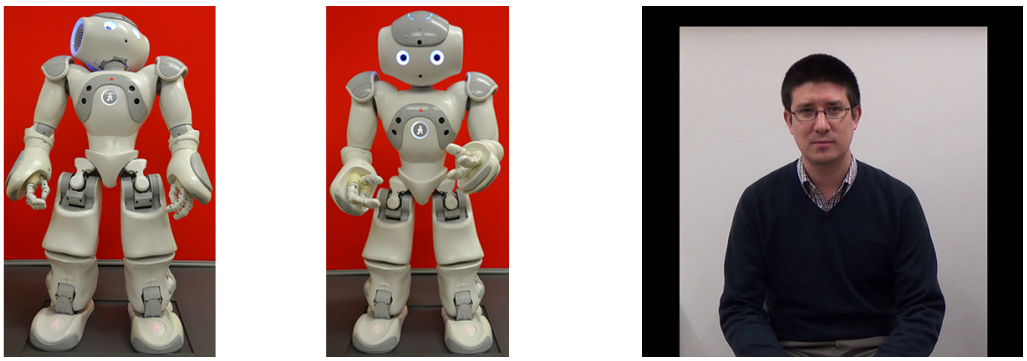

Figure 4.1: Still images from the conditions used in the evaluation; *left* to *right*: (1) low nonverbal immediacy robot, (2) high nonverbal immediacy robot, (3) human. Red backgrounds for the robot were not used in practice and are just used to ease visibility here; the video was shown in widescreen format, with a black background covering the unused space, as in the figure.

for data analysis. All participants consented to participation in the study and all children had parental permission to take part. The children were recruited from one school year group of a primary school in the U.K.; the children were split across conditions based on their usual school classes, which ensures an appropriate balance for gender and academic ability. Adults in the robot conditions were recruited through regular lectures, and through online advertising for the human condition.

4.2.2 Short Story

A short story was created for the purpose of the recall test. The story was largely based on one freely available from a website containing many short stories for children[1]. This was done to make sure that the language and content was appropriate for children. Some elements were added or modified in order to create opportunities for recall questions, and some of the phrasing was modified so that the robot text-to-speech sounded more accurate. The final version of the story created can be seen in Appendix D and lasts for just under four minutes when read in the experimental conditions. None of the participants reported to have heard or read the story before.

4.2.3 Measures

Two measures were used: a nonverbal immediacy observer report questionnaire and a recall test. The nonverbal immediacy questionnaire is described in Chapter 3 and can be seen in Appendix E. The recall test was devised based on information provided in the

[1]http://freestoriesforkids.com/children/stories-and-tales/robot-virus

short story and consisted of 10 multiple choice questions, with a final free text answer about the moral of the story. The full list of questions and answer options can be seen in Appendix D. The questions were designed to vary in difficulty based on how many times the piece of information had been stated, how central it was to the plot, and how many answer options were similar to the correct one. An additional question was added to the adult human condition regarding the colour of the background in the video; this was part of a series of checks to ensure that the video had actually been watched.

4.2.4 Conditions

In order to address the hypotheses for the study, three conditions were devised which were shown to both children and adults. Two conditions use a robot that has the social behaviour manipulated along the nonverbal immediacy scale, while the third uses a human to provide a control condition, given the novelty of applying nonverbal immediacy to robot social behaviour.

1. **High nonverbal immediacy robot** (Fig. 4.1 *centre*) - the robot behaviour was maximised for immediacy where possible; full details of the robot behaviour can be seen in the following paragraph. Child n=27; adult n=9.

2. **Low nonverbal immediacy robot** (Fig. 4.1 *left*) - the robot behaviour was minimised for immediacy where possible; full details of the robot behaviour can be seen in the following paragraph. Child n=28; adult n=9.

3. **Human** (Fig 4.1 *right*) - a human was recorded on video reading the story to provide a control condition. Video was used to ensure identical behaviour between child and adult groups and to time the story to be at the same pace as the robot conditions in order to have equivalent exposure time and reading speeds (which can impact recall; Hulme & Tordoff, 1989; Simonds et al., 2006). The human was not given explicit instructions in terms of nonverbal behaviour, as their immediacy level is not under consideration, but whether the children and adults perceive their immediacy level in the same way is. Therefore, the behaviour itself is not of concern, provided that it is identical between conditions. Child n=28; adult n=13.

Behavioural Dimension	High Nonverbal Immediacy	Low Nonverbal Immediacy
Gesture	Frequent gestures, occurring approximately every 12 seconds during the story. Slight randomness added to joints to provide small constant movement.	No gestures, no joint random movement.
Gaze	Head gaze directed forwards randomly at approximately the same height as the robot towards the centre of the movement range (towards observers).	Head gaze directed randomly up and towards the corners of movement range (over/away from observers).
Vocal prosody	No modifications to standard text-to-speech (TTS) engine, allowing shaping of sentences and responsiveness to punctuation.	All strings passed to TTS have punctuation stripped and are forced to be spoken with no context of the sentence (resulting in words sounding identical every time they are said). Additionally, vocal shaping was reduced via a TTS parameter.
Body orientation	Leans towards observers by approximately 15 degrees.	Leans away from observers by approximately 15 degrees.

Table 4.1: Operationalisation of behavioural manipulations between robot immediacy conditions

4.2.5 Robot Behaviour

The high and low nonverbal immediacy robot conditions were developed by maximising and minimising dimensions of the nonverbal immediacy scale, as described in Chapter 3. The conditions sought to maximise the differences between the behavioural dimensions and therefore also the dimensions measured by the nonverbal immediacy scale. Some dimensions were not varied due to limitations in the experimental set-up. Facial expressions were not varied as the robot being used for the study, an Aldebaran NAO, is not capable of producing facial expressions such as frowning or smiling. Proximity was not varied due to the group setting in which the study was being conducted. When the robot is telling the story to a classroom of children it is not feasible, or safe, to incorporate touch or to approach the children. The operationalisation of behavioural manipulations that were carried out can be seen in Table 4.1.

4.2.6 Procedure

For the robot conditions, the robot was placed at the front of the classroom on a table to be roughly at the head height of observers (either children or adults). The experimenter would then explain that the robot would read a story and that afterwards they would be required to fill in a questionnaire about what they thought of the robot. The recall test was explicitly not mentioned to prevent participants from actively trying to memorise the story. The experimenter then pressed a button on the robot's head to start the story. Once the story was complete, the nonverbal immediacy questionnaires were provided to all participants. When the whole group had completed this questionnaire, the recall test was introduced and given to participants. For the children, this was followed by a short demonstration of the robot. The human video condition procedure was the same for the children. The video was resized to match the size of the robot as closely as possible, and the volume was adjusted to be approximately the same as well.

As the children did not know this person, the adults should not either so that the reported immediacy score is based purely on the behaviour seen in the video and not prior interaction. The subjects for the video condition were recruited online and completed a custom web form which prevented the video from being paused or played more than once, and recorded timestamps for the start of the video, the end of the video, and the completion of the questions. An additional question about the background of the video

was also added to the recall test to verify that the participants had actually watched the video (as opposed to the rest of the recall questions which can be answered through listening alone; described in Section 4.2.3). One participant was excluded from analysis as the timestamps for the start and end of the video indicated too little time for the full video to have been viewed and another participant was excluded as the time between watching the video and completing the questions was in the order of hours (all other participants completed all questions in under 10 minutes), indicating that the intended protocol had been violated.

4.3 Results

4.3.1 Nonverbal Immediacy Results

Cronbach's alpha values were calculated for the nonverbal immediacy questionnaire for adults and children, splitting the human condition and the robot conditions. All alpha values are based on the 16 item scale. The reliability rating for the adults with the robot is high ($\alpha = .79$), whereas in the human condition it is quite a bit lower ($\alpha = .45$). This difference may be an effect of embodiment, and will be explored further in the discussion Section 4.4.4. Reliability scores for children are relatively low in both cases (human $\alpha = .55$; robot $\alpha = .30$). The implications of this are also discussed in Section 4.4.4.

Nonverbal immediacy scores were calculated from the questionnaires and produce a number which can be between 16 and 80. Immediacy scores and confidence intervals can be seen for each condition in Table 4.2. Whilst these scores might initially appear to be relatively low given the possibility of scores as high as 80, the scores do fall in the range expected. Due to the exclusion of certain aspects of the immediacy inventory in the robot conditions in terms of moving towards and touching observers, as well as producing facial expressions, it is unlikely that the score would raise above 56. It is however possible to be perceived differently and score more highly (for example the robot could have been perceived to have produced a smile, even though the mouth cannot move).

For the robot conditions, two groups were used with independence of observations, and a continuous measure for nonverbal immediacy. Distributions did not significantly deviate from normality (Kolmogorov-Smirnov test; $p > .05$) and had homogeneity of variances (Levene's test; $p > .05$). For this reason, two-tailed independent samples t-tests are used

to analyse the results for both children and adults. A two-tailed *t*-test on the adult data reveals a significant difference between the nonverbal immediacy score for the high immediacy robot (M=50.2, 95% CI [47.0,53.5]) and the low immediacy robot (M=36.3, 95% CI [33.5,39.1]); $t(16)$=7.460, p<.001. The same test on the child data also reveals a significant difference between the nonverbal immediacy score for the high immediacy robot (M=50.8, 95% CI [48.6,53.0]) and the low immediacy robot (M=46.5, 95% CI [44.2,48.8]); $t(53)$=2.793, p=.007 (Figure 4.2). These results confirm that the manipulation was successful: the robot behaviour designed to be more or less immediate is perceived as such when measured using the nonverbal immediacy scale. This provides a useful check that the behaviour of the robot has been interpreted as intended by both children and adults.

Support can be seen for hypothesis H2, that children and adults will perceive nonverbal immediacy in the same manner for both robots and humans (Table 4.2). The results show that both children and adults score the high immediacy robot very similarly, with almost identical means. The relative ranking of immediacy between conditions is also the same, with the high immediacy robot being perceived as most immediate, then the human, followed by the low immediacy robot condition.

However, there are also some differences as the child scores are more tightly bunched together; this could reflect their different (yet consistent) interpretation of negatively formulated questions (Borgers et al., 2004), or more limited language understanding impeding the data quality (Borgers et al., 2000). A two-way ANOVA was conducted to examine the effect of age group (child/adult) and condition (high/low robot, human) on the immediacy rating. These groups were used with two independent variables, independence of observations between groups, and a continuous measure. Distributions did not significantly deviate from normality (Kolmogorov-Smirnov test; $p > .05$) and had homogeneity of variances (Levene's test; $p > .05$). A significant interaction effect was found between age group and condition: $F(2,108)$=5.29, p=.006. Whilst both age groups rated the human to have lower NVI than the high immediacy robot, and the low immediacy robot in turn lower than the human, this difference was more pronounced for the adult age group. Significant main effects were found for condition ($F(2,108)$=16.96, p<.001) and age ($F(1,108)$=26.51, p<.001). However, due to the interaction effect between age group and condition, exploration of simple main effects splitting the conditions is also required to correctly interpret the results. Significant simple main effects are found for condition within each level of age group (child/adult): adults – Wilks' Lambda=.796,

Condition	Adult *M*	95% CI	Child *M*	95% CI
High immediacy robot	50.2	[47.0,53.5]	50.8	[48.6,53.0]
Low immediacy robot	36.3	[33.5,39.1]	46.5	[44.2,48.8]
Human	41.5	[38.4,44.5]	49.7	[47.0,52.4]

Table 4.2: Mean nonverbal immediacy scores by condition

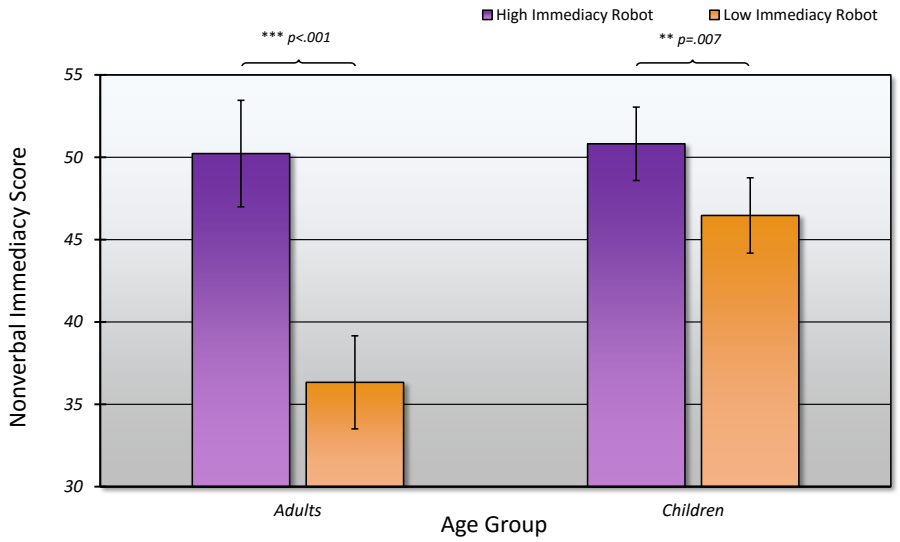

Figure 4.2: Robot nonverbal immediacy scores as rated by children and adults, relating to hypothesis H2. Both children and adults perceive the difference in nonverbal immediacy between the two robot conditions. ** indicates significance at the $p<.01$ level, and *** indiciates significance at the $p<.001$ level. *Error bars* show the 95% Confidence Interval

$F(4,214)=6.46$, $p<.001$; children – Wilks' Lambda=.798, $F(4,214)=6.38$, $p<.001$. Significant simple main effects are also found for age group within each condition: low immediacy robot – Wilks' Lambda=.664, $F(2,107)=27.11$, $p<.001$; high immediacy robot – Wilks' Lambda=.862, $F(2,107)=8.54$, $p<.001$; human –Wilks' Lambda=.811, $F(2,107)=12.49$, $p<.001$.

These findings suggest that some differences are present in the way that children perceive (or at least report) the immediacy of the characters when compared to adults. This is not surprising given the tighter bunching of child nonverbal immediacy scores. Nevertheless, there is a strong positive correlation between the child scores and the adult scores, $r(1)=0.91$, although this is not significant ($p=.272$) due to the low number of comparisons (3 conditions). Overall, due to the strong positive correlation and the same ranking of the conditions, it would seem that children perceive nonverbal immediacy in a similar manner as adults, but there are clearly some differences at least in terms of reporting. We would argue that there is a strong enough link to deem nonverbal immediacy an appropriate measure to use with children (and to tie the findings here to the adult human immediacy literature), but this is an area that would benefit from further research.

4.3.2 Recall Results

Recall results are based on the 10 recall questions presented to all participants; scores are given as the correct proportion of answers, i.e., 8 correct answers = 0.8. Recall scores and confidence intervals can be seen for each condition in Table 4.3 and are represented graphically in Figure 4.3.

To explore hypothesis H1, a comparison in recall was made on the adult data between those observing the high and low immediacy robot conditions. Two groups were used with independence of observations, and a continuous measure. Distributions did not significantly deviate from normality (Kolmogorov-Smirnov test; $p > .05$) and had homogeneity of variances (Levene's test; $p > .05$). For this reason, two-tailed independent samples *t*-tests are used to analyse the results. No significant differences at the $p<.05$ level were found; $t(16)=-0.577$, $p=.572$. However, significant differences are found for the child data. A two-tailed independent samples *t*-test reveals that recall is higher in the high immediacy robot condition ($M=0.58$, 95% CI [0.52,0.64]) than in the low immediacy robot condition ($M=0.49$, 95% CI [0.46,0.53]); $t(53)=2.006$, $p=.011$.

These results provide partial support for hypothesis H1: recall will be greater when the character reading the story is more nonverbally immediate. It can be seen that this holds true for the children, where recall is greater in the high immediacy robot condition than in the low immediacy robot condition, in accordance with this condition being perceived as more immediate. However, there are no significant differences in recall between the conditions for adults. This is likely due to a ceiling effect with adults because the recall questions were designed so that they were suitable for children. This may have made them too easy for adults overall, leaving limited space to show differences between conditions. If the questions were more difficult and exclusively targeted towards adults then it is possible that differences would be found. The partial support for H1 and replication of findings from previous studies of nonverbal immediacy – using robots – provides a proof-of-concept for the approach.

No support is found for hypothesis H3: that higher individual perception of nonverbal immediacy will lead to greater recall for that individual. Correlations between nonverbal immediacy ratings and recall scores are not significant for children ($r(81)=-0.047$; $p=.673$) or adults ($r(29)=-0.188$; $p=.311$). Indeed the correlations themselves are in the opposite direction (although only with a small magnitude) to that which was expected. This would

Condition	Adult *M*	95% CI	Child *M*	95% CI
High immediacy robot	0.80	[0.69,0.91]	0.58	[0.52,0.64]
Low immediacy robot	0.83	[0.76,0.91]	0.49	[0.46,0.53]
Human	0.79	[0.73,0.84]	0.63	[0.56,0.70]

Table 4.3: Mean recall scores by condition

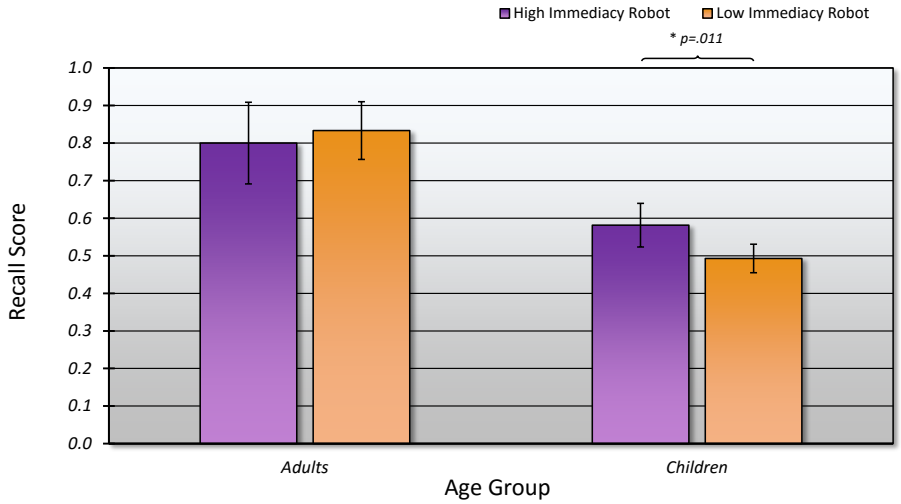

Figure 4.3: Recall scores for high and low nonverbal immediacy robot conditions relating to hypothesis H1. Children recall significantly more information when the story is read by a robot with higher nonverbal immediacy. * indicates significance at the p<.05 level. *Error bars* show the 95% Confidence Interval

suggest that in this study, the rating of immediacy at the individual level has less of a bearing on recall than the average as judged by the group, but there is not enough evidence here to explain why this occurred.

4.4 Discussion

The established research field of *nonverbal immediacy* links behaviour to learning gains in a measurable and comparable manner (Chapter 2). The evaluation in this chapter applied behaviour based on nonverbal immediacy cues to a social robot. It was found that both children and adults perceive the immediacy of a robot designed to have low and high nonverbal immediacy behaviours as intended, which confirms and extends prior work in HRI (Szafir & Mutlu, 2012). Additionally, both children and adults ranked the nonverbal immediacy of robots and humans in the same order, although children's raw scores were more tightly grouped. This gives rise to the possibility that much of the nonverbal immediacy literature, which has mostly been conducted with adults, would also apply to children.

Recall of a short story improved significantly for children when the robot reading the story was more immediate in behaviour, which does indeed confirm the hypothesis derived from nonverbal immediacy literature, based on human-human studies showing the same effect (Goodboy et al., 2009; Witt & Wheeless, 2001). No significant difference in recall was observed in the adult data, but this may be due to the relative lack of difficulty of the recall test, which had been designed specifically for children.

The following subsections will discuss the findings here in the wider context of research conducted in HRI and HHI. First, the impact of individual characteristics will be discussed in relation to hypothesis H3, which was not supported. Secondly, the possible impact of novelty on the perception of behaviour and recall will be explored. Thirdly, potential shortcomings of nonverbal immediacy as a measure for characterising interactions are raised. Finally, the lessons learnt from this study in applying nonverbal immediacy measures to HRI will be presented, along with a consideration of the influence of the study design on the findings.

4.4.1 Students as Individuals

Out of necessity, most experiments observe the learning of large samples of students, meaning that the effect is seen on average, but does not necessarily apply to all students. All children are individuals, with their own characteristics, preferences for subjects and learning styles. It may be that there are some educational scenarios, topics, or children, with which technology is more suited to assisting (Dede, 2009). Some children may be impacted to a degree related to their personality (and their 'need to belong'; Pickett et al., 2004), or their learning style (Witkin et al., 1977), which can affect their sensitivity to social cues.

Gender could also have an impact on learning and the use of social cues. It has been found in both virtual environments (Bailenson et al., 2001, 2003, 2005) and physical environments (Bull & Gibson-Robinson, 1981) that males do not utilise gaze cues in the same way as females; or if they do, it does not manifest in behaviour change or learning. The gender of the teacher, at least in virtual environments, does not however seem to impact on the learning which takes place (Baylor & Kim, 2004).

In the evaluation, support was not found for hypothesis H3, which sought to link individual perceptions of the robot behaviour (as measured through nonverbal immediacy)

to recall scores. It is suggested that this may be because the nonverbal immediacy scale does not cater for the many other variables between individuals that may influence their learning. However, this does not reduce the utility of nonverbal immediacy as a characterisation of robot social behaviour, with differences in robot behaviours clearly demonstrated as part of the manipulation check. Instead, there is possibly the need to further develop means of including perceptions of robot behaviour into broader models of learner characteristics (discussed further in Chapter 11).

4.4.2 The Novelty Aspect

It is necessary to acknowledge that the use of social cues is only partially responsible for positive learning outcomes. The approach, content and assessment of teaching contributes significantly to the learning process (Coe et al., 2014), as does the knowledge of the teacher (Hill et al., 2005) and their beliefs towards learning (Askew et al., 1997). Of course, the students play an equal part in learning too, with aspects such as their emotion playing a role in the process (Garner, 2010). Teachers and students often have long-standing relationships; these relationships allow for familiarisation with teaching styles and learning styles, which is beneficial for learning: when teacher turnover increases, attainment scores have been shown to drop, evidencing the importance of consistent relationships (Ronfeldt et al., 2012). This highlights the need for long-term interaction if using social robots to assist in education, alongside thorough development of learning materials.

The majority of the studies considered as part of the background for this work (Chapter 2) only look at single interactions, rather than interactions over time. There is evidence for changing preferences (and thus possibly changes in subsequent learning outcomes) over time, as seen in Wang et al. (2010). Of course, a relative lack of long-term data in HRI is understandable because of the immense challenge in enforcing methodological rigour over extended periods of time and the ethical implications of using atypical conditions (such as the low immediacy robot condition from the evaluation in this paper) in real-world learning.

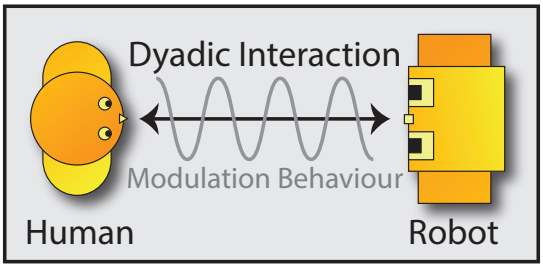

Figure 4.4: Representation of the role of social cues in dyadic HRI. Social cues are used as modulation behaviour within the interaction.

4.4.3 Nonverbal Immediacy and Interaction

Due to the potentially great benefits of using robots as tutors in one-on-one interactions (Bloom, 1984; VanLehn, 2011), and the possibility of personalisation in such contexts, this seems to be an apt means of applying robots in education. Whilst nonverbal immediacy addresses how competent a speaker is at communicating towards others, i.e., how well a teacher can convey information to students, in one-to-one tutoring, it is important to be competent at two-way communication as well. As such, it may be that the approach taken in this chapter would need adapting for one-to-one tutoring, incorporating more principles from dyadic interaction work.

Social behaviour plays a key role in dyadic interaction and on the outcome of communication within a dyad. The role of communication, or the social interaction within the dyad, in such a scenario is posited to be "the mutual modification of two ongoing streams of behaviour of two persons" (Beebe et al., 1992). The behaviour of one party affects the behaviour of the other. In this view, social cues are used as part of the modulating behaviour in this process (Figure 4.4) and can therefore be utilised in many processes influencing education.

The joint modification of behaviour within the dyad gives rise to the need for regulation and alignment of behaviour in order to simultaneously transmit and receive information (Jaffe et al., 2001). All parties engaging in a social interaction must continually adapt the social cues they are using in order to effectively construct the interaction (Green & Weade, 1985); for example, verbal turn-taking must be regulated through the use of various social cues (Beebe et al., 1992). Such regulation is important in learning interactions, indicating when it is appropriate for learners to ask questions, and when it is time for them to receive information; learning is more challenging without social cues or conventions to manage this turn-taking (Nicol et al., 2003). This simple coordination in interaction is vital and has

been shown to influence cognition from infancy (Jaffe et al., 2001). Even in unstructured interactions with robots, children appear to actively seek such turn-taking in interactions (Baxter et al., 2013).

These kinds of interaction phenomena are not catered for in nonverbal immediacy measures. The evaluation in this chapter saw positive results, but the interaction between the robot and the humans was largely in one direction (the robot instructing the humans); the robot was not responsive to human social cues or behaviour. This is an area which needs further exploration in HRI: the question is when the interaction becomes more interactional than those presentational behaviours considered in the present study, do immediacy principles hold, or are additional behaviours (such as turn-taking policies) required? In the absence of further evidence in such contexts, the application of the nonverbal immediacy metric provides a suitable basis for initial investigation; this will be conducted in later chapters.

4.4.4 Using Nonverbal Immediacy in HRI

Whilst the evaluation in this paper had positive results and confirmed (or partially confirmed) two of the three hypotheses, it should be made clear that there are limitations imposed by the study design which could inhibit how well these findings translate to other scenarios. The human condition was shown through a video, whereas the robots were physically present. This means that a comparison between the recall and nonverbal immediacy scores from the human and the robot conditions could be influenced by embodiment, or social facilitation effects (Zajonc, 1965). It should be noted that in this study, there is no direct comparison between these conditions: comparisons are made within robot conditions, or from children and adults, but not between the human and robot conditions.

Generally, the adult raters have high reliability levels, which reflects the behaviour seen in the literature. That this applies to ratings of robot behaviours indicates the applicability of the metric. Whereas the alpha statistic is lower for children, there are two points of note. Firstly, there remains a reasonable consistency for the ratings of the human condition – this extends the literature by showing the ability of children (in addition to adults) to use the nonverbal immediacy metric. Secondly, for both children and adults, there was agreement in the ordering of relative immediacy levels between the conditions – this indicates that

the nonverbal immediacy scale is sensitive enough for the present study, for both adults and children.

A number of caveats apply however that require further investigation. A high reliability score is found for the adults who saw a robot condition, but this is not so high for those who saw the human condition. This may be due to relatively low subject numbers when considering only the human condition (13 subjects), where inconsistency from one or two individual subjects could have a large impact on the alpha value. The reliability for the human is higher for children than for adults, suggesting the difference in subject numbers could be a factor. Alternatively, it could be a result of embodiment: the robot conditions were seen in person, whereas the human was shown on screen, which may have influenced the reporting of social behaviour on the questionnaire.

The Cronbach's alpha statistic for the children who saw a robot condition is considerably lower than that of the adults. This is not so surprising, given the complications highlighted in the literature of using questionnaires with children (Borgers et al., 2000). However, it may also be a product of limitations in robot social behaviour. Cronbach's alpha measures the internal consistency of questionnaire items. Whilst some inconsistency is likely due in part to child interpretations of negatively worded items (Borgers et al., 2004), there are some items within the questionnaire that the robot behaviour itself is probably not consistent in. For example, the questions related to smiling and frowning are opposites of each other in terms of calculating a value for the scale, but could both be answered as 'never' performed, as the robot does not have moveable facial features. Such a response would provide maximum inconsistency between these items. This would not necessarily reflect the reliability of the questionnaire, but a limitation in the ability of the robot to implement all of the questionnaire items. The same argument could be made for the items concerning touch – it could be considered that the robot never touches the observer, whilst also not 'avoiding' touch, as the question is worded. Inclusion of these two behavioural elements (that were not possible in the evaluation here) in subsequent work exploring the use of nonverbal immediacy for characterising robot social behaviour would likely yield higher reliability scores. For these reasons, it is argued that nonverbal immediacy provides a suitable metric for use with children in this context, given that is has the sensitivity to detect differences between groups of children, and the advantages provided by the metric in providing a guideline set of cues for robot behaviour design (as described in Chapter 2).

The interaction was also over a very short period of time (approximately 4-5 minutes) and the measurement of learning was through recall. Although recall is a fundamental element of learning, it is very different from understanding or applying knowledge, or from the higher dimensions of learning as defined in the revised version of Bloom's taxonomy (Krathwohl, 2002). Later chapters of this thesis will investigate the use of nonverbal immediacy in slightly longer interactions, and in dyadic contexts, with generalised cognitive learning as a measure (Chapter 7; Kennedy et al., 2015e).

4.5 Summary

Nonverbal immediacy can be used to characterise social behaviour through observer-reports on the use of social cues, such as gaze and gesture. This chapter built on HHI and HRI literature introduced in Chapter 2 relating to these cues, which were implemented in an evaluation that compared an intended high nonverbal immediacy and a low nonverbal immediacy robot. A human condition was also included to link the work here to existing nonverbal immediacy literature. Several hypotheses derived from the nonverbal immediacy literature were confirmed. Both children and adults judge the immediacy of humans and robots in a similar manner. The children's responses were more varied than the adults, but it was still possible to identify a significant difference in their perception of the social behaviour between the two robot conditions. Children also recalled more of the story when the robot used more nonverbal immediacy behaviours, which demonstrates an effect predicted by the literature. While there are some limitations in the measure, it is proposed that nonverbal immediacy could be used as an effective means of characterising robot social behaviour for human-robot interaction, for both adult and child subjects. Additionally, the findings here support the approach taken in Chapter 3: to use adults as post-hoc evaluators of robot nonverbal immediacy, given that their perceptions reflect those of the children to some extent.

Chapter 5

Robot Embodiment and Child Learning

Key points:

- An experiment was designed to explore differences in child learning and social behaviour in response to a physical robot, or a virtual robot (on-screen).
- A novel dataset was created for use as part of a category-based sorting game.
- The dataset used introduced a bias in the results, but this was accounted for.
- Results do not reveal any differences in learning between the two conditions, but differences are found in children's social behaviour towards the robots. The physical robot attracts more gaze from the children than the virtual robot, demonstrating the promise of physically present social robots for dyadic tutoring.

Parts of the work presented in this chapter have been published in Kennedy et al. (2014a) and Kennedy et al. (2015b)[1]. The final publications are available from Springer via http://dx.doi.org/10.1007/s12369-014-0277-4 and the ACM via: http://dx.doi.org/10.1145/2559636.2559820

[1]Note about technical contributions in this chapter: the author used software from the ALIZ-E project for the touchscreen, basic robot functionality for pointing at and selecting images on the screen, and low-level robot control. The author contributed the material used within the touchscreen software, some of the robot to touchscreen communication, the high-level robot social behaviour programming, and the porting of the robot software to the virtual robot.

As stated in Chapter 2, the application of social robots to the domain of education is becoming more prevalent. However, there remain a wide range of open issues, such as the effectiveness of robots as tutors on student learning outcomes, the role of social behaviour in teaching interactions, and how the embodiment of a robot influences the interaction. This chapter presents a study intended to explore children's behaviour towards a robot tutor for children in a novel guided discovery learning interaction.

Chapter 2 introduced literature relating to the embodiment of robots. It was suggested that the physical presence of a robot can have a large impact on human behaviour, possibly due to the social presence theory (Biocca et al., 2003) and a social facilitation effect (Triplett, 1898; Zajonc, 1965). A positive effect of using a physical robot on performance has been shown in different task contexts with adults, including in a negotiation game (Bartneck, 2003) and solving logic puzzles (Leyzberg et al., 2012). However, with children, studies have shown no significant differences in learning (Kory Westlund et al., 2015), do not control for prior knowledge (Han et al., 2008), or do not control for exposure time to material between embodiment conditions (Hyun et al., 2008). However, social behaviour responses of children have been observed due to differing robot embodiments Looije et al. (2012). The positive effects observed with adults in addition to the social response of children to physical robots suggests that this is an area worth further experimentation.

To address this gap in the literature, the study in this chapter seeks to directly compare robot behaviour when instantiated on screen against on a physical robot. Prior knowledge and exposure time are controlled for. While constrained compared to that of the human, the robot tutor behaviour demonstrates sensitivity to the behaviour of the child, and emphasises the structured self-discovery of the subject matter to be learned by the child. A novel set of information for the children to learn was devised to ensure that the children involved would have no prior knowledge and so would start at the same experience level. There are two primary aspects of interest: firstly, whether the embodiment of the robot impacts on how much children learn, and secondly, how the children behave in the interaction in response to the tutoring strategy of the robot.

In many HRI studies there has been a focus on prescriptive tutoring, with the robot providing instructional lessons to subjects; a 'teacher-centred' approach to learning, for example Leyzberg et al. (2014). However, educational literature suggests that a 'learner-centred' approach confers many advantages; for example, learners can gain a deeper understanding

of the material and can be more motivated due to an increased responsibility for their own learning (Spencer & Jordan, 1999). Such an approach is taken in Kanda et al. (2012), for example, where children undergo collaborative learning with a robot in a variety of group and individual lessons.

The learner centred approach taken in this study, 'guided discovery learning', has overlaps with the collaborative learning seen in Kanda et al. (2004), but also some important differences. In collaborative learning, interacting partners are often peers. However in guided discovery learning, one of the interacting partners has more knowledge and can therefore guide the learner towards a correct solution. Learners must generate their own hypotheses, which they then test, and analyse the results, which uses skills that would not be developed when the necessary information is simply presented by a teacher (De Jong & Van Joolingen, 1998). It has also been suggested that this type of learning can promote a better understanding of a domain when compared to teacher-centred learning (Alfieri et al., 2011; Van Joolingen, 1998).

In a similar manner to collaborative learning, the teacher initially delivers enough instruction for problem solving to commence. However, instead of providing a lesson when learners get stuck, the teacher will help to guide the learner towards the correct solution by scaffolding analysis of the decisions the learner made surrounding hypothesis generation and analysis, with the aim of improving in the next "hypothesis–test–analyse" cycle (De Jong & Van Joolingen, 1998; Harvel, 2010). Such an approach requires the robot behaviour to be responsive to the child not only in the task-related actions, but also in social behaviour. The robot needs to be sufficiently social to maintain the child's engagement throughout the interaction as the exploration is led by the child. The robot has key guidance for the child, and if they lose interest in the robot, then this could negatively impact their ability to succeed in the learning task. The robot behaviour must also be appropriate given the context, and to this end, a human model is used to derive the social behaviour; details are provided in Section 5.2.5.

The age of the subjects used in this experiment had to be carefully selected in order to make sure that the children had the cognitive skills to direct the exploration and motivate themselves to solve the problem presented. With the assistance of teaching professionals, it was decided to use children of around 8 years old. This age is quite novel in educational interactions, with most studies using subjects aged 10 and older (for

example Kanda et al., 2012; Leyzberg et al., 2014; Pereira et al., 2008; Saerbeck et al., 2010). Since the necessity of real robots (as opposed to virtual agents) in education has not been definitively established in the literature, the effect of robot embodiment was assessed here. The aim of the experiment was to evaluate differences in child learning and social behaviour in response to a physical robot, or a virtual robot (on-screen), i.e., the robot embodiment.

5.1 Hypotheses

The purpose of the study conducted here was primarily to explore children's responses to robot behaviour across different embodiments in a novel guided discovery learning task. This means that the central hypotheses are based around the child's behaviour. Given that the interactions are educational, part of the validation also lies in how well the children learn, particularly with regards to differences between embodiments. Other work, such as Looije et al. (2012), has found differences in gaze behaviour between embodiment conditions and it would be reasonable to predict that the same will be found in this study, despite substantial differences in context. Prior studies with a similar task structure and hardware configuration (robot with a large touchscreen), for example Baxter et al. (2013), have found that the children will gaze more towards the touchscreen than the robot. However, if the robot behaviour is sufficiently social, then the children will remain engaged with the robot throughout the interaction. This can be examined through child responses to the robot's behaviour and activity in the task. Other studies have also found that the physical presence of a robot causes an increase in learning gains, for example Bainbridge et al. (2008); Leyzberg et al. (2012). Again, we predict the same effect here. From this background, the hypotheses derived for the study are enumerated below:

H1: The real robot will attract more gaze than the virtual robot from the children.

H2: The children will remain engaged with both the robot and the task throughout the interaction.

H3: Children will gaze more towards the touchscreen than the robot, but they will still pay attention to, and respond to, the robot's behaviour.

H4: Children will exhibit greater learning gains after interacting with a physically embodied robot.

5.2 Experimental Setup

The study design compares an Aldebaran NAO under two embodiment conditions: physically present and displayed virtually on a large screen. The conditions are employed in a between-subjects design. The study was informed by numerous pilot studies which explored the assessment of children's learning when interacting with each other and an interaction mediator, the Sandtray touchscreen (Baxter et al., 2012). It was decided that the most appropriate task to assess learning would be an adaptation of the sorting task with which several other experiments have been run (Baxter et al., 2013; Kennedy et al., 2013). Previous experience of using this task for cHRI has led to the development of a practiced experimental protocol which serves as a solid foundation for use in this work.

5.2.1 Participants

Full permission to take part and be recorded on video was acquired for all participants. In total, 37 interactions took place, however, nine of these were not suitable for analysis. One child asked to stop before the interaction was completed, whilst in the other eight cases the experimental protocol was not followed. The breaks in protocol included technical issues with the robot/mediator, one child leaving to go to the toilet and one instance of Wizard error. As a result, 28 child-robot interactions were completed and recorded (11M, 17F, aged 7 or 8 years old, M=7.9, SD=0.31).

A further two interactions could not be included in the learning assessment because an incorrect dataset was displayed, or logging did not complete correctly during either the pre- or post-test. 26 pairs of pre- and post-tests were analysed in terms of learning and correlating social behaviour to learning outcomes (11M, 15F, age M=7.9, SD=0.33).

5.2.2 Conditions

The address the hypotheses, a two condition, between-subject design was employed for the study. The embodiment of the robot was swapped between the two conditions:

1. A 'real', physical Aldebaran NAO robot (n=15 for interaction analysis; 14 for learning assessment)

2. A large monitor displaying the virtual robot, an Aldebaran NAO from the Chore-

Figure 5.1: Side-by-side images of the virtual (*left*) and real (*right*) robots used for this study: the Aldebaran Nao. The images are stills taken from one of the cameras used for filming. Scaling has been kept consistent between the two images so that size comparisons can be made.

graphe software (n=13 interaction; 12 learning)

The use of these two embodiment conditions allows evaluation of the experiment hypotheses, which all assume the presence of an agent (be it real or virtual) to interact with the child. The virtual robot acts as a control for the real robot, making it possible to explore the differences between the embodiments.

The robots in the two different conditions were made as close to the same size as possible (Figure 5.1). It has been found previously that when children interact with a virtual robot with the same morphology as a physical robot that they have already interacted with, they can see this as the same character (Segura et al., 2012). Although the children were interacting with the robot in only one of the embodiment conditions, there was a concern that their peers could tell them about the robot in the other condition. As such, the robots used were arbitrarily given different identities, using different gender-neutral names and different colours. The real robot had grey features and was named 'Pop', whilst the virtual robot had blue features and was named 'Crackle' (Figure 5.1).

5.2.3 Interaction Scenario

The experiment took place in a primary school in the U.K. The room used was a classroom that the children were familiar with, but was not in use by a regular class. As such, there was a large amount of space available to the experimenters who were also in the room at the time of the interaction. The experiment involved two pieces of novel technology for the children: the Sandtray and the Aldebaran Nao robot. The behaviour of these two

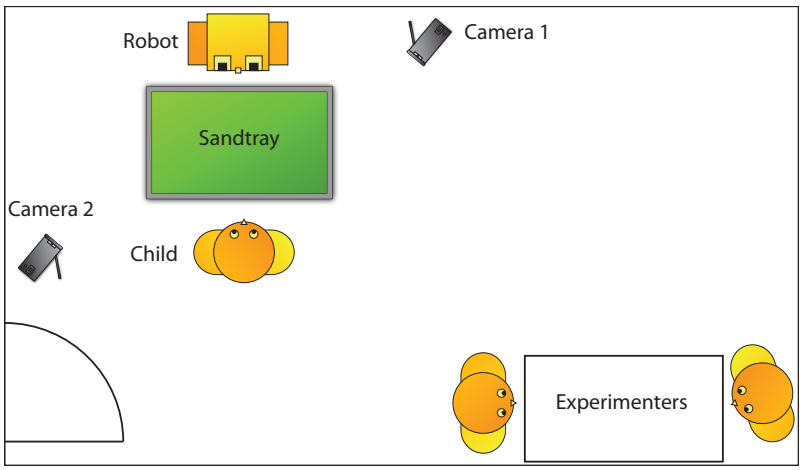

Figure 5.2: Schematic overview of the mediation device-centred interactions under investigation in this chapter. Two interactants (the child and the robot) face one another over the touchscreen. Two video cameras record the interactants during the studies. The experimenters are also in the room, but out of view of the child. Figure not to scale.

devices are outlined in Sections 5.2.4 and 5.2.5 respectively. Both the child and the robot can manipulate objects on the Sandtray. The Sandtray and robot were positioned such that children passing by the room could not see them and the child taking part in the interaction could not see the hallway or the experimenters, who were sat behind the child on the other side of the room. Two cameras were positioned around the Sandtray so that the behaviour of both the child and the robot could be recorded (Figure 5.2).

5.2.4 Alien Sorting Task

Pilot studies showed that overriding children's prior knowledge in a relatively short (5-10 minute) interaction time is extremely challenging; prior knowledge can play a large part in learning (Smith III et al., 1994). It has been shown that differing knowledge levels lead to different interpretations of a problem (Chi et al., 1981) and also require substantially different teaching formats to cope with this (Kalyuga, 2008). To remove these factors, a task with which all children have the same level of knowledge is therefore desirable. As such, the most practical solution is a task that the children have no prior knowledge about. To this end, a novel sorting dataset was created using aliens (as inspired by Lupyan et al., 2007).

An online 'alien creator' associated with a children's television programme was used to produce aliens of different morphologies and colours. Each alien consisted of six main body parts: a torso, a head, two legs, two arms, wings and a tail. There were three body types which each had different options for the body parts. Approximately 220,000 different

Figure 5.3: From *left* to *right*: (1) the 'orange planet' category image, (2) an outline of one of the aliens with shading to differentiate the body parts which can be manipulated (an actual image has not been used due to copyright restrictions), (3) the 'purple planet' category image

combinations could be created. Each body part can then be coloured using the full RGB colour space. From this, a random subset of 96 unique alien images were created. These were split into six training sets and two test sets of equal sizes (12 images per set).

The sorting task required the children to make a binary categorisation on the Sandtray touchscreen. A sorting rule was formulated that was based on one feature: given the wide range of possible features (and indeed combinations of features) that the rule could be based on, it is unlikely that the rule would be discovered by chance in the short period of the interaction. In this case, aliens with yellow legs would be correct if placed in the 'purple planet' category; all aliens which did not have yellow legs belonged to the 'orange planet' category (Figure 5.3). Twelve aliens would be presented to the children in each image set for categorisation and would be split equally between the categories; 6 aliens of each set would belong each category. Children could drag an alien across the screen and release it over the category that they thought it belonged to. The category icon would then change to display either a large green tick, or a large red cross depending on whether the categorisation was correct or not.

5.2.5 Robot Behaviour

A behaviour for the robot was created by analysing the behaviour of a human tutor (an education professional) when guiding a student unknown to them through a sorting task on the Sandtray, as in Kennedy et al. (2013). The human tutor was told that they would be assisting the child in guided discovery learning and that they could use any technique to guide the child, provided that they did not explicitly state the categorisation rule (as this would then no longer constitute discovery learning).

Two preparatory sessions were conducted with different children interacting with a human tutor. These underwent video analysis in order to get an objective measure of particular movements and vocalisations made in the interaction by the tutor. The most common verbal phrases, along with the timing and types of screen movements made were used as a basis for the robot behaviour. The result was a script that the robot would follow to introduce itself and the task to the child (full transcript available in Appendix G), along with a guiding behaviour for the discovery learning part of the task. The guided discovery assisting behaviour of the robot consisted of the following elements:

1. Verbal feedback specific to the image categorised by the child whenever a categorisation was made.

2. Advancing the screen library when all of the images in a particular set had been categorised, along with a general hint about the pattern.

3. If the child did not make a categorisation for 6 seconds, the robot would select an image, move it to the centre of the screen and make a verbal comment to the child about the item. This will be referred to as the robot 'highlighting' an image.

4. A gaze towards the child was also made when making a comment and highlighting a possible move.

Given that these child-responsive robot behaviours are directly inspired by the behaviour of the human teacher, we contend that it therefore demonstrates some key aspects of social behaviour in a tutoring context. The touchscreen provided the robot with information about the visible images, allowing the robot to make comments such as "why don't you try this one with purple wings?", or "pink legs worked in that one". This was the mechanism by which children were encouraged to think about the properties of the aliens that they were categorising and to lead them towards the correct solution. Of course, it would be straightforward to inform the children of the pattern and then see how well they recall it, but the benefits from guided learning, as outlined in Kuhlthau et al. (2007); Merrill et al. (1995) and the introduction to this chapter, would not be leveraged in this case.

The robot behaviour was structured such that the speech could be blocked depending on its importance and events on screen. The aim was for the robot to provide feedback on every move made by the child, however if the child then categorised images at a very quick rate, the robot speech would not be able to keep up. To solve this, a blocking period

of 2 seconds was put in place after each robot vocalisation. In cases where the child was making extremely fast categorisations (approximately one per second), two phrases could follow one another before the blocking period would begin. The speech blocking period could be ignored if the intended speech had been marked as *important* in the code. Speech which was part of the robot script and the general comments made at the end of each library (often key hints for solving the task) were classed as important speech which could ignore the blocking period in order to ensure equality across conditions and interactions. This speech planning strategy ensured that all children experienced the same structure to the interaction, whilst remaining adaptive to individual behaviour.

At the start and end of the interaction, alongside the scripted speech, the robot would make a number of predefined gestures and gaze upwards, towards the child, in a similar way to the human teacher. For the rest of the interaction, the robot would randomly move its head and body to give it a 'lifelike' feel. The random gaze was restricted to operate within a rectangular volume roughly directed towards the touchscreen while the child was moving images, as seen in previous human-human and human-robot studies with this task and the touchscreen (Baxter et al., 2013; Kennedy et al., 2013).

The robot behaviour was almost fully autonomous, with input required only to start the interaction and to start the post-test at the appropriate time. Following the protocol for a large number of HRI studies, a Wizard-of-Oz (WoZ) experimental technique was adopted to serve this purpose (definitions and descriptions of WoZ use in HRI can be found in Riek, 2012). A Wizard was needed simply to click a button to start the interaction once the child was present and to start the post-test once the time limit had been reached for the teaching behaviour (see Section 5.2.8 for more details of this). The Wizard was one of the experimenters located in the room with the child, as described in Section 5.2.3.

5.2.6 The Learning Task

The learning task required children to explore the images presented on-screen and discover, through trial and error, the rule that yellow legged aliens belonged on the purple planet. The robot would assist by making suggestions and providing hints about features to test, as described in Section 5.2.5. Without the robot's assistance, the children would only have ticks and crosses displayed on screen for each categorisation as feedback. This would make the task one of reinforcement learning; the screen providing the positive or negative

reinforcement on each categorisation. Children do not respond to feedback as effectively as adults and take many more trials to incorporate feedback into their strategy-making (Crone et al., 2004).

Additionally, given the balancing of the task, half of the information they see belongs in one planet, and half in the other. With no knowledge of the rule before they start, they are likely to get some categorisations wrong. This can lead to acquisition of erroneous information simply because they have had that thought before, even if negative feedback is provided; the 'mere-exposure' effect (Begg et al., 1985; Roediger & Karpicke, 2006).

Section 5.2.4 showed that there are around 220,000 alien body combinations that could be created, with each of the 6 body parts on each alien coloured differently, and each alien of a different size. This presents an overwhelming number of possible features on which to categorise the aliens. If utilised, the hints from the robot substantially reduce this search space, making the solution then possible to reach within the time provided. Given the short interaction time, the complexity of the dataset and the way children learn with just reinforcement feedback, it would be highly unlikely for them to find the correct solution without the help of the robot.

5.2.7 Measuring Learning

As is commonly applied in HRI studies examining learning, pre- and post-tests (as described in Dimitrov & Rumrill, 2003) were used to measure the learning of the child. The children were given as long as they liked to complete the pre- and post-tests, so that there was no time pressure. The pre- and post-tests were novel from the learning data and from each other. Using two different tests, the images of which were not present in the training data, means that learning is measured on novel data in both pre- and post-test conditions. This eliminates any biasing because of familiarity with the data.

The tests used each consisted of 12 aliens which had to be categorised into either the purple or the orange planet, as per the alien sorting task described in Section 5.2.4. Prior to the pre-test, the children had been introduced to the task by the robot; for the full script used here, please refer to Appendix A. The children had been instructed as to the nature of the task - sorting aliens into planets - but had no further indications as to what the categorisation rule may be based on.

The two tests were used in a cross-testing strategy; test 'A' was administered to half of the

children as a pre-test, who then took test 'B' as a post-test. The other half of the children took the tests the other way around. The test used for the pre- and post-test was swapped between each interaction, i.e., Child 1 would take Test A as a pre-test and Test B as a post-test, then Child 2 would take Test B as a pre-test and Test A as a post-test. Given the novelty of the material to be learned, this strategy allows analysis to determine whether learning gains can be attributed to differences in difficulty between tests, should any such differences unintentionally arise due to unknown aggravating factors. The category positions would also switch between tests (and indeed in the training sets) to avoid any conditioning to spatial locations.

5.2.8 Interaction Scenario

The interaction is largely unstructured and the children are free to work at their own pace. The robot provided verbal feedback on moves that the child made, and would suggest a move to the child if 6 seconds passed without the child making a categorisation. This allowed the child to involve the robot as much, or as little, as they desired based on how long they waited between moves.

The following interaction scenario was created by combining the human teacher model with the lessons learnt from earlier work:

1. The robot and touchscreen are introduced to the child by the experimenter. The child is told that they are free to stop at any time, or ask questions of the experimenters.
2. The robot introduces itself to the child and outlines the task to be completed.
3. The child completes the pre-test on the Sandtray.
4. The robot provides a 'clue' for the child and begins the guided discovery behaviour while the child categorises further image libraries.
5. After 5 minutes, the robot brings the guided session to a close and asks the child to complete the post-test, again on-screen.
6. Once the child has completed the post-test, the robot thanks the child and says goodbye.
7. The child is debriefed by the experimenters.

Due to the unstructured nature of the task, strict time limits could not be set for the interaction. As the guided discovery behaviour of the robot was the main variable being measured, an effort was made to keep this a consistent length of time. The target length of time was set at 5 minutes, as this was estimated to make the total interaction around 7-10 minutes long; an appropriate length as identified in the pilot studies. The Wizard would have a button to click once the child was nearing the end of an image library in the fourth minute of the teaching behaviour. This would then trigger the post-test script at the end of the current library.

The average length of an interaction was M=533 seconds, SD=58s. This was measured from the moment the child entered the experiment room, until the moment that they left. The average length of the learning phase was M=308s, SD=45s.

5.2.9 Video Data

All 28 videos were coded by one coder, with validation from a second coder. It is not practical to second code all of the video due to the amount of time this takes. Therefore, a proportion (18%) of the videos were second coded to validate the first coder, following the example set by Moshkina et al. (2014), Nalin et al. (2012) and Stanton et al. (2008). The videos for second coding were randomly selected from groups which ensured proportional representation between experimental conditions, experimental days and genders. The overall inter-coder agreement level, Cohen's kappa, across all tracks was an average of 0.78, which indicates *substantial* agreement (Landis & Koch, 1977). The coding scheme used was as objective as possible, based solely on overt child or robot behaviour and can be seen below. Table 5.1 shows the agreement for the tracks which will be used for analysis in Section 5.4.

- Interaction stage
- Child gaze
- Child gestures
- Child vocalisations
- Robot gaze
- Robot gestures
- Robot vocalisations

Track	Cohen's Kappa
Overall agreement	0.78
Child gaze	0.89
Child gestures	0.84
Robot gaze	0.63
Robot gestures	0.76

Table 5.1: Inter-coder agreement by track coded.

5.3 Learning Results

Twenty-six pairs of pre- and post-tests were logged during the interactions for analysis of learning. Two different tests were used as described in Section 5.2.7, named *Test A* and *Test B* for ease of discussion here.

When considering the population as a whole, a significant learning effect is found between the pre- and post-tests. Two groups are compared with paired values on a continuous measure, distributions did not significantly deviate from normality (Kolmogorov-Smirnov test; $p > .05$). As such, a paired samples t-test is used. A paired t-test reveals that the post-test score (M=9.12, 95% CI [8.13,10.10]) was significantly higher than the pre-test score (M=7.08, 95% CI [6.34,7.82]); $t(25)$=3.016, p=.006. However, when the learning effect is examined in more detail, a more complex story is revealed. Children who completed Test A as a pre-test and Test B as a post-test did not exhibit significant learning, whereas for the reverse (Test B to A), extremely significant learning was found (Table 5.2). This complication is due to the comparative 'difficulty' of the tests given no knowledge of the data and biases which are present, to be discussed in Section 5.3.1.

When considering the calculated bias values in the context of the tests, if the biases are followed, then a child would get 9 out of 12 correct on Test A and 7 out of 12 correct on Test B. This is reflected in the actual pre-test scores acquired: children scored average of 7.93 (95% CI [6.88,8.98]) for Test A and an average of 6.08 (95% CI [5.25,6.92]) for Test B. Additionally, this explains why learning effects measured from Test A to Test B may be hidden, and from Test B to Test A, may be amplified. Whilst the 'gain', the improvement in the score from pre- to post-test, is higher on average for the virtual robot than the real robot, this is not statistically significant (Table 5.2).

Condition A	**Condition B**	***t*-test used**	**Condition A *M* (*n*, 95% CI)**	**Condition B *M* (*n*, 95% CI)**	***p* value**	***t* (*df*)**
Pre-test A	Post-test B	Two tailed, paired	7.93 (14, [6.88,8.98])	8.43 (14, [6.74,10.12])	.627	$t(13)=0.498$
Pre-test B	Post-test A	Two tailed, paired	6.08 (12, [5.25,6.92])	9.92 (12, [9.00,10.83])	$<.001$*	$t(11)=6.823$
Virtual gain	Real gain	Two tailed, unpaired	2.42 (12, [0.65,4.18])	1.71 (14, [-0.60,4.03])	.614	$t(24)=0.510$
Pre-test A	Post-test A	Two tailed, unpaired	7.93 (14, [6.88,8.98])	9.92 (12, [9.00,10.83])	.005*	$t(24)=3.051$
Pre-test B	Post-test B	Two tailed, unpaired	6.08 (12, [5.25,6.92])	8.43 (14, [6.74,10.12])	.017*	$t(24)=2.558$

Table 5.2: Learning effect *t*-test results, comparing many different variables. 'Gain' refers to the increase in score between pre- and post-tests. The maximum score for all conditions is 12. * indicates a significant *p* value at the .05 level. All distributions do not significantly deviate from normality (Kolmogorov-Smirnov test; $p > .05$). For unpaired tests, observations are independent. All measures are continuous.

5.3.1 Learning Bias

Significant learning effects are observed when Test B was used as the pre-test, but the same was not found for Test A as the pre-test. To explore why this occurred, every pre-test image categorisation was analysed. It became apparent that whilst the datasets used had been carefully designed to be novel and prevent children having preconceptions, they were not immune to biasing effects. Clear patterns emerged in the way that the children categorised the aliens in the pre-test; the point at which they had no knowledge of the dataset material.

Upon further investigation of the literature it was discovered that children start to use colour as a predictor of category membership from an early age, as shown in Macario (1991). Combining this indication from the literature and examining the consistently incorrectly categorised aliens, the following hypothesis was formulated about the bias of colour in the dataset:

- Given no knowledge of the dataset, the greater the proportion of purple on an alien, the more likely it is to be categorised on the purple planet.
- Similarly, the same is true for orange on an alien and the orange planet.

Children were consistent at applying this bias and it was concluded that a bias-free dataset would be almost impossible to create. If the bias could be quantified then learning effects in spite of the bias, or on minimally biased images, could be evaluated. The equation shown in 5.1 was formulated as a measure of bias.

$$\mathtt{bias} = \%\mathrm{P} - \%\mathrm{O} \tag{5.1}$$

Where:

- %P = the percentage of pixels perceived to be purple out of the total number of coloured pixels in the image.
- %O = the same as the above for orange pixels.

This results in a bias value between -1 and 1. A value of 0 represents no bias, a negative value is a bias towards the orange planet and a positive value is a bias towards the purple

planet. The greater the magnitude of the number, the greater the bias effect. This equation assumes no bias when neither purple or orange are present in an image and takes into account the relative balance between purple and orange in an image; if they are equal then they will cancel each other out.

In order to evaluate the effect of the colour bias, a series of paper-based tests were given to a different group of children. Three different tests were used: two were the test sets from the main study and a third test was created to investigate aspects of the biasing hypothesis. Each test had twelve images of aliens in a vertical line in the centre, with the planets aligned to the right and left edges of the page. The side on which the purple and orange planets were placed was varied between the tests. A total of 54 tests were completed; 18 of each different test. 24 males and 30 females completed the tests, the children were aged 7 or 8 years old (M=7.2, SD=0.54).

The percentage of pixels which were purple and orange was counted for each of the 36 images used across the three tests. These values were then inserted into Equation 5.1 to provide a bias value for each of the images. Correlation between the calculated bias values and the actual percentage of children which categorised the image as purple or orange was then measured. Pearson's product-moment correlation coefficient shows a very strong correlation of $r(34)=0.761$ ($p<.001$) between bias value and percentage categorised as purple, and $r(34)=-0.761$ ($p<.001$) between bias value and percentage categorised as orange. The correlations are the inverse of one another due to the inverse relationship between the percentage categorised as purple and orange.

As the bias value holds a strong correlation to the actual child behaviour, this can be used to divide the test sets into groups based on their relative biasing. This is useful because it allows learning effects to be considered in the context of the bias. There are some clear clusters and a division of 0.2 and -0.2 was used to split the groups (Figure 5.4):

- $\mathtt{bias} > 0.2 = 9$ images biased towards purple
- $-0.2 < \mathtt{bias} < 0.2 = 9$ minimally biased images
- $\mathtt{bias} < -0.2 = 6$ images biased towards orange

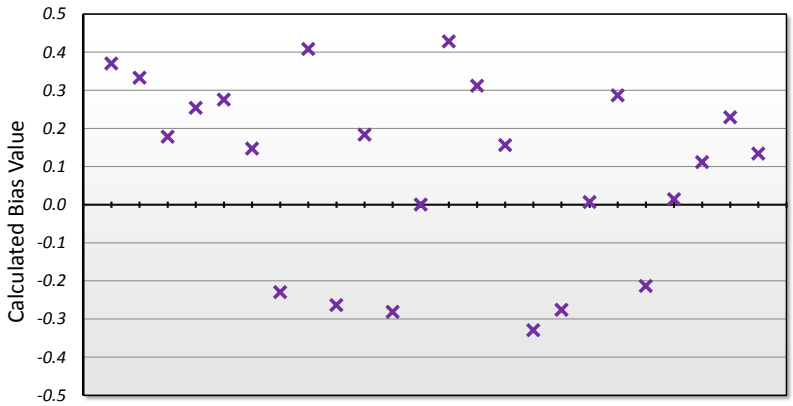

Figure 5.4: Calculated bias value for each of the images used in the experimental pre- and post-tests using Equation 5.1.

5.3.2 Accounting for Bias

Table 5.2 shows a significant increase between those children taking pre-test A and those taking post-test A *and* a significant increase between pre-test B and post-test B. This cross-comparison could be used to make an argument for overall significance of learning effects in spite of differences between the tests. However, it is more convincing to consider learning effects taking into account the known biases. The children's tests will now be evaluated in the context of the bias groups as laid out in Section 5.3.1. If learning is indeed present, the following hypotheses would be true:

H_0: Images with minimal biasing will be categorised more correctly in the post-test than the pre-test.

H_1: Images with large biasing towards a correct category remain unchanged.

H_2: The bias of images with large biasing towards an incorrect category will be reversed.

H_1 and H_2 make it necessary to divide the images with a large bias into two smaller groups: those where the bias leads to a correct categorisation and those where it leads to an incorrect categorisation. In regard to H_1, the number of correct categorisations will not change if learning is present, but it is hoped that the reasoning behind the categorisation changes from bias-influenced to knowledge-based. Unfortunately, it is not possible to measure the reasoning behind a classification given the current task.

In order to test these hypotheses, the percentage of correctly categorised images in the pre- and post-tests were grouped together based on the strength and direction of the bias. Four images are biased towards an incorrect categorisation, 11 images are biased towards

Bias group	Pre-test (95% CI)	Post-test (95% CI)	p value	U value
Minimal	62% [52,72]	72% [61,82]	.202	70.5
Bias to correct	72% [64,80]	79% [71,87]	.197	70.5
Bias to incorrect	9% [1,17]	77% [62,92]	<.001*	12.5

Table 5.3: Learning effect Mann-Whitney U test results grouped by bias. All observations are independent, but some distributions deviate significantly from normality (Kolmogorov-Smirnov test; $p < .05$), making the U test more efficient than t-tests. Average percent correct is shown for both pre- and post-tests. * indicates a significant p value at the .05 level.

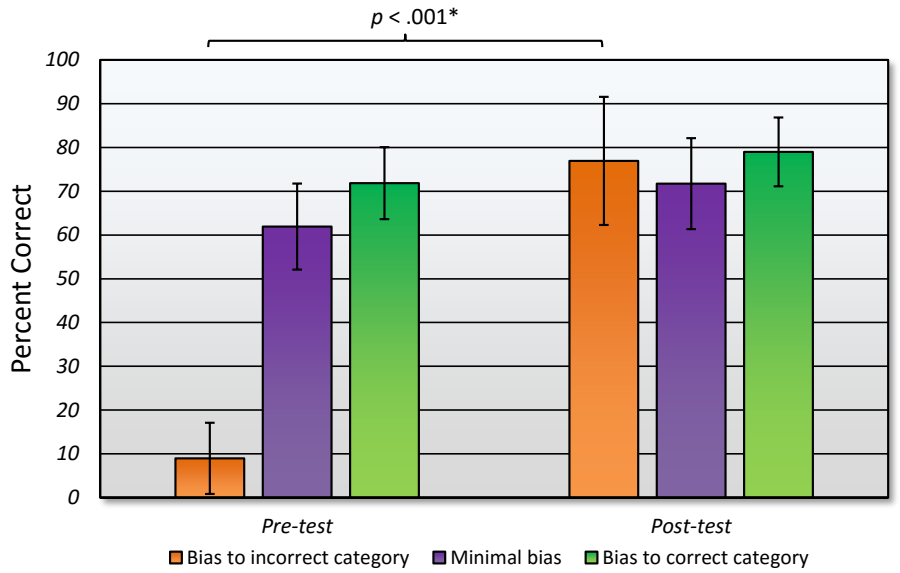

Figure 5.5: Pre-test and post-test percentage of correct answers for images, grouped by bias type and direction. *Error bars* show the 95% Confidence Interval.

correct categorisation and 9 images are minimally biased. Each image is categorised between 12 and 14 times.

The increase for minimally biased images is not significant, which does not support hypothesis H_0. The increase for images with a large bias towards a correct classification is also not significant, meaning that H_1 is supported. H_2 is also supported; a significant effect is found between pre- and post-test scores for those images biased towards an incorrect categorisation (Figure 5.5, Table 5.3). The bias groupings were considered across the two embodiment conditions, but no significant differences were found.

5.4 Behaviour and Embodiment Results

This section will analyse the relationship between the behaviour of the children, behaviour of the robot and the embodiment condition. Previous work suggests that embodiment will have an effect on the children's behaviour, as discussed in the introduction to this

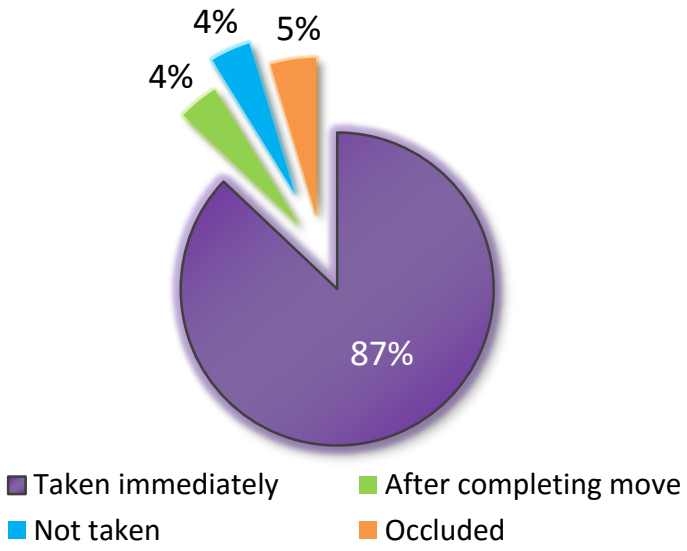

Figure 5.6: Child responses to robot suggestions to make a move. Children clearly comply with the robot's requests.

chapter and Chapter 2. This analysis is necessary in order to explore Hypotheses 2 and 3. This section will first analyse the children's compliance with the robot's suggestions, which relates to Hypothesis 2. The two subsections after will consider different aspects of the gaze behaviour of the child, which both relate to Hypothesis 3.

5.4.1 Compliance

The children clearly complied with the robot's suggestions for moves, as the percentages of responses below show (Figure 5.6). Even when the children were in the process of completing a move themselves, they were significantly more likely to stop their current move and follow the robot's suggestion than not (Kennedy et al., 2014a).

- 87% of the moves which the robot suggested were taken immediately by the children
- 4% were taken after the child had finished any move that they had already begun
- 4% of the suggested moves were ignored by the children
- 5% were occluded in the video analysis

Two groups were used with independence of observations, and a continuous measure. Distributions did not significantly deviate from normality (Kolmogorov-Smirnov test; $p > .05$) and had homogeneity of variances (Levene's test; $p > .05$). For this reason, two-tailed independent samples t-tests are used to analyse the results. When performing this

test, there were no significant differences between the two embodiment conditions for the number of moves taken immediately ($t(26)=1.568$, $p=.129$), although the real robot had a slightly higher average ($M=89.5\%$, 95% CI [77.7,101.3]) compared to the virtual robot ($M=77.7\%$, 95% CI [66.9,88.6]). No significant correlation was found between the number of suggested moves by the robot and the improvement in score between pre- and post-tests; Pearson's $r(24)=-0.137$, $p=.507$.

The high level of compliance with the robot's suggestions provides an indication that the children were engaged with the robot as well as the task throughout the interaction. This provides partial support for Hypothesis 2, although further evidence is required to fully support this hypothesis.

5.4.2 Gaze and Embodiment

When considering the full length of the interaction, there were a number of significant findings in the differences between the children's gaze and touchscreen gestures towards the real and virtual robot. Two groups were used with independence of observations, and a continuous measure. Distributions did not significantly deviate from normality (Kolmogorov-Smirnov test; $p > .05$) and had homogeneity of variances (Levene's test; $p > .05$). For this reason, two-tailed independent samples t-tests are used to analyse the results. Children interacting with the real robot ($M=5.19$, 95% CI [4.48,5.91]) make significantly more gazes towards the robot per minute than those in the virtual robot condition ($M=4.13$, 95% CI [4.48,5.91]), $t(26)=2.296$, $p=.030$. The length of each individual gaze is similar between conditions, so those in the real robot condition ($M=9.40$, 95% CI [8.36,10.44]) spend significantly more seconds per minute of interaction gazing towards the robot than those interacting with the virtual robot ($M=7.53$, 95% CI [6.37,8.70]), $t(26)=2.586$, $p=.016$. This result confirms findings from Looije et al. (2012) in a new context, and also supports Hypothesis 1.

Repeated measures ANOVAs are used to compare the effect of interaction time on the childrens' gaze towards the robot. Greenhouse-Geisser correction is applied due to violation of the assumption of sphericity (Mauchly's test $p < .05$). For the *virtual* robot condition, there is a significant effect of interaction segment on child gaze towards the robot at the $p<.05$ level for the three segments [$F(1.10,13.15)=7.504$, $p=.015$]. However, post-hoc pairwise tests with Bonferroni correction find no significant differences between the

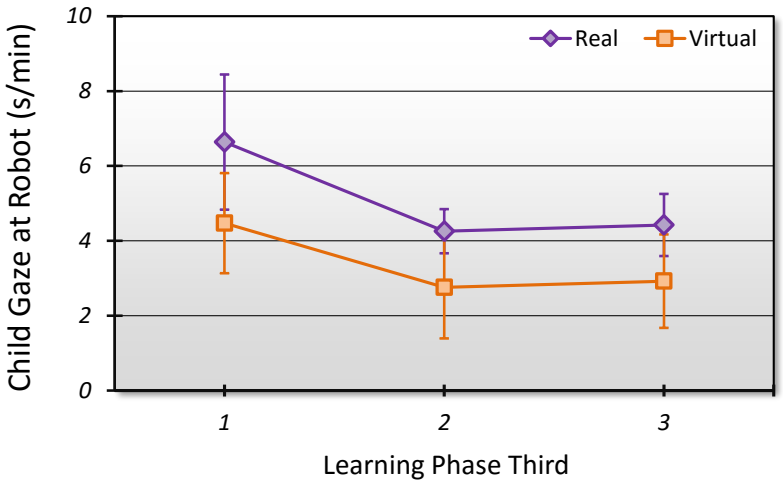

Figure 5.7: Seconds per minute that the child spends gazing towards the robot, split by learning phase third, comparing embodiment conditions. *Error bars* show 95% Confidence Interval.

individual stages (all $p>.05$). For the *real* robot, there was a significant effect of interaction segment on child gaze towards the robot at the $p<.05$ level for the three conditions [$F(1.16,16.23)=7.089$, $p=.014$]. Post-hoc comparisons with Bonferroni correction indicated that the mean score for the first segment ($M=6.64$, 95% CI [4.83,8.44]) was significantly different to the second segment ($M=4.26$, 95% CI [3.67,4.85]) and to the third ($M=4.42$, 95% CI [3.59,5.25]); $p=.013$ and $p=.023$ respectively. The second and third segments had no significant difference, $p=1$. This means that the gaze significantly dropped from the first to the second interaction segment for the real robot and then remained at roughly the same level as the second for the third. For the virtual robot, the same pattern is seen, but the changes are not as large. The comparison of these two curves can be seen in Figure 5.7.

It is suggested that the drop in gaze for the virtual robot is smaller than for the real robot because the starting level is lower than that of the real robot. Because of this lower starting point there is less of a reduction in gaze which is possible (a floor effect), whereas the relatively high starting point for the real robot gaze level allows for a greater drop. In the third segment, the gaze remains and roughly the same level as in the second segment for both conditions. This is an indication that once the children become accustomed to the social behaviour of the robot their interest in the robot drops off, reflected by their reduced gaze towards it (Baxter et al., 2014). When novel social behaviour is re-introduced for the post-test, the engagement level then raises again, in agreement with Szafir & Mutlu (2012).

5.4.3 Gaze and Robot Behaviour

Considering the interaction as a whole reveals a number of interesting results, but considering the interaction in terms of its component parts, as laid out in Section 5.2.8, allows a more thorough analysis and the exploration of behaviour over time. This has previously been suggested for use as a "proxy for engagement in the interaction or for the human's attribution of social agency to the robot" (Baxter et al., 2014). Gaze can be converted into seconds per minute values in order to normalise between individuals and allow for direct comparison.

The amount of gaze towards the robot varies a lot between the different segments (Figure 5.8). When the robot is directly addressing the child, the child gazes more towards the robot than when the robot is not addressing the child at all; a good example is the difference between when the robot is providing instructions and when the child is completing the pre-test. The gaze for the learning phase appears to be quite low in comparison to some of the other sections; whilst it is, this does not mean that the child is not paying attention to the robot. It is possible for the child to observe the robot's actions on screen and to get feedback from the screen, whilst also listening to the robot; this could explain the relatively low level of gaze towards the robot throughout this phase of the interaction. Another notable difference is seen between the gaze towards the robot during the pre- and post-tests; this will be discussed further in Section 5.5.2.

Of particular interest is the behaviour of the child during the main learning phase. As the stages are 3 related measures, a within-subjects repeated measures ANOVA was conducted to compare the effect of interaction time on the child's gaze towards the robot (both conditions combined). Greenhouse-Geisser correction is applied due to violation of the assumption of sphericity (Mauchly's test $p < .05$). The learning phase was split into thirds for comparison (Baxter et al., 2014). There was a significant effect of interaction segment on child gaze towards the robot at the $p<.05$ level for the three thirds [$F(1.14,30.87)=13.920$, $p<.001$]. Post-hoc comparisons with Bonferroni correction indicate that the mean score for the first segment (M=5.63, 95% CI [4.48,6.79]) was significantly different to the second segment (M=3.56, 95% CI [2.84,4.28]) and the third segment (M=3.73, 95% CI [2.99,4.47]), with p=.004 and p.008 respectively. No significant difference was found between the second and third segments, p=1. Therefore, children look at the robot significantly more in the first third of the learning phase, before dropping for the rest of the interaction.

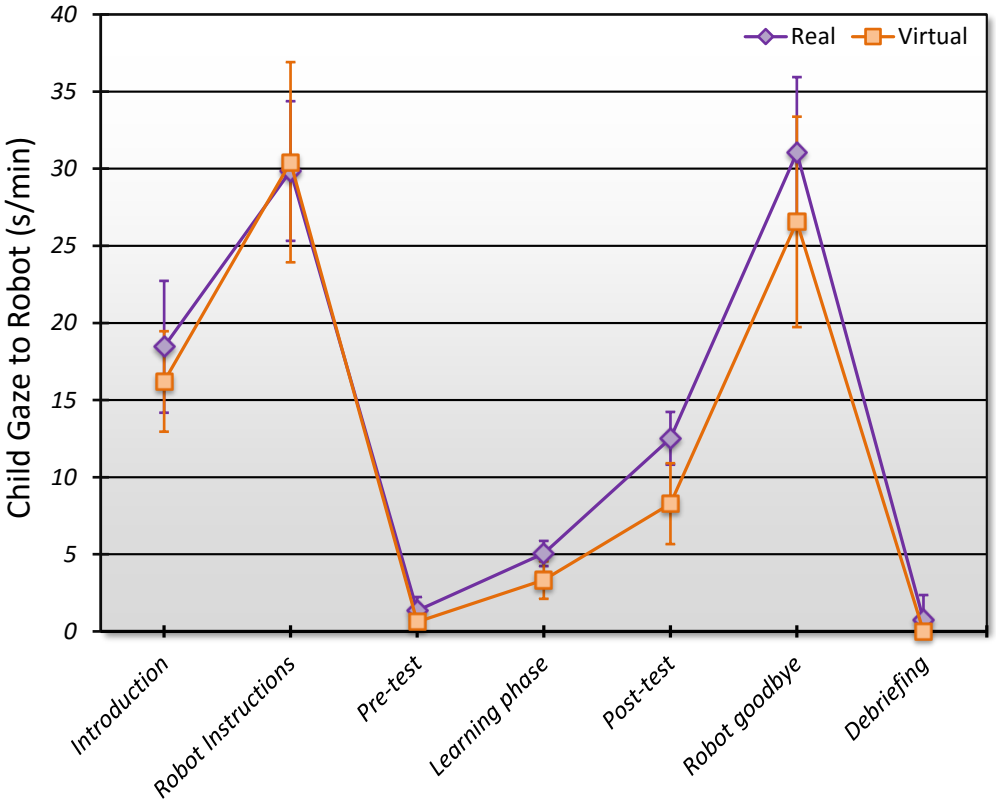

Figure 5.8: Seconds per minute that the child spends gazing towards the real and virtual robots, split by interaction segment. *Error bars* show 95% Confidence Interval.

5.5 Discussion

This section will discuss the overall learning significance in relation to the task and the robot, the lack of learning differences between the real and virtual robot conditions, and the significant behavioural differences in the response of children to the real and virtual robots. Addressing these points allows conclusions to be made in response to the hypotheses laid out in Section 5.1.

5.5.1 Embodiment and Learning

No significant differences were found in learning between those children who interacted with the real robot and those who interacted with the virtual robot. Other studies have found significant differences between different robot embodiments, for example Huang & Mutlu (2013); Leyzberg et al. (2012). In this case these effects were not found; thereby not supporting Hypothesis 4. However, this is in line with some work, for example Hasegawa et al. (2010). Given the context of the interaction, it is suggested that the robot's behaviour had a greater impact on learning than its embodiment.

The length of the interactions may have caused the lack of difference between the embodiment conditions. The learning phase of the interaction was an average of 309s; just over 5 minutes. This is a very short amount of time for learning, meaning that the task had to be simple. Additionally, the social behaviour of the robot was limited. The scripted elements of the interaction were relatively rich, but the main learning phase was repetitive and it indeed appears that the children lost interest in the robot as they became aware that the robot was not socially responsive (Section 5.4.3). If the robot could exhibit richer social behaviours during the learning phase then it may be that a greater difference between the children's reactions towards the robot would be observed, which may improve the learning outcome.

The learning differences between the real and virtual robots may also have been influenced by the novelty effect. Not only were the children facing a novel robot, in whichever form that may take, but also a large touchscreen. It is likely that even if there are differences in how children would respond to either of the robot embodiment conditions, they would be excited by the novel technology in either condition and therefore more likely to give the task their maximum attention, reducing any potential for difference between conditions in task performance.

One way to disentangle the novelty effect would be to carry out the experiment over a longer length of time so that the effect wears off. However, this would certainly require a change in the complexity of the task to prevent it quickly becoming boring for the children. Although a study of this nature could be interesting, it may not yet be as useful as it could be. It would be beneficial to establish the importance of a more socially responsive behaviour for the robot, and then consider the impact of this over time.

With a larger sample size it is possible that learning differences may then become more pronounced and could also be generalised. The difficulty would then be in recruiting enough subjects of the correct age, particularly given the challenges in recruiting subjects and running studies outside of the lab, as highlighted in Ros et al. (2011) and Walters et al. (2005b).

5.5.2 Social Behaviour and Embodiment

When considering the social behaviour exhibited by the children between the two conditions, the main difference was in the amount of time the children spent looking at the real

robot; they look at the real robot significantly more than the virtual robot. An increased amount of gaze towards a real robot when compared to a virtual one has been seen in other studies as well, for example Looije et al. (2012), and the increased mobility of the real robot has been suggested as an explanation, as in Breazeal (2004).

The ability of the real robot to enter into the child's field of vision whilst they are looking at the touchscreen is a great advantage. This can be used as a technique to direct the child's attention during the task, or to make sure that the child is paying attention (Szafir & Mutlu, 2012). This could be particularly useful in a more complex task where the robot's input is more tightly coupled to the learning outcome.

Because of the differences in embodiment and the subsequent lack of depth information when looking at the virtual robot, the virtual robot (M=15.4, 95% CI [12.8,17.9]) appears to gaze at the child significantly more than the real robot (M=10.5, 95% CI [8.2,12.9]) when normalised to s/min, $t(26)=3.029$, $p=.005$. It is surprising that this does not cause the child to look at the virtual robot more often in order to reciprocate this gaze. The real robot attracts significantly more gaze from the child than the virtual robot and if the robot behaviour were to be more varied then it is possible that the heightened levels of gaze seen at the start of the learning phase could be maintained. This could be used to argue that the real robot has the potential to be more socially engaging than the virtual robot and that this may lead to increased task performance in the future.

The mediator has a large effect on the social interaction which takes place; the mediator attracts the majority of the gaze from the child and in it's current form, the children can get all the information they need to play purely from the screen once the task has started. Something which may reduce the large disparity between gaze towards the robot and the touchscreen could be to remove any feedback elements from the touchscreen; if the feedback on the screen were to be removed, then the robot would be the child's only source of feedback, which may facilitate social engagement. An increased reliance on the robot for feedback could lead to more engagement with the robot from the child, which could improve learning gains.

Additionally, once the child understands the concept of the task, the robot's input is not necessary for completing the task presented here because of the mediator. The children can choose to exclude the robot from parts of the learning phase by taking moves quickly and preventing the robot from suggesting a move. If the coupling between the task and

the robot were tighter, it is likely that the behaviour of the robot would have a greater impact on the outcome of the task. From this perspective, the impact may also be more consistent, as the robot involvement would likely be more consistent as well, as opposed to the current setup where the robot input varies depending on how the child behaves. It should be noted that it is not being suggested that this consistency would result in a higher average performance increase; the hypothesis would remain that a more adaptive robot would result in greater child performance.

The results show that the children spend more time looking at the real robot than the virtual robot. It is likely that if the task was more spatially orientated, or required joint reference, the real robot would hold an advantage. In the task used, when the robot suggested a move, it was clear on the touchscreen which image the robot was suggesting, so the gaze of the robot was not needed to identify the object of reference. It may be that if the touchscreen did not make it clear which of the images the robot was pointing to, then the increased gaze towards the real robot may play a more important role in the learning outcome.

The amount of time per minute that the children gaze towards the robot during the pre- and post-tests has previously been highlighted as an interesting difference to discuss. Figure 5.8 suggests that there would be a significant difference between the amount of time the children spend looking towards the robot during the two testing phases. This is due to the inclusion of the post-test instructions from the robot to the child in the post-test segment. When splitting out this instructional phase, there is very little difference between the gaze towards the robot during the tests. During the pre-test, the children spend an average of M=1.0 seconds per minute (*s/m*) gazing towards the robot (95% CI [0.5,1.5]); this rises very slightly to M=1.3 s/m (95% CI [0.8,1.8]) during the post-test. The post-test gaze towards robot value had been inflated by the inclusion of the instructions in this phase. Whilst the instructions are being given by the robot, the child spends M=25.0 s/m (95% CI [21.1,28.8]) gazing towards the robot. This is almost half of the time and further supports the point made in Section 5.4.3 suggesting that the children gaze more towards the robot when it is directly addressing them and exhibiting novel behaviour.

Schermerhorn & Scheutz (2011) have demonstrated the complex interactions which occur between embodiment and other elements of robot behaviour. Whilst the analysis of an integrated system is always desirable, we would suggest that varying just one of the

dimensions at a time affords the ability to establish a hierarchy between factors which are hypothesised to have an impact on the results and also to make direct attributions between variables and outcomes.

In a similar way, Huang & Mutlu (2013) adopt a multivariate analysis approach to study the impact of a specific behavioural variable, in this case gesture, on knowledge recall. This attractive approach affords the ability to study several variables whilst keeping subject numbers low; often a great challenge for HRI research. However, when dealing with social behaviour, it remains to be seen whether these specific sub-behaviours being varied can be statistically extracted from a more complex behaviour and then successfully implemented into a new 'optimised' model.

5.5.3 Task Characteristics

The task that the children completed will have had a large effect on the learning which took place. The task is simple so that it is possible to be learnt within the short interaction time. As such, there is a very limited gradient in terms of the learning which can take place; children either figure out the pattern and do very well on the post-test, or they don't and they continue to sort the images according to the colour biases identified in Section 5.3.1. This means that subtle differences in learning are unlikely; the learning is often binary, which limits the variability between the post-test scores and therefore between conditions. A task which has a greater gradient of learning to measure on has a much greater resolution of measurement and can therefore provide more variability to make comparisons where subtle differences can be more pronounced. Equally, learning could easily be assessed over the full course of the interaction, rather than in just a pre- and post-test; this type of continual assessment is supported in educational literature (Guskey, 2003; Myers & Myers, 2007).

Furthermore, the unit of measure for learning and how learning is defined is important. Completion time of a puzzle has been used, as in Leyzberg et al. (2014). Whilst time provides a good resolution of learning steps, it is possibly too closely related to motivation, rather than knowledge gain (although there is an undeniable connection between the two) in the context of the task used for the experiment in this paper. Evaluation of skill over the course of the interaction, rather than just in a pre- and post-test setup, may provide more insight into the learning process.

The position of the robot around the mediator may also impact upon the learning outcome. The position of people around a surface has been studied elsewhere and correlations between seating positions and interaction styles have been made, e.g., Scott et al. (2004); Tang et al. (2006). In this case it is proposed that the seating position has not made a significant impact as the studies showing differences have been human-human and have been ambiguous in the way that the interaction partners have been presented to one another. In this experiment, the children were expecting to play a game with a robot which would be there to help them, so competitive behaviours seen elsewhere when interactants face each other are probably overridden by the context here. This is reinforced by the teacher and student roles assumed by the robot and the child.

The task used in this study was designed so that it would be completely novel to the children. The aim was to prevent preconceptions from influencing the learning taking place, as inspired by Lupyan et al. (2007). However, a side effect of this was an introduction of a colour bias which complicated the results. Additionally, the development of an entirely novel task in this manner removes a lot of the context from the learning. It is thought that context has a great impact on learning and transfer of knowledge (Tessmer & Richey, 1997). It is possible that the removal of a real-world context in the learning task inhibited the amount that children could learn, possibly contributing to the lack of learning difference between conditions.

The biases present in the dataset used highlight a trade-off between development of a novel task and the introduction of bias. A novel task was desirable here so that learning could easily be measured independently of preconceptions. However, in creating a novel task, biases were inadvertently introduced. For future work it is important that any novel dataset created is tested and validated to remove, or at least balance, any possible bias. Moreover, work in subsequent chapters will instead move away from the use of a novel task due to the complications that this introduces, and the lack of context surrounding the learning. Instead, a task which children could learn from in an interaction that more closely follows their academic curriculum, whilst not overlapping and confounding the measurement of learning is used.

5.5.4 Nonverbal Immediacy

The nonverbal immediacy ratings of the robot behaviour provide a further perspective on the learning results that were found in this study. 37 ratings of the real robot, and 35 of the virtual were acquired (for demographic details please see Appendix B). The real robot received an average nonverbal immediacy score of M=51.9 (95% CI [50.5,53.3]), with the virtual M=50.2 (95% CI [48.5,51.9]). These scores are not found to be significantly different when using an independent samples, two-tailed t-test; $t(70)$=1.517, p=.134.

The nonverbal immediacy metric considers overt social cues and provides validation that the social behaviour between the conditions could be perceived in the same manner despite the differences in embodiment (and therefore possible issues with identifying gaze direction in the virtual condition). Prior work has found that embodiment can make a large difference to task outcomes (Bainbridge et al., 2008; Leyzberg et al., 2014). The non-significant difference between the nonverbal immediacy scores of the conditions supports the suggestion that the robot behaviour is largely responsible for the learning outcome here, rather than the embodiment. This may have been due to task characteristics or novelty, as previously discussed, or possibly due to age. Previous studies that find differences due to embodiment largely deal with adults rather than children. It may be that embodiment effects are not so pronounced with children, however, further evidence would be required to confirm this suggestion.

5.5.5 Sample Size

The sample size used in this chapter is 28 after exclusions. When split across 2 conditions, this is half of the sample size per condition compared to those from the previous chapter (28 per condition). Even if no exclusions had taken place, there would still be fewer subjects in the study here when compared to the previous chapter. This has implications in terms of the confidence one can have in the results found here. Children are known to be highly variable in their performance in tasks (Siegler, 2007), and for this reason, to make strong conclusions from a study involving children, a larger sample size is often desired.

However, a larger sample size does not come without other considerations. Throughout this thesis, children are recruited through visits to their schools. It is often the case that

attitudes towards technology vary by teacher and the teacher's willingness to engage with technology will influence the attitudes of children towards the technology (Sariçoban, 2013; Teo, 2006). Teacher attitudes are in turn influenced by training (Teo, 2008), and the school ethos towards technology use (Sipilä, 2010). For this reason, it is preferable to conduct experiments with a single year group (or class) at a school. This maximises the chance of having children with similar attitudes towards technology, and also offers the possibility to balance for cognitive ability and prior knowledge.

The year group used in Chapter 4 was large, whereas the one in this chapter is more typical, hence the drop in sample size. Given the novelty of the robot technology, and the impact that the teacher can have on how children perceive their interaction with the robot (Baxter et al., 2015), using a single class or year group can provide tighter experimental control despite lower subject numbers. If children from more than one school were to be used, then it is suggested that several schools should be used instead and subject numbers would need to be greatly increased to average out interaction effects introduced from comparisons between children from different schools. This would be an ideal scenario, but it then takes substantially more time to run and analyse each experiment. For a lone researcher, while required to be constantly present in case of problems with the robot and to brief and debrief children to ensure experimental protocols are followed, such an endeavour would greatly limit the number of experiments that could be conducted and research questions that could be addressed. As such, the decision was made here to answer a greater number of research questions, but at the expense of larger sample sizes. This does have the effect of limiting confidence in the results as it becomes less clear that they are generalisable. It also becomes a greater challenge to detect subtle differences between experimental groups, with statistical significance less likely to be observed for small effect sizes. Again, this is a limiting factor when making firm conclusions from experiments.

5.6 Summary

It is clear that although the study was designed to prevent children from having prior knowledge about the test sets, there are still biases present in the test material. It is suggested that in a sorting task of this manner, it would be almost impossible to eliminate all of the possible biases. In this instance, the significant cause of bias could be accounted

for and quantified, thus allowing robust analysis in spite of these biases. It is important for HRI experimenters to consider the effect that such biases and preconceptions may have on any learning effects that they are trying to measure.

Varying the social behaviours exhibited by the robot during the learning phase could be a useful extension to this experiment. It was observed that the children seemed to lose interest as the main learning phase progressed and it became apparent to them that the robot behaviour was not socially responsive to their behaviour (like speech or gaze). This drop-off in apparent engagement signifies that the robot behaviour needs improvement, highlighted by Hypothesis 2 from Section 5.1.

In support of Hypothesis 3, regarding gaze and attention, children's compliance with the robot's requests (as shown in Section 5.4) demonstrates that they were paying attention to the robot despite most of their gaze being towards the touchscreen. This was entirely as expected, as informed by prior studies, e.g., Baxter et al. (2012), Kennedy et al. (2013).

Although no differences in learning between the embodiment conditions were found in this study, a number of reasons have been suggested as to why this was the case. These results did not support Hypothesis 4, that significant learning differences would be observed between embodiment conditions. This could be used as evidence for the robot behaviour over-riding embodiment effects (as supported through the nonverbal immediacy ratings of adults), or perhaps environmental factors impeding learning. Alternatively, it is suggested that if the task were to be more spatially dependent or have a greater resolution for measuring learning then differences in learning between real and virtual robot conditions may become more apparent.

To conclude, this chapter has contributed to the existing literature in the domain of HRI in educational interactions by exploring the effect of embodiment in a novel learning context. It is found that the real robot attracts more gaze than the virtual robot, but that there are no learning differences between these conditions. Additionally, results here have confirmed the value in considering child behaviour over the course of an interaction as a means of characterising the effectiveness of the robot behaviour. Further work needs to be done in order to improve the social behaviour for the robot to maintain the initially high levels of attention from the child throughout the interaction.

Chapter 6

Robot Social Behaviour and Child Learning

Key points:

- An experimental methodology was devised to teach and measure children's performance in prime number identification.
- 'Social' and 'asocial' robot conditions were derived from observation of a human tutor. These conditions were compared with a no-robot control, and a no-robot, no-lesson control.
- Children learn a significant amount from the robot, but not from the no-robot or no-lesson controls.
- However, children in the 'asocial' robot condition learn a significant amount, whilst those in the 'social' condition do not. Whilst initially surprising, nonverbal immediacy ratings may provide an explanation for these differences.

Part of the work presented in this chapter has been published in Kennedy et al. (2015d) [1].
The final publication is available from the ACM via:
http://dx.doi.org/10.1145/2696454.2696457

[1] Note about technical contributions in this chapter: the author wrote original software for the touchscreen, replacing the software previously used from the ALIZ-E project. This new software (or a variation thereof) continues to be used in Chapters 7 and 8. The robot code was also upgraded from Urbi v2.7.5 to v3.0, this was performed in part by the author and in part by the ALIZ-E project. All high-level robot social behaviour programming was done by the author.

The previous chapter explored the impact of robot embodiment on child learning. No learning differences were observed, but this may have been a product of complications introduced by the novel dataset used. However, there were clear differences in the social behaviour of the children towards the physically present, real, robot. This bodes well for the use of robots in educational interactions, given the connection between social behaviour of this nature and learning in the HHI literature (Richmond et al., 2003).

The influence of social behaviour on learning was explored in Chapter 2 from a mix of HHI and HRI literature. It was established that between humans, social behaviour has a large impact on learning. Both verbal (Gorham, 1988) and non-verbal (Richmond et al., 2003) aspects of behaviour can influence learning outcomes, with various guidelines presented for appropriate teaching behaviour. However, these guidelines are relatively broad and high-level, without specifically indicating certain aspects that social roboticists require for behavioural design. For example, Richmond et al. (2003) suggests that performing more gestures should lead to more positive learning outcomes, but the specific type of gesture and timing of these gestures is not specified.

Research conducted in HRI has sought to confirm some of these observations from human studies, but with results at a greater degree of specificity as required for behavioural design (Huang & Mutlu, 2013). Other findings from HHI literature have been confirmed with robots and adults, for example Szafir & Mutlu (2012), but it remains to be seen as to whether these findings also translate to domains with children. The robot behaviour in this chapter is derived from a human model completing the learning task with children so that an appropriate behaviour could be devised. The behaviour produced is then evaluate against an 'inverse' to provide a counter to what is considered as an optimal social behaviour (assuming that a human is an optimal model for social behaviour).

This chapter seeks to not only explore the effect that the presence that a robot has on the interaction taking place, but to also assess how the behaviour that this robot employs influences child social responses and learning outcomes. Whilst the previous chapter actively sought to avoid covering material included the school curriculum, this chapter instead places the work into the context of the curriculum. This is done to provide a clear relevance for the findings, and to maximise the chance of the children understanding why the subject is being taught (thereby improving their motivation to learn). Of course, this also requires prior ability to be more carefully controlled, but a novel topic within

the curriculum is selected. The methodology is based on the teaching strategy used by teachers to provide further validity to the approach.

6.1 Hypotheses

The aim for the study was to explore both embodiment and the social behaviour employed by a robot. It is useful to verify whether a robot brings benefits to the learning in this specific context: a sorting task, using the NAO robot, with children, in addition to evaluating the impact of social behaviour on learning (as a step towards establishing appropriate social behaviour for robots in such tasks). Prior work has suggested that the presence of a robot will result in greater learning gains (Han et al., 2005; Leyzberg et al., 2012); this drives the hypothesis for the embodiment differences (H1). Positive effects of social behaviour have been observed in other HRI studies such as Leyzberg et al. (2014) and Saerbeck et al. (2010), and psychology studies such as Atkinson et al. (2005) and Mayer et al. (2004). These prior works are used as the basis for the hypothesis about social behaviour (H2). The hypotheses are as follows:

H1: The presence of a robot will result in greater learning gains than when a robot is not present given equal information content.

H2: A robot with greater social skills will result in greater learning gains than a counter social robot.

6.2 Experimental Setup

Four conditions are used in a between-subjects design to explore the social presence of a robot and the effect of social behaviour: 2 without a robot, and 2 using a robot with differences in the social behaviour. The methodology of the experiment was designed based on previous work with child learning in sorting tasks, as in Chapter 5, and with sub-tasks leading to a combination of knowledge in a larger primary learning objective, as in Leyzberg et al. (2014). The following section will detail the participants involved, the interaction scenario, the task structure, the robot behaviour and the conditions used in the experiment.

6.2.1 Participants

A total of 53 children had permission to take part in the study. Due to technical issues, 8 of the children's data had to be excluded, leaving 45 children included in the study (23F, 22M). All participants were aged 7 or 8 and from the same year group at a primary school in the U.K. Participants were randomly distributed between conditions, whilst maintaining a balance of gender and mathematics ability (based on their teacher's assessment) between the groups. For the split between conditions, please see Section 6.2.3. Those in the robot conditions were requested for permission to film, which was granted in all but 2 of the cases. One video had to be excluded from analysis as it was not possible to see the child's eyes. Therefore, video analysis was conducted on 20 interactions.

6.2.2 Interaction Scenario

Interactions took place either in an unused classroom, or a relatively quiet public space in a primary school in the United Kingdom. The child was brought into the experiment area and would be sat facing a robot, an Aldebaran Nao, with a 27 inch touchscreen horizontally between them (Figure 6.1). A Microsoft Kinect was placed above and behind the robot to track the child's face. Two video cameras were also positioned around the setup: one to record the child's face and actions and another to record the robot's actions. The use of a touchscreen mediator (Baxter et al., 2012) allows a consistent, constrained environment, so the robot's social behaviour can be manipulated without impacting on the nature of the task or the content of the learning (Kennedy et al., 2013).

The learning content for the interaction was devised with the help of primary school teachers from a different school to the one where the study took place. The aim was to select a topic with which children had no prior exposure, but could be learnt in a relatively short time. Prime numbers were determined to be an ideal solution. Calculation of whether a number is prime can be performed by using division (for more detail see Section 6.2.2). Children of the age used in the study are familiar with division, but have not been taught what a prime number is at this stage of their education. The study was conducted at the end of an academic year. The children would typically learn about prime numbers in the following academic year, which would only be 3 months later. As such, the concepts involved, and difficulty of the task, are suitable for children at this developmental stage. The experimental context is also more relevant to the school curriculum, which

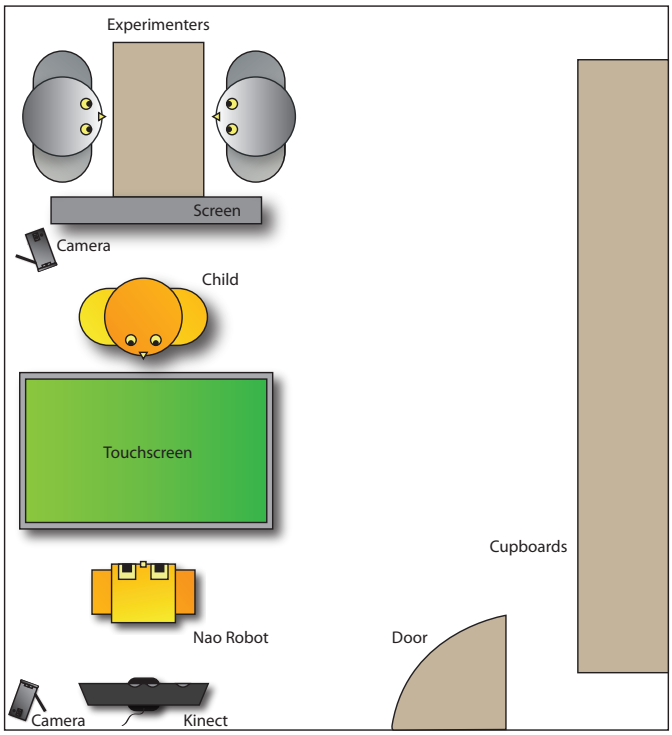

Figure 6.1: Schematic overview of the interactions under investigation in this paper. Two interactants (the child and the robot) face one another over the touchscreen. Two video cameras record the interactants during the studies. A Microsoft Kinect tracks the child's face. Two experimenters are in the room, but out of view of the child. Figure not to scale.

offers the findings greater veracity.

The touchscreen presents different numbers for sorting. The child can touch the numbers to drag and drop them into categories. An example library here would display text labels at opposite sides of the screen, such as 'prime' and 'not prime', with some numbers in between (Figure 6.2). The child can touch these numbers and drag them to the label for categorisation. The touchscreen sends all state information to the robot so that the robot knows the child's moves, and the robot can make moves itself by synchronising movement with on-screen animation (the robot does not physically touch the screen; Baxter et al., 2012).

Task Structure

The structure for the task was created partly through necessity for measuring learning and partly through a logical method of calculating primes known as the Sieve of Eratosthenes (O'Neill, 2009). The Sieve of Eratosthenes works through a group of numbers, eliminating non-prime numbers in a methodical manner to leave only the prime numbers. For the number range used in this study all composites can be eliminated by dividing by 2, 3, 5

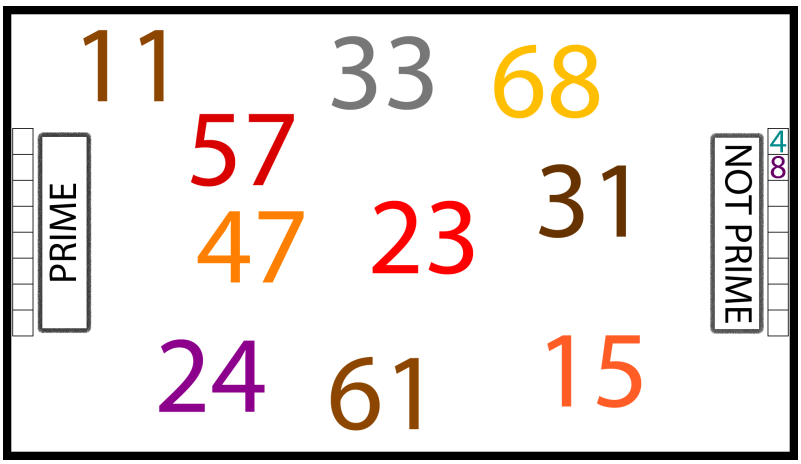

Figure 6.2: Example of the sorting task used. This is a screenshot of one of the tests used in the experiment. Children can touch a number, drag it over the 'prime' or 'not prime' label and release to make a categorisation. The number will then shrink and move into the boxes beside the category label.

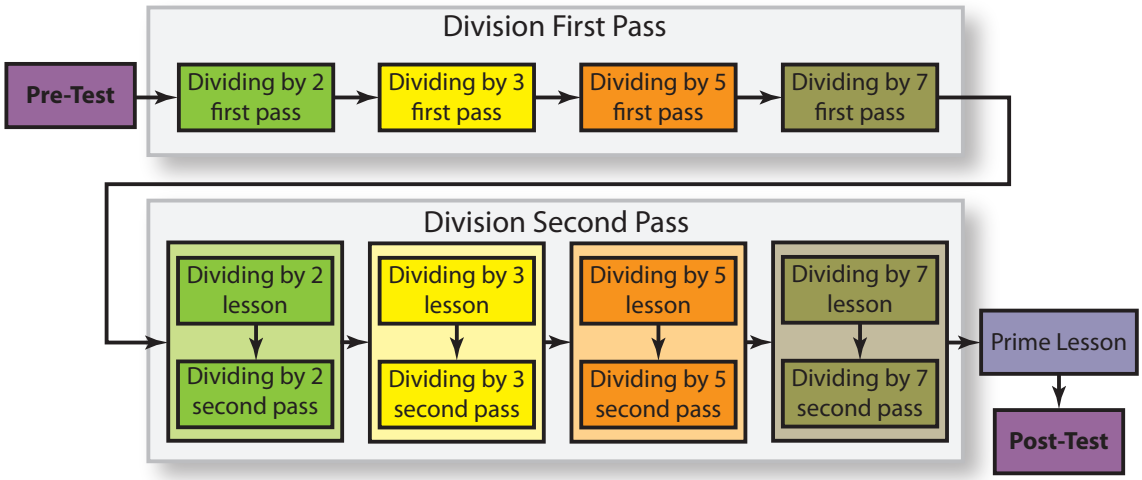

Figure 6.3: Structure of the task used in the interactions, showing robot lesson positions.

and then 7.

The task was structured so that appropriate measures could be taken for both prime number learning and division learning. Additionally, the task structure allows the examination of the children's division skills prior to the prime number post-test, which is important as these division skills are necessary for the calculation of primes (Figure 6.3).

Pre- and post-tests each consisted of 12 numbers being presented on screen; 6 were prime and 6 were non-prime (Figure 6.2 shows an example test). Both tests avoided numbers from the prime lesson and had balanced distributions of numbers across the range being used (10-70).

Each practice library in the division 'pass one' consisted of 8 examples - 4 of which could

be divided with no integer remainder by the number in question, and 4 which could not. This first pass was used to obtain a measurement for each child in how well they could divide by each of the divisors required for the main goal of calculating prime numbers. The number of examples in division 'pass two' totalled 24, but the distribution between each of the 4 divisors (2, 3, 5 and 7) was dependent upon the condition and performance in pass one (see Section 6.2.4).

Lesson Content

In the second division pass, a lesson was provided for each of the divisors: 2, 3, 5 and 7. This involved verbal instructions and categorisations on-screen. Each lesson consisted of a short verbal overview of the technique, followed by categorisation of 2 examples with verbal narration explaining the application of the technique to the examples. One example could be divided with no remainder, and the other could not. The lessons were not to teach the concept of the division, but often to provide a 'trick' whereby the division could be accomplished more easily. The lessons were explanations of the following concepts:

- Divisible by 2 - the number is even (ends in 0, 2, 4, 6 or 8)
- Divisible by 3 - sum the digits of the number and test if that divides by 3
- Divisible by 5 - the number ends in 0 or 5
- Divisible by 7 - no trick available; a reminder that a number in the 7 times table will be divisible by 7

The lesson about primes which took place after the second division pass used the information from the earlier division lessons to draw together the practice the child had with dividing by 2, 3, 5 and 7 into calculating whether numbers were prime. The concept of primes was explained (a number divisible, with no remainder, by only 1 and itself) before two worked examples were completed on-screen - one prime and one not prime. The Sieve of Eratosthenes was adapted to eliminate numbers one-by-one for categorisation. Children were instructed to consider each number to be categorised in turn, attempting to divide it by 2, 3, 5 and 7. If the number divided by any of these then it was not prime, otherwise it was prime.

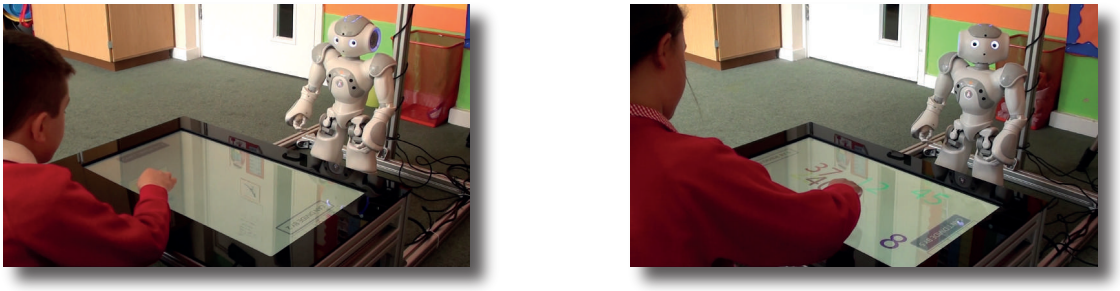

Figure 6.4: Snapshots taken from the video recordings of interactions. Both the social (*left*, looking at the child) and asocial robot (*right*, actively avoiding the gaze of the child) conditions are pictured to show the difference in gaze behaviour between them.

6.2.3 Conditions

In order to address the hypotheses, four conditions were devised. The 'division only' condition is used to provide the manipulation check; it is useful to verify that the lessons provided do indeed facilitate learning, with no robot present. The 'asocial non-personalised robot' condition described below provides an overlap between the embodiment and social behaviour research questions. The content is identical to the 'screen only' condition, thereby providing the comparison for physical embodiment. The 'social personalised robot' condition is then used as a contrast in terms of social behaviour.

1. **Division only** [n=11] - division pass one, followed by division pass two without any lessons. Conducted on the touchscreen only, with no robot present.

2. **Screen only** [n=11] - the full interaction as described in Section 6.2.2, but with no robot present. All feedback and lesson content is delivered by the speakers in the screen.

3. **Asocial non-personalised robot** [n=11] - identical script to the 'screen only' condition, but with the robot delivering the content. All verbal content and feedback is given by the robot; the screen now only displays the numbers for the task. Robot behaviour is designed to be non-social (see Section 6.2.4 for full details).

4. **Social personalised robot** [n=12] - a social version of the full interaction. All lesson content is kept the same as the asocial robot condition, but the non-lesson speech is adjusted to be more social. Robot non-verbal behaviour is also designed to be social.

6.2.4 Robot Behaviour

Human tutors are known to be effective, using social behaviour and adapting to the learning needs of the child. As such, the social robot behaviour was based on a human tutor's behaviour when taking five children through the task on the touchscreen. Section 6.2.4 outlines four observed behavioural dimensions that were implemented on the robot. Whilst maintaining balance between the conditions, the inverse for each dimension is used for the asocial robot behaviour in order to evaluate Hypothesis 2.

The phrases and actions used by the human were observed and implemented in the social robot model. It is posited that behaviour is perceived by the child as an integration of cues (Zaki, 2013), meaning that each dimension must be considered in context of the others. Consequently, personalisation and social behaviour are considered inseparable in assessing Hypothesis 2 for this study, following the human model.

Condition Independent Behaviour

Both robot conditions adopted the following basic behaviour during the image categorisation portions of the task:

Move Suggestions - During each stage of the interaction, if the child was hesitant in making moves then the robot would move a number to the centre of the screen and suggest that the child work on that number next. The decision about when to move was probabilistic and cued by the child's behaviour. If the child did not make a categorisation for 6 seconds, then there was a 25% chance that the robot would move, with the decision repeated every 2 seconds until a move was made - the 6 second timer would then start again.

Categorisation Feedback - The robot would provide verbal feedback on the child's categorisations. Not every categorisation received feedback; there was a 25% chance of feedback on each categorisation - following the human tutor model.

Robot Condition Differences

Verbal Content - The script for the social robot speech was taken from the human tutor; this was then modified for the asocial robot by removing any personalisation, i.e., "Johnny, we'll do dividing by 2 next" becomes "You'll do dividing by 2 next". The total amount of speech was kept as close as possible between the conditions, and the lesson content was

the same.

When providing speech alongside a suggestion, or when providing feedback, a number of phrases were available and selected at random. The asocial robot had only 2 options for each event (compared to the social robot's 8), thereby making it very repetitive.

Gestures - The social robot script used for the introduction and some of the lessons included iconic gestures. In the asocial condition, these were placed at inappropriate times, for example, the robot would wave its arm to greet the child half way through a sentence, rather than when it says hello at the start. The same gestures were used in both conditions, the only difference was their position in the script.

Personalisation - The social robot would use the child's name in greeting, just before the post-test and in the goodbye script. The asocial robot would not use the child's name at all. Personalisation of learning content was also provided by the social robot.

The performance of the child in the first division pass would dictate how many examples of each division library they would do in the second pass. A total of 24 numbers were always used in the second division pass. For the asocial condition, these were split equally between divisors, so 6 numbers for each of dividing by 2, 3, 5 and 7. In the social condition a minimum of 3 numbers were used per divisor, but the remaining 12 numbers were distributed between the divisors based on how many of each divisor the child got wrong in the first pass. Therefore, they had more practice on numbers that they were weaker at in the second pass.

In the second division pass, for each divisor library, there was also a reminder of the lesson available. In the asocial condition, this reminder would be delivered by the robot half way through the categorisations for that library (i.e., after the 3rd of the 6 categorisations to be made). In the social condition, the reminder was given after the first incorrectly categorised image.

Gaze - The social robot gaze was constrained so that it would generally be looking towards the touchscreen or in the direction of the child. Additionally, a Microsoft Kinect was used for tracking the child's head pose. If the child's head pose was directed towards the robot, then the robot would respond by looking back at the child. In the asocial condition, the robot was intentionally programmed to look up and to the side so that the gaze would avoid the child (Figure 6.4).

6.2.5 Procedure

One of the experimenters shown in Figure 6.1 controlled the start and end of the autonomous behaviour. This individual had three responsibilities: *1)* to type in the name of the child for the social robot condition before the child arrived in the room, *2)* to click a button once the child was sat down in front of the robot to denote the start of the interaction, and *3)* to click an 'emergency' button if anything went wrong, where the robot would gracefully end the interaction. All other robot behaviour was fully autonomous.

6.3 Results

This section will present the results from each of the conditions in relation to the hypotheses. Learning will be considered either between the pre-and post-test improvements, or for division, between the total percent correct in division pass one and division pass two. The behavioural analysis is derived from video coding of the child's gaze as previous work has highlighted gaze as the primary behaviour of interest in interactions of this nature (Kennedy et al., 2015b). The video coding was completed by one coder for all videos. Coding was verified by second-coding 20% of the videos, as in Moshkina et al. (2014), with an average Cohen's Kappa of 0.80 signifying substantial agreement (Landis & Koch, 1977).

The conditions were split to have an equal balance of ability based on an estimate by the children's teacher (higher, middle and lower tiers). Comparing the approximate ability level of the children that was provided by their teacher against their performance in the first division pass (at which point they've had no lesson input), Pearson's r correlation is 0.638. This is a good correlation, which confirms that the teacher's estimate is reflected in the results of this study and therefore that the conditions are balanced for ability.

The mean average length of the interactions were: 974s (95% CI [750s,1199s]) in the asocial robot condition, 1011s (95% CI [786s,1236s]) in the social robot condition, and 873s (95% CI [680s,1066s]) in the screen only condition. The average length of the division only condition (M=452s, 95% CI [277s,629s]) was much shorter as the robot lessons, pre-test and post-test add a lot of time.

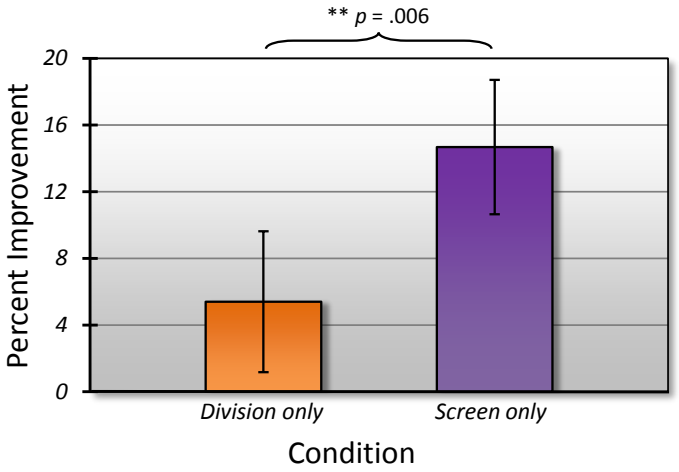

Figure 6.5: Improvement between division pass one and division pass two in percent for the division only and screen only conditions. Significantly greater improvement occurred in the 'screen only' condition (where division lessons were present) when compared to the 'division only' condition (where division lessons were not present), indicating that the lessons have a significant, positive impact on child division. *Error bars* show 95% Confidence Interval, ** indicates significance at the p<.01 level.

6.3.1 Learning from Lessons

A 2 tailed, unpaired t-test was conducted to compare the improvement between division pass one and division pass two in the division only (no lessons) and screen only (with lessons) conditions (with no robot present). The groups were used with independence of observations, and a continuous measure. Distributions did not significantly deviate from normality (Kolmogorov-Smirnov test; $p > .05$) and had homogeneity of variances (Levene's test; $p > .05$). There was a significant difference in the scores for division only (M=5.40, 95% CI [1.17,9.63]) and screen only (M=14.68, 95% CI [10.65,18.71]) conditions; $t(20)$=3.114, p=.006 (Figure 6.5). This manipulation check shows that the improvement was significantly higher when the lessons were present. The result here is not surprising, but it is beneficial to show the effectiveness of the division lessons.

6.3.2 Robot Presence

To examine how the robot affects the learning of the child, the improvement from pre-test to post-test scores was compared between the screen only and (combined) robot conditions. All pre-test and post-test scores are out of 12. None of the children who took part in the study reported to know what a prime number was before the interaction. As a result, based on 2 options for each categorisation, it would be expected that the pre-test scores would be around chance (50% out of 12 correct). Pre- and post-tests are paired observations, on a

continuous measure, with distributions that did not significantly deviate from normality (Kolmogorov-Smirnov test; $p > .05$), so paired samples *t*-tests are used for their analysis. Unpaired *t*-tests are used to compare between robot conditions as the observations are then independent.

In the screen only condition a 2 tailed, paired *t*-test reveals no significant difference between the scores for the pre-test (M=5.91, 95% CI [4.68,7.13]) and post-test (M=7.36, 95% CI [5.49,9.24]); $t(10)$=1.027, p=.329. However, when a robot is present there is a significant difference in scores between the pre-test (M=6.04, 95% CI [5.15,6.94]) and the post-test (M=7.78, 95% CI [6.61,8.95]); $t(22)$=2.997, p=.007. To further explore this result, the screen only pre-test and post-test scores were compared with those in the asocial robot condition. These two conditions are identical in the script that is used (the screen plays recorded clips of the robot voice) and the lack of personalisation. For the asocial robot, the difference is significant when the same test is run between scores for the pre-test (M=6.27, 95% CI [5.00,7.54]) and post-test (M=8.45, 95% CI [6.84,10.07]); $t(10)$=2.597, p=.027. However, it should be noted that there was no significant difference in the improvement between the screen only (M=1.46, 95% CI [-1.32,4.23]) and asocial robot (M=2.18, 95% CI [0.54,3.83]) conditions; $t(20)$=0.442, p=.664, when using a 2 tailed, unpaired *t*-test. Despite the asocial robot leading to significant learning gains and the screen only condition not, the lack of significant difference between the conditions means that there is no conclusive statistical evidence for learning improvement due to the robot. This means that Hypothesis 1 is not supported: the learning gains were indeed greater when a physical robot was present, but this difference was not statistically significant.

6.3.3 Social Condition

As shown in the previous section, the learning gains for the asocial robot were significant. When conducting a 2 tailed, paired *t*-test for the social robot condition there is no significant difference between the pre-test (M=5.83, 95% CI [4.54,7.13]) and post-test (M=7.17, 95% CI [5.50,8.84]); $t(11)$=1.627, p=.132. Whilst all conditions show improvement between the pre-test and the post-test, the only condition where the learning gain is significant is with the asocial robot; both the social robot and screen only conditions show non-significant improvement (Figure 6.6). This result contradicts Hypothesis 2, that a more social robot will result in greater learning gains than a less social robot.

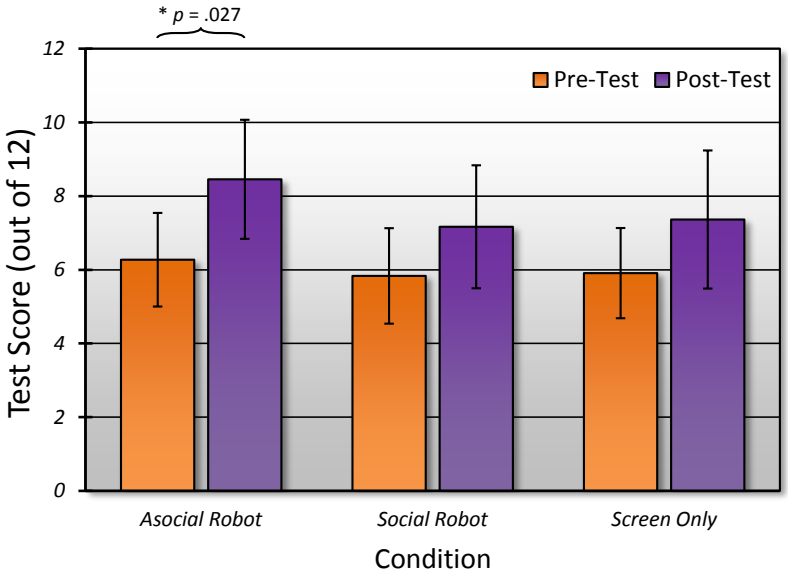

Figure 6.6: Pre-test and post-test scores for the asocial robot, social robot and screen only conditions. *Error bars* show 95% Confidence Interval, * indicates significance at the $p<.05$ level.

To explore the impact that the learning personalisation may have had on the results, the lesson reminders and practice of numbers in the second division pass are considered. In the asocial condition a reminder of the lesson is given for each divisor, whereas in the social condition, reminders are only given when the child makes a mistake. This meant that in the asocial condition a total of 44 reminders were given (M=4.00 per interaction; no deviation), whereas in the social condition a total of 22 reminders are delivered (M=1.83, 95% CI [0.94,2.73] per interaction). This is not surprising, as most children can comfortably divide by 2 and 5 at this age; thereby eliminating the need for around half of the reminders. When correlating the number of reminders provided to children in the social robot condition with their improvement between pre- and post-test score, Pearson's correlation $r(10)$=-0.418, p=.176. This is a moderate negative correlation, which suggests that receiving fewer reminders does not reduce the child's performance.

Additionally, children in the social robot condition were given the opportunity to practice more of the numbers that they were weaker at (following the human model described in Section 6.2.4). There is a possibility that this could have been a de-motivator if they then performed poorly in this phase of the interaction. However, this seems unlikely as it is found that there is no significant difference between the performance in the second division pass between children in the social condition (M=82% correct, 95% CI [71%,93%]) and those in the asocial condition (M=85% correct, 95% CI [75%,95%]); $t(21)$=0.432, p=.670.

In order to investigate the reasons behind why children's learning gains are not as great

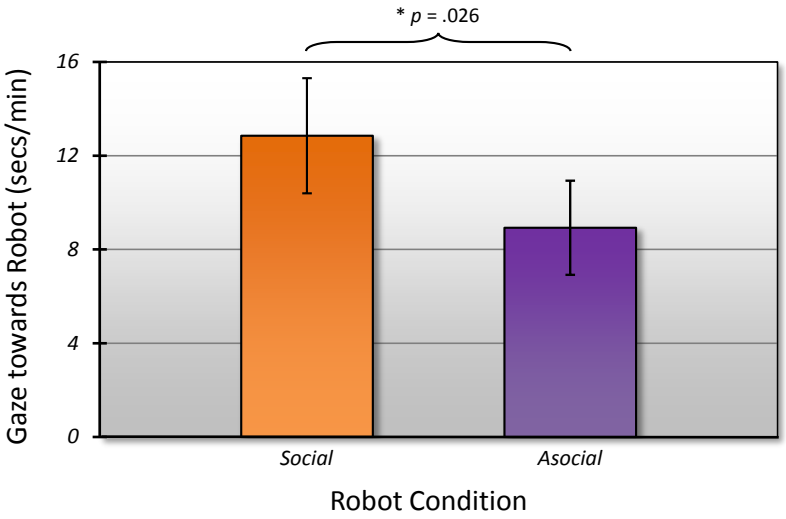

Figure 6.7: Child gaze towards the robot in seconds per minute, split by robot condition. *Error bars* show 95% Confidence Interval, * indicates significance at the $p<.05$ level.

when the robot is social compared to when it is asocial, the children's behaviour and self-reported view of the robot were analysed. From video coding of the interactions, it was found that children look significantly more often at the social robot (M=12.9, 95% CI [10.4,15.3]) than at the asocial robot (M=8.9, 95% CI [6.9,10.9]); $t(18)=2.425$, $p=.026$ (Figure 6.7). Values are provided in seconds of gaze at the robot per minute of interaction; this normalisation allows for comparison across interactions of different lengths.

The children completed a pre-questionnaire and a post-questionnaire before and after the interaction. These were very short, with just 4 questions in the pre-questionnaire and 2 questions for the post-questionnaire. The questionnaires were used to see what the children expected from the interactions, and subsequently how they viewed the robot afterwards. Despite being told by the experimenters several times before their interactions that they would be *taught* by a robot *teacher*, with the robot script emphasising this point too, the children in the social robot condition consistently reported that they thought the robot was a 'friend' after the interaction. The question asked "For me, I think the robot was like a -" with 8 options available: brother or sister, classmate, stranger, relative (e.g., cousin or aunt), friend, parent, teacher, and neighbour.

It was expected that the children would report the robot to be a teacher (as this is what they had been told), so their responses were grouped into either 'teacher' or 'not teacher'. In the social condition, 17% of the children reported the robot to have been like a teacher, compared to 64% in the asocial condition. As the data consists of two groups of categorical data, Fisher's exact test is used to show that the responses differ significantly by condition,

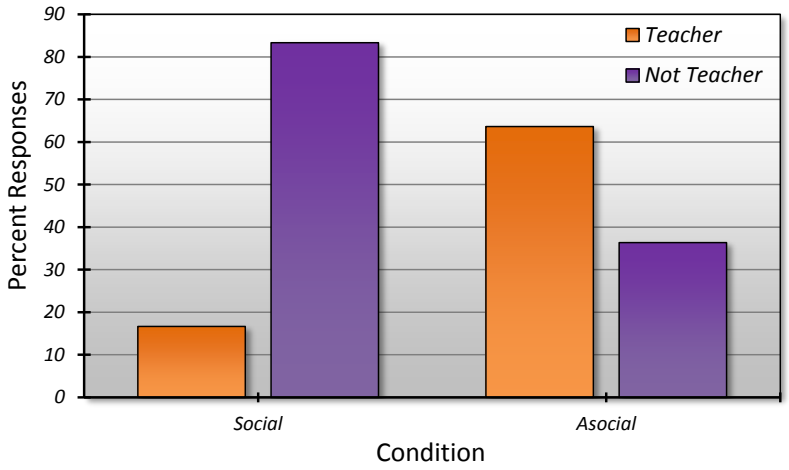

Figure 6.8: Post-questionnaire responses of the children when asked what they thought the robot was like. Eight options, including 'teacher' were available.

p=.036 (Figure 6.8).

It is clear from the children's gaze and self-reported responses that the difference in robot behaviour between conditions has an effect on the children's behaviour and attitudes towards the robot. It is suggested that these differences could account for the difference in learning gains observed between the social and asocial robot conditions. Whilst the robot is providing the lesson about prime numbers, it demonstrates two examples on the screen by highlighting the numbers, discussing them and correctly categorising them. Therefore, during this period it is useful to look at the screen. During the prime lesson the average amount of gaze towards the social robot (M=26.9 secs/min, 95% CI [22.9,30.9]) is significantly higher than the gaze towards the asocial robot (M=17.0 secs/min, 95% CI [11.0,23.1]); $t(18)$=2.669, p=.016 (unpaired t-test). It is suggested that the additional attention directed towards the social robot's behaviour could distract the children from the content that it is delivering; this possibility is further discussed below (Section 6.4).

6.4 Discussion

From the analysis of the results it is clear that the lessons for division have a positive effect on the children's performance. This validates part of the teaching behaviour and demonstrates that the children have the ability to understand the robot's voice and apply knowledge gained from the lessons in the task on-screen.

When the asocial robot is present, despite having the same content as the screen only condition, the improvement between pre-test and post-test is significant. This is a demon-

stration of the social presence effect; the addition of an agent into the interaction leads to improvement in task performance (although non-significant in this case), as observed before in other contexts, for example Kose-Bagci et al. (2009) and Leyzberg et al. (2012). However, the improvement is lost when the robot behaviour is changed to become more social. This is a surprising result, which contradicts both Hypotheses 1 (that a robot will provide greater learning gains than the screen alone) and 2 (that a more social robot will result in improved learning gains).

This result is in contrast to existing studies in the literature that Hypothesis 2 was based on. As described in Section 6.2.4, the robot behaviour was derived directly from that of a human tutor. This necessitates a perspective that integrates behavioural dimensions (Zaki, 2013) that emphasises sets of behavioural competencies (similar to the use by Saerbeck et al. 2010). This differs from the more typical focus on individual social cues, as in Mayer et al. (2004) and Szafir & Mutlu (2012). With the interaction context (child-robot interactions in a school) and task content (learning mathematical concepts) also differentiating the work here from previous studies, this integrated cues perspective may merit further investigation in terms of the effects on the perceptions and performance of human interactants.

One factor that should be considered in the analysis of the results with regards to embodiment is the behaviour employed by the 'screen only' condition. The 'asocial robot' condition was used as the overlap between the embodiment and social behaviour research questions in this work. As such, the screen only condition used the same script as the asocial robot condition. This was done in part due to the personalisation aspects of the social robot condition; it was considered that the screen telling the child its name may be confusing given the lack of a visible character. The asocial robot condition did not contain these personalisation aspects, so was deemed a more appropriate choice. As a result, a comparison only between the screen only and asocial robot conditions can be made in terms of embodiment; a comparison with the social robot condition incurs a confound of social behaviour. Ideally, a 2x2 between-subjects design would have been used (i.e., through the addition of a 'social screen' condition), but subject numbers were limited by class sizes, and a more thorough manipulation check would be required to understand how children perceived the social behaviour from the screen (particularly in light of the personalisation aspect). It could be the case that when the social skills variable is included in the design and the behaviour from the screen changes, different results

would be observed.

One possible explanation for the unexpected findings with respect to learning is that although the children looked at the social robot significantly more than the asocial robot during the lesson phase (which could be considered advantageous as the robot provides the lessons), they were paying attention to the social behaviour instead of the lesson content. An alternate explanation is that the social behaviour presented by the social robot places more cognitive load on the children, which may inhibit their capacity to process information related to the task (Sweller, 1994). It may be that in the long-term, as the novelty of the social behaviour wears off, the social robot would then elicit better learning, as indicated by Kanda et al. (2004). However, further research is required to explore these ideas explicitly and in more detail.

6.4.1 Child Perception and Ability

In Section 6.3.3 it was shown that children in the asocial robot condition were more likely to report that they viewed the robot as a teacher than those in the social robot condition. The infrequency with which those in the social condition reported the robot to be like a teacher was surprising. The children were told several times before and during the interaction by both the robot and the experimenters that the robot was a teacher. It is suggested that there may be two reasons as to why this was the case. Firstly, it may be that the directness of the asocial robot conformed more to their expectations of what a robot teacher would be like than the social robot, which was less direct in its instructions. Secondly, the behaviour of the asocial robot may not have had enough character to change the children's perception of the robot as a teacher, whereas the social robot did. Unfortunately, no holistic measure was used to assess how the manipulation of the robot's social behaviour was perceived by the children. As such, it is not possible to evaluate whether the children saw either condition as more or less social. This limits the findings here as it is not possible to verify whether the differences in the child perceptions were directly linked to the social behaviour manipulation. It would additionally have been useful to verify that the children detected the differences between the robot social behaviour conditions. This is an aspect of the study design that is rectified in subsequent chapters. Interestingly, there was almost no correlation between the children's perception of the robot as a teacher and their performance; Pearson's $r(20)$=-0.11, p=.626.

There was only a weak correlation (Pearson's $r(20)=0.13$, $p=.558$) between the teacher-provided mathematics ability levels of the children and their subsequent improvement between pre-test and post-test. This is somewhat surprising, as one would expect the higher ability students to progress more given the same practice as those who were lower ability. This may highlight a limitation in the adaptiveness of the robot's behaviour used in this study. It is possible that a robot which is more adaptive could better respond to each individuals' needs and push them more effectively through the Zone of Proximal Development (Vygotsky, 1980).

Due to the relatively small sample sizes used here, it only requires 2 or 3 subjects to perform particularly poorly or well to impact on the significance of the results. However, there is a trade-off between trying to carefully control the experiment and get greater subject numbers. Subjects were selected from the same school and year group so that they would have similar educational experiences and backgrounds. Due to limits on the sizes of school classes, it is likely that to get greater numbers would mean selecting subjects across multiple schools. This then introduces the risk of large variability between subjects' mathematical ability and the environment in which the experiment is conducted.

6.4.2 Gender Differences

One interesting aside that was noticed through additional exploratory analysis are differences between the genders. These results were not included in Section 6.3 as they were not part of the original hypotheses for this study. However, as an interesting observation they have been included here, with the suggestion that they may be worth further research. A significant difference is found between the improvement between pre-test and post-test of girls ($M=2.77$, 95% CI [1.18,4.36]) and boys ($M=0.40$, 95% CI [-0.85,1.65]) when interacting with a robot present (both social and asocial conditions combined); $t(21)=2.192$, $p=.040$. These results show that the boys barely improved with a robot, whilst the girls improved quite substantially.

Additionally, girls who interacted with a robot present ($M=2.77$, 95% CI [1.18,4.36]) improved more than those without a robot present ($M=-0.40$, 95% CI [-3.71,2.91]). Whilst this difference is not quite significant ($t(16)=1.907$, $p=.075$), it seems as though there may be a possible trend. Gender differences due to social presence have been observed in other contexts in HRI, such as Schermerhorn et al. (2008), where females saw a robot as more

machine-like. This could support the argument that the robot social behaviour distracts from the lesson content that it is delivering; girls, who may perceive the robot as less social, therefore outperformed the boys. Whilst there is not enough evidence here to make firm conclusions about this point, the effect of gender possibly merits more research in the context of educational interactions.

6.4.3 Nonverbal Immediacy

The adult crowdsourced nonverbal immediacy ratings of the robot behaviour provide further insight as to why the anticipated learning differences were not found. 30 ratings of the 'asocial' robot, and 33 of the 'social' were acquired (for demographic details please see Appendix B). The asocial robot received an average nonverbal immediacy score of M=48.5 (95% CI [46.1,50.8]), with the social M=49.0 (95% CI [47.6,50.4]). These scores are not found to be significantly different when using an independent samples, two-tailed t-test; $t(61)$=0.372, p=.711. At first glance, this is a surprising lack of difference, however, this is likely due to the way in which the nonverbal immediacy questionnaire quantifies social behaviour. It does not take into account all aspects of interaction (as discussed in Chapter 4), but just the quantity of overt nonverbal cues used. As the manipulation in this study was largely concerned with timing and verbal phrasing differences, these would not be picked up in the nonverbal immediacy measure.

Section 6.4.1 showed that there were clear differences in the perception of the robot by children, but these differences are not reflected in the nonverbal immediacy scores as rated by adults. Nevertheless, the correlation between nonverbal immediacy and cognitive learning gains would hypothesise no learning differences between the conditions here, given the near-identical nonverbal immediacy scores. To further explore this relationship in interaction-based HRI scenarios it would be useful to intentionally create and compare robot conditions with more contrasting nonverbal immediacy behaviours.

6.5 Summary

As expected, the use of lessons improved the children's performance between the first division pass and the second division pass, as shown in Section 6.3.1. Partial evidence was found in support of the social presence effect. Section 6.3.2 showed that when a

robot delivered the lessons to the child, the learning was significant, whereas when the same information was provided by just a screen, without a robot, it was not, but the difference between the conditions was not statistically significant. By further breaking down the robot results into the two different behavioural conditions, it was found that the learning remains significant with the asocial robot, where the script is identical to the condition without the robot present (where the learning was not significant). However, these positive effects were not maintained when the robot was more social.

The results here have shown that a robot which is not appropriately social led to greater learning gains of children in a maths task than a robot with appropriate social behaviours. This result contradicts expectations and predications made based on other studies in the literature (for example Mayer et al., 2004 and Saerbeck et al., 2010). It is hypothesised that the social behaviour of the socially appropriate robot may distract from the content it is delivering with regards to the learning task, whilst the asocial robot leads to disinterest, and therefore less distraction from the learning task. Gaze behaviour of the children throughout the interaction and specifically during the prime numbers lesson is used to provide evidence for this suggestion.

Chapter 7

Robot Nonverbal Immediacy and Child Learning

Key points:

- An experiment is devised to test the effects of high and low nonverbal immediacy behaviours when applied to robot social behaviour.
- The methodology builds on the one used in the last chapter: teaching children how to identify prime numbers. The robot behaviour in this chapter is derived from principles of nonverbal immediacy, rather than from a human model. This allows for a stronger theoretical underpinning in terms of what constitutes social behaviour, and provides a directly measurable manipulation of robot behaviour.
- Children learn a significant amount from a robot perceived by the children as having higher nonverbal immediacy, but do not undergo significant learning with a lower nonverbal immediacy robot.
- These findings confirm predictions from the human-human immediacy literature and provide a basis for robot social behaviour implementation in educational interactions.

Part of the work presented in this chapter has been published in Kennedy et al. (2015e). The final publication is available from Springer via:
http://dx.doi.org/10.1007/978-3-319-25554-5_33

Chapter 5 demonstrated the social advantages that using a physically present social robot can bring to interactions with children. These findings were furthered in Chapter 6 where children learnt a significant amount from a social robot and touchscreen, but did not experience such learning gains when provided with the same information on a touchscreen alone. However, Chapter 6 also revealed surprising learning results in response to different robot social behaviours. These behaviours were based on a human model completing the same task, and an 'inverse' set devised by the author. These behaviours were intended to provide maximal differences between conditions to demonstrate the positive effect of appropriate social behaviour. However, the results were not as expected, with the inverse set of behaviours (expected to be inappropriate for inducing learning) instead leading to significant learning.

This chapter seeks to further explore the impact of robot social behaviour on child learning. The lesson content from the previous chapter is maintained, but the robot behaviour used here is different. The aim is to remove any confounds relating to the personalisation contained in the behaviour of the previous chapter (these personalisations were part of the human model and were not distinctly separable from the general social behaviour). In order to do this, the social behaviour derived from the human model is substituted for behaviour derived from the principles of nonverbal immediacy as described in the literature and specified in the nonverbal immediacy scale (Richmond & McCroskey, 1998).

7.1 Hypotheses

The HHI literature has shown that greater instructor NVI leads to increased cognitive learning gains (Witt et al., 2004). These findings have been partially confirmed in HRI (Szafir & Mutlu, 2012), but using only 2 modalities (speech and gesture). Nonetheless, survey data showed that participants could perceive such behavioural differences. Previous work in Chapter 6 conducted in a similar context to this study found that children gazed more at a 'more social' robot tutor during lessons, and were more likely to report it to be like a friend than an equally active, but counter-social, robot tutor. It could be argued that an increase in NVI behaviour is analogous to an increase in social skills, so the same perceptual and behavioural differences of children could be predicted here, leading to H2 and H3. Based on these prior findings, the following hypotheses were devised:

H1: A robot with higher nonverbal immediacy will lead to greater child cognitive learning gains.

H2: Children will regard a robot with high nonverbal immediacy more like a friend than one with low nonverbal immediacy.

H3: Children will gaze at a robot with high nonverbal immediacy more during the prime lesson period than at a robot with low nonverbal immediacy.

7.2 Experimental Setup

The study here uses 2 conditions in a between-subjects design, both using a robot, with the nonverbal immediacy manipulated to create contrasting robot social behaviours. The methodology used in this study is as established in prior studies and Chapter 3. A robot is used as a tutor in one-to-one interactions to teach children how to identify prime numbers between 10 and 100 based on whether they are divisible by 2, 3, 5 or 7. The children participating in the interactions do not have prior knowledge of prime numbers, but have the skills to do the division (albeit with imperfect performance), making the combination of these skills into a rule for categorising primes possible in a short interaction. Prior knowledge is assessed with a pre-test. The novel difference between the present study and previous work is in the robot behaviour. Previously the social behaviour was based on a human model, whereas in this study the robot behaviour is derived from NVI concepts (detailed in Chapter 2).

7.2.1 Participants

The study was conducted in a class of children aged 8-9 years old. All children interacted with the robot, but due to breaks in protocol, and one statistical outlier (Grubbs' test), several interactions were excluded from the analysis. A total of 23 interactions were considered (16F/7M, age M=8.74, SD=0.45). All subjects had permission from a parent/guardian to participate in the study. Of the 23 subjects, 21 also had permission to be filmed for video analysis.

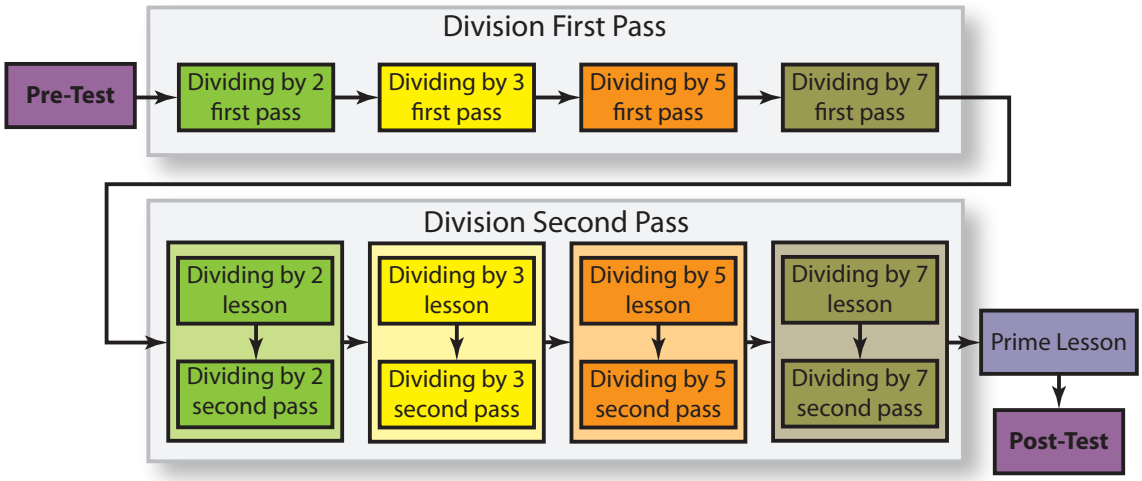

Figure 7.1: Structure of the task used in the interactions, showing robot lesson positions (identical to Chapter 6 Figure 6.3).

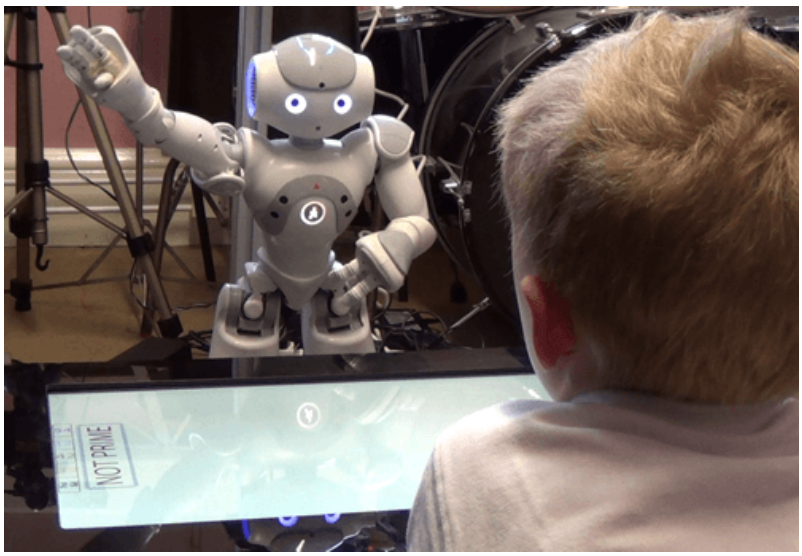

Figure 7.2: A snapshot from the 'high' nonverbal immediacy condition.

7.2.2 Interaction Protocol

Interactions took place in an empty room familiar to the children near to their classroom. The children were briefed by one of the experimenters before entering the room. Two experimenters were present in the room, out of view of the child whilst they interacted with the robot. The child sat across a large touchscreen from an Aldebaran NAO robot (Figures 7.2 and 7.3). A Microsoft Kinect was placed behind the robot to track the child's head gaze. Video cameras were placed behind the robot and behind the child to record the interaction. The average time spent interacting with the robot was M=14m19s, SD=3m27s. The average interaction time from the videos (from entering the experiment room, to exiting – therefore including questionnaire time) was M=19m19s, SD=3m43s.

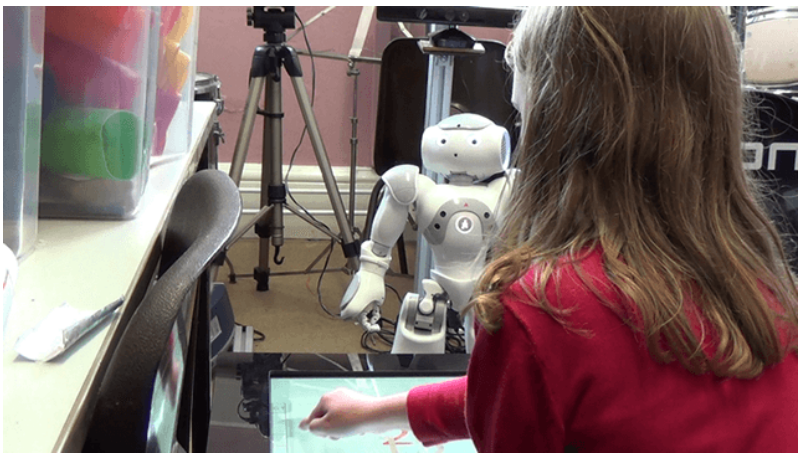

Figure 7.3: A snapshot from the 'low' nonverbal immediacy condition.

The robot would first introduce itself and ask the children to complete a pre-test on the touchscreen for prime numbers, and then pre-tests for each of the divisors (2, 3, 5 and 7). The robot would then deliver a lesson for each of the divisors and ask the child to complete a post-test following this lesson, i.e., the robot gives a lesson on dividing by 2 and then the child does a dividing by 2 post-test, followed by dividing by 3 lesson and post-test, and so on. After this had been completed, the robot delivered a lesson about prime numbers which combined the lessons for the divisors into a rule for determining whether a number between 10 and 100 is prime (a variation on the *Sieve of Eratosthenes* method). The interaction with the robot would finish with a prime number post-test.

The prime number pre- and post-tests both consist of 12 numbers which must be categorised as 'prime' or 'not prime'. Two sets of numbers were used for these tests, which are alternately used as the pre- and post-tests in a cross testing strategy to control for potential difference in test difficulty. The tests were balanced in terms of the number size, as it was assumed that higher numbers would be harder for the children to work with. The divisor pre-tests consist of 8 numbers which must be categorised as either 'can divide by X', or 'can't divide by X' (where X is 2, 3, 5 or 7, Figure 7.1). The divisor post-tests are the same, but with 6 numbers instead of 8. In all pre- and post-tests, an equal quantity of numbers belong to each category.

After the interaction with the robot is finished, the child is asked by the experimenter to complete two questionnaires. The first questionnaire was a Robot Nonverbal Immediacy Quesionnaire (RNIQ), adapted from the short-form NVI questionnaire (Richmond & McCroskey, 1998), described in Chapter 3 and available in Appendix E. The second questionnaire consisted of two multiple choice questions, asking the children what they

High Nonverbal Immediacy	Low Nonverbal Immediacy
Leans forwards	Leans backwards
Actively gazes at child (with frequent movement)	Looks up and away from child (with occasional movement)
Frequent gestures while talking	No gestures while talking
Standard TTS	TTS modified to make voice "dull"
Continuous small upper body movements (relaxed upper body)	Rigid/tense upper body with no movement

Table 7.1: Robot behaviour for high and low nonverbal immediacy (NVI) conditions.

thought the robot was like (8 options including friend and teacher), and what they thought playing with the robot was like (4 options, plus a free text box) – available in Appendix I.

7.2.3 Robot Behaviour

The robot social behaviour was generated by considering the NVI questionnaire measures, as seen in Richmond et al. (2003). The intention was to create high and low NVI conditions in order to address the hypotheses for the study (Section 7.1). Each of the modalities rated in the RNIQ was considered for the Aldebaran NAO robot. Some of the modalities are not possible to manipulate (for example the NAO cannot perform facial expressions), but the other modalities were considered in turn and designed to be either maximally or minimally immediate. Table 7.1 shows the differences between the two robot conditions. All robot behaviour was autonomous, a 'Wizard-of-Oz' was only employed to click a button to begin the behaviour once the child was in position in front of the robot/screen.

Children were assigned to conditions randomly, whilst balancing for gender and mathematical ability (as judged by the class teacher). This led to 12 children in the low NVI condition (9F, 3M) and 11 children in the high NVI condition (7F, 4M) after exclusions.

7.3 Results

7.3.1 Learning Gains

To test the impact of the robot's lessons on the children's division skills, the percentage score of division across all pre-tests was compared with the score across all division post-tests. The two pre- and post-test groups were compared with independence of observati-

ons (as there were a different number of items in the pre- and post-tests), and a continuous measure. Distributions did not significantly deviate from normality (Kolmogorov-Smirnov test; $p > .05$) and had homogeneity of variances (Levene's test; $p > .05$). For this reason, a two-tailed independent samples *t*-test is used. A significant difference is found between the division pre-test percentage (M=0.84, 95% CI [0.80,0.88]) and the post-test percentage (M=0.89, 95% CI [0.85,0.92]); $t(22)$=2.081, p=.049. This demonstrates that the children can learn from the robot and suggests that the lessons that the robot delivers are appropriate.

All scores for the prime number pre- and post-tests are out of 12. Given that the children have no prior knowledge of prime numbers and there are 2 potential categories for each image, a pre-test score of 6 (50%) would be expected from random behaviour. Two groups (pre-test and post-test) are compared with paired values on a continuous measure, distributions did not significantly deviate from normality (Kolmogorov-Smirnov test; $p >$ .05). As such, a paired samples *t*-test is used. In the low NVI condition the improvement from pre-test (M=7.08, 95% CI [5.01,9.15]) to post-test (M=8.00, 95% CI [6.24,9.76]) is not statistically significant; $t(11)$=0.754, p=.466. However, in the high NVI condition the difference from pre-test (M=5.09, 95% CI [3.43,6.75]) to post-test (M=7.00, 95% CI [4.88,9.12]) is statistically significant at the p<.05 level; $t(10)$=3.057, p=.012 (Figure 7.4).

The pre-test score appears to be very different between the conditions, however this was not found to be significant; $t(21)$=1.640, p=.116. The 95% confidence interval for the pre-test in both conditions covers the expected value of 6, which reassures that the children did not know what primes were before the intervention. Additionally, there is no significant difference between the two different pre-test scores, or of the improvement between pre- and post-test, regardless of which of the two pre-tests were taken; this shows that the tests can be considered of equal difficulty. Therefore, support has been shown for Hypothesis 1: children interacting with the high NVI robot benefit from increased cognitive learning gains. However, this is slightly tempered, as there is no significant difference between conditions. Children in both conditions are likely to improve (which isn't surprising given practice and teaching input), but those in the high NVI condition undergo significant improvement, whereas those in the low immediacy condition do not.

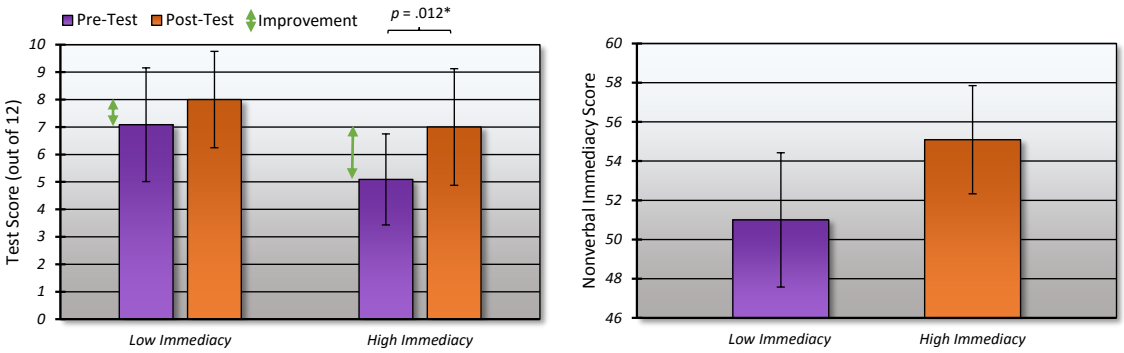

Figure 7.4: Pre- and post-test scores on recognising prime numbers for the low and high nonverbal immediacy (NVI) conditions (*left*); NVI scores for the designed low and high NVI conditions (*right*). Children improve more in recognising prime numbers when taught by a high immediacy robot. *Error bars* show 95% CI.

7.3.2 Questionnaire Data

After the children had interacted with the robot they were asked to complete the RNIQ on paper. Immediacy scores are calculated from the answers to the RNIQ questions: the higher the resulting number, the higher the perceived immediacy. The score can be up to 80, but there are a number of measures for which there are no equivalent robot behaviours (e.g., touching the child). Therefore, a score of around 56 would indicate a rating of near-maximal NVI given the modalities which are manipulated. This reduction in the expected score also inhibits the potential for difference between conditions, as for many of the questionnaire elements, the behaviour is the same (e.g., the lack of facial expressions).

The designed low immediacy condition received a mean NVI score of M=51.0 (95% CI [47.6,54.4]). The designed high immediacy condition received a mean score of M=55.1 (95% CI [52.3.57.9]). An unpaired t-test is used as these groups are independent observations. The test reveals that this difference falls just outside of significance at the p<.05 level; $t(21)$=2.031, p=.055 (Figure 7.4). However, the subject numbers are relatively low and if the trends seen were to continue, then this difference would become significant. Indeed, the differences between the means only just includes no difference; $-0.10 \leqslant \mu_{\mathrm{HNVI}} - \mu_{\mathrm{LNVI}} \leqslant 8.28$. This provides reasonable support for the manipulation check; children perceive a robot designed to be more nonverbally immediate as such.

The second questionnaire that the children completed asked them what the robot was like, and what playing with the robot was like. The children were asked "For me, I think the robot was like a -", and had 8 options to choose from (brother or sister, classmate, stranger, relative (e.g., cousin or aunt), friend, parent, teacher, neighbour). Given Hypothesis 2 (that

children in the high immediacy condition will more frequently report the robot to be like a friend) the responses were sorted into whether the children responded that the robot was like a friend, or not. In the high immediacy condition 6 children reported the robot to be like a friend and 5 not (with all selecting 'teacher'), whereas in the low immediacy condition 1 child reported the robot to be like a friend and 11 not (1 'classmate', 10 'teacher'). Fisher's exact test reveals a significant difference between the conditions, with those in the high immediacy condition significantly more likely to report that the robot is like a friend than those in the low immediacy condition; p=.027. Therefore Hypothesis 2 is supported.

This result is surprising as the children are told multiple times by both the experimenter and the robot that the robot is a *teacher robot* which will teach them some maths. However, the same result has previously been found, which led to the formulation of Hypothesis 2: children interacting with a 'more social' robot reported more frequently that the robot is like a friend (as in the previous chapter; published in Kennedy et al., 2015d).

7.3.3 Gaze Analysis

The 21 videos from the interactions were manually coded for child gaze during the prime lesson segment as this has previously been found to be indicative of overall gaze patterns in the interaction (Chapter 6, and the prime lesson constitutes a key part of the interaction in terms of the learning outcome. One of the 21 videos was excluded due to occlusions, leaving 9 videos from the high immediacy condition and 11 from the low. 20% of the remaining videos were second coded to verify validity, with a mean Cohen's Kappa of 0.83, indicating *almost perfect* agreement.

No significant difference was found between the length of time children gaze at the robot in seconds per minute of the prime lesson segment between the high NVI condition (M=15.9, 95% CI [11.3,20.5]) and the low NVI condition (M=15.4, 95% CI [11.9,18.6]); $t(18)$=0.214, p=.832. Nor is there a significant difference in the number of times children gaze at the robot per minute of the prime lesson segment between the high NVI condition (M=15.2, 95% CI [12.2,18.2]) and the low NVI condition (M=14.7, 95% CI [11.7,17.7]); $t(18)$=0.234, p=.817. Therefore, Hypothesis 3 (children will gaze more at the high NVI robot) is not supported. This is a surprising result, which possibly strengthens the link between robot behaviour and learning. If gaze is considered to be a reflection of child attention, then

despite equivalent attention during the key piece of learning input from the robot the learning results still vary, suggesting that the robot social behaviour could be responsible. Of course, this is just one of many possibilities and the gaze could be equal simply because the behaviour during this phase is quite novel compared to the rest of the interaction.

7.4 Discussion

The manipulation check revealed that children do recognise when a robot has higher or lower nonverbal immediacy. The difference was close to significance at the 5% level, with the difference between the means only just including no difference. This finding shows that the robot behaviour is largely interpreted by the children as intended by the designer, despite the children not seeing the other robot condition for comparison (which would likely further emphasise the difference between conditions). However, the variation in the children's answers is quite high, possibly due to a tendency to categorise at the extremes of scales (Borgers et al., 2004), misunderstanding of some negatively worded questions, or over-attribution of robot competencies.

The results also partially confirm Hypothesis 1; that a robot perceived as more nonverbally immediate will lead to greater cognitive learning gains. This prediction was made based on HHI data (Witt et al., 2004) and HRI data (Szafir & Mutlu, 2012), which seem to agree with the present findings. It should be noted that the effect size is relatively small: although there is significant improvement, the post-test mean 95% confidence interval still covers the expected 50% score of 6 which could be achieved through random action. Interestingly, there is a moderate positive correlation between immediacy score and cognitive learning gains for the high NVI condition ($r(9)=0.22$), which is remarkably close to that which is found in HHI literature ($r=0.17$; Witt et al., 2004). Conversely, there is a negative correlation for the low NVI condition and learning gains ($r(10)=-0.32$), indicating that as children rate the low NVI robot as more immediate, their learning tends to decrease. It should be noted that neither of these correlations are significant at the $p<.05$ level, but this is in part due to the low degrees of freedom for such a test.

It is therefore suggested that other factors besides robot behaviour could have a greater impact on the learning taking place at the individual level, particularly for those in the low NVI condition. From exploratory analysis of the data in this study, gender, teacher predicted maths ability, and age were all controlled for, with none being revealed *post*

hoc as a significant factor. Novelty is often raised as a potential issue when performing single interactions of this nature (Gockley et al., 2005; Kanda et al., 2004), and it could indeed be a factor here, with the novelty of the robot impacting some of the children more than others, although the influence of novelty could be expected to be similar in both conditions. Another possible factor could be in the character of the children themselves. Whilst the children are familiar with the environment, they are not familiar with the two experimenters in the room, which may impede their performance, or affect their questionnaire responses (Leite et al., 2012). Children who are more timid may be affected by this to a greater extent than those who are more confident.

Finally, it should be noted that the interactions in this study are relatively short, and the pre- and post-tests were conducted immediately before and after the learning input. Therefore, whether the learning gains observed here are retained over a longer period of time (and thus the concepts are truly *learned*) remains to be seen. This is an important factor which should be addressed in future work. Research from HHI, which has been conducted over the period of academic terms with adults, has found that high NVI behaviour can confer a greater advantage in terms of learning gains (Witt et al., 2004), although it must be noted that this is not always the case (Comstock et al., 1995). It could be hypothesised that over a longer time period with a robot tutor the differences between high and low immediacy conditions would increase as novelty would wear off and more of the potential benefit commonly gained in HHI from more immediate behaviour could occur.

7.4.1 Adult Immediacy Ratings

The crowdsourced adult nonverbal immediacy ratings reflect the ratings provided by the children in this study. The high NVI robot received an average nonverbal immediacy score of M=48.4 (95% CI [46.9,50.0]), with the low NVI M=40.2 (95% CI [38.1,42.2]). These ratings are found to be significantly different at the $p<.05$ level when conducting a two-tailed, independent samples t-test; $t(62)=6.173$, $p<.001$. The adult ratings have a smaller variance than the child ratings, and also have a larger difference between the conditions. The findings further support the differences between the conditions, and the suggestion that the children could perceive the difference in robot behaviour. As in Chapter 4, the comparison between child and adult immediacy ratings finds the child ratings higher and more tightly bunched, but with clear similarities in terms of relative ranking. The

adult scores also support the positive correlation observed in the HHI literature between perceptions of nonverbal immediacy and cognitive learning (Richmond & McCroskey, 1998).

7.5 Summary

This study has shown a strong trend towards children perceiving robots designed to have high and low nonverbal immediacy behaviours as such when measured using a short-form robot nonverbal immediacy questionnaire (RNIQ). These findings are in agreement with those from Chapter 4, but are now confirmed in a one-to-one interaction context, rather than a lecture-style environment. They are further supported by the crowdsourced adult nonverbal immediacy ratings. This perceived difference was also supported by the children's interpretation of the robot's relation to them, with significantly more children in the high nonverbal immediacy condition reporting the robot to be like a friend. There were no observable differences in gaze behaviour during the prime lesson period of the interaction, indicating that learning differences remain despite possibly equal amounts of attention being paid by the child to the robot during the lesson.

This chapter has shown that children who interact with the robot exhibiting more immediate nonverbal behaviour make significant cognitive learning gains, whereas those interacting with the less immediate robot do not. A strong trend is found in the difference between the conditions suggesting that the children can perceive the differences between conditions, which supports results with adults. While further work is required to assess the strength of the learning effects over longer time scales, and the effect of individual differences beyond academic competence, these results have demonstrated the utility of high nonverbal immediacy robot behaviours in a tutoring context.

Chapter 8

Robot Verbal Immediacy and Child Learning

Key points:

- High nonverbal immediacy behaviours applied to social robot tutors have been found to lead to significant child learning in the previous chapter. This chapter now considers the other part of the immediacy construct: verbal immediacy.
- A novel French language learning scenario is devised. Building on methodological approaches from previous chapters, a recall test is completed by the children the following week in addition to the pre-test and post-test as part of the interaction. A greater number of subjects are also considered in this experiment.
- It is found that children learn a significant amount in both high and low verbal immediacy conditions. However, despite being able to perceive differences in verbal immediacy, child learning is equivalent between conditions.
- Children retain their learning the following week and acquire vocabulary during the interaction, highlighting the promise of applying social robots in the context of language learning.

Part of the work presented in this chapter has been published in Kennedy et al. (2016b). The final publication is available from the IEEE via:
http://dx.doi.org/10.1109/HRI.2016.7451757

Previous chapters have explored the impact of nonverbal social behaviour and nonverbal immediacy on learning. Of course, this only covers part of the immediacy construct, with a range of verbal immediacy behaviours remaining unexplored in HRI. This chapter seeks to evaluate the impact of verbal social behaviour, as motivated by verbal immediacy, on child learning. An experiment is devised to test this, which builds upon the methodological approaches taken in previous chapters. The learning of the children is measured in the short-term (immediately after the interaction), and also the following week to check that the learned information is retained.

This chapter also introduces a novel language learning task. Child language learning provides an ideal domain for social HRI to contribute to. In the case of language, children learn better than adults, despite the increased cognitive capacity of adults. Language learning has a 'critical period' in neurobiology (Kuhl, 2010), which means that there is a window in which it is best learned. As such, evaluating with children aged 8 and 9 years old as in previous chapters is ideal. At this age, the children are still within the critical period, but have sufficient skill to read novel words without assistance. Language learning is an inherently social process (Kuhl, 2007a), which places it in a strong position to evaluate social behaviour. It is hoped that by making the task rely more heavily on social behaviour, the robot can have a greater influence on the outcome of the learning.

The background for this thesis provided in Chapter 2 introduced a variety of work conducted in the field of HRI with a focus on the impact of robot social behaviour, and where possible, how this related to learning outcomes. As the work in this chapter moves into the domain of language learning, literature from HRI focusing on this will now be discussed to inform the experimental design here. Some of the literature has previously been referenced in Chapter 2, but will be considered from a different perspective here, with a focus on the types of tasks and learning taking place.

8.1 Language Learning and Social Robots

Using social robots to teach a foreign language has been explored by different researchers with various aims. This section will discuss studies with children for the purpose of informing the design of the experiment in this chapter, and to provide context for the findings here. A summary of the studies considered in this section can be seen in Table 8.1.

Reference	Language	Aspect	Child Age	n	Research Topic	Details/outcomes
Alemi et al. (2014)	English (L2)	Vocabulary	12	60	Robot presence	5 week, multi session study. Children learn significantly more when robot is present
Gordon et al. (2015)	English (L1)	Vocabulary	3-8	48	Social behaviour (curiosity), Embodiment (robot/tablet)	Learning was significant (above chance), but averaged 1 word per interaction
Herberg et al. (2015)	French, Latin	Verb conjugations	10-12	23	Social behaviour (watchfulness)	Being watched by the robot leads to less learning, but there is a potential confound in terms of general motion
Kanda et al. (2004)	English (L2)	Vocabulary	6-7, 11-12	119, 109	Robot presence	2 week study. No overall learning effect, but effect for children who interact more with the robot in the second week.
Kory Westlund et al. (2015)	English (L1)	Vocabulary	4-6	19	Embodiment (robot/tablet/human)	Learning equal, but only 6 words are used
Saerbeck et al. (2010)	Toki Pona	Pronunciation, Vocabulary, Grammar	10-11	16	Social behaviour (socially supportive)	Significant learning is observed for the robot with socially supportive behaviours
Tanaka & Matsuzoe (2012)	English (L2)	Vocabulary	3-6	17	Robot presence	Significantly greater learning when interacting with a robot compared to when not

Table 8.1: A summary of studies conducted in HRI to investigate different aspects of robots on child language learning. *L1* indicates where English was being taught to native speakers, whereas *L2* refers to cases where English was being taught as a foreign language.

Alemi et al. (2014) employed a social robot as an assistant to a teacher over a 5 week period to teach English vocabulary to Iranian students. It was found that the class with the robot assistant learnt significantly more than those with just the human teacher. In addition, this vocabulary was retained, as measured through a retention test. This builds on earlier findings by Kanda et al. (2004) where a 2 week study with a robot situated in the classroom revealed a connection between interacting with a robot and vocabulary acquisition. Further results by Tanaka & Matsuzoe (2012) also confirm that the presence of a robot leads to significant learning of vocabulary. All three of these studies used children learning English as a second language and took place 'in the wild', with child ages ranging from 3 to 12.

Other researchers have explored the impact of embodiment and social behaviour for children learning English as a first language in a laboratory setting. Neither Gordon et al. (2015) or Kory Westlund et al. (2015) found significant differences due to embodiment in their studies of vocabulary acquisition by children. However, this may be due in part to methodological limitations. Gordon et al. (2015) only find an average of 1 word learnt per interaction, leaving very little room for observing differences; similarly Kory Westlund et al. (2015) only compares the learning of 6 words. These studies were conducted with children aged between 3 and 8. The relatively small gains are therefore quite surprising, due to the speed at which children of this age can theoretically acquire language (Kuhl, 2010). Due to the novelty of methodologies and comparison with both tablets and humans, it is suggested that the small learning gains are likely not due to the use of a robot.

Social behaviour has previously been studied in the context of children learning languages that were novel to them. Saerbeck et al. (2010) explored the impact of 'socially supportive' behaviours on child learning of the Toki Pona language, using an iCat as the tutor. These behaviours included verbal and non-verbal manipulations which aimed to influence feedback provision, attention guiding, empathy and communicativeness. It was found that the tutor with these socially supportive behaviours led to significantly greater child learning when compared to a neutral tutor. This study used a variety of measures including vocabulary acquisition, as other studies have, but also included pronunciation and grammar tests. Another study which did not consider only vocabulary acquisition can be seen in Herberg et al. (2015). French and Latin verb conjugations were taught by an Aldebaran NAO to children aged 10-12. In one condition, the robot would look towards the student whilst they completed worksheets, but in the other, the robot would look away.

While gaze towards the child was predicted to lead to greater social facilitation effects (and therefore higher performance), this was not found.

As discussed in Chapter 2, certain aspects of verbal immediacy have previously been explored in HRI. However, these research efforts did not frame the work in terms of immediacy, or focus on learning. Nevertheless, they provide promising indications of positive outcomes when applying high verbal immediacy social behaviour on robots interacting with children. It has been found that 'off-activity talk' - dialogue with a robot which does not concern the task being completed - encourages compliance in children in a therapeutic setting (Kruijff-Korbayova et al., 2014). Personalisation in therapeutic contexts has also been considered. Children were asked a number of questions about their preferences and the robot then mentioned these in an interaction, the children who interacted with a personalised robot enjoyed the interaction more, but subject numbers were too low for statistical comparisons (Blanson Henkemans et al., 2013).

In summary, many promising results have been found when robots have been used to tutor children aspects of language. The presence of a robot produces clear advantages when tutoring (Alemi et al., 2014; Kanda et al., 2004; Tanaka & Matsuzoe, 2012). However, the impact of social behaviour is less clear, with some positive results (Saerbeck et al., 2010), but other negative ones (Herberg et al., 2015). Verbal immediacy will be used in this chapter to further explore the impact of robot social behaviour on child language learning. Despite an increasing interest, there are still relatively few studies that have considered robot language tutoring, leaving space to explore novel aspects of language learning, which will be done here.

8.2 Hypothesis

Following on from previous research with humans (Gorham, 1988) and robots (Blanson Henkemans et al., 2013; Kruijff-Korbayova et al., 2014) this chapter aims to test whether robot verbal immediacy has a positive impact on children's second language learning as predicted by the literature. In order to make such an assessment, it first needs to be clear that children perceive the behaviour of the robot as intended, so a manipulation check will be conducted. Verbal immediacy provides a basis for measuring the children's perceptions and also for motivating differences between robot conditions. To ensure that any observed learning effects are retained and not just the product of short-term memory recall, the aim

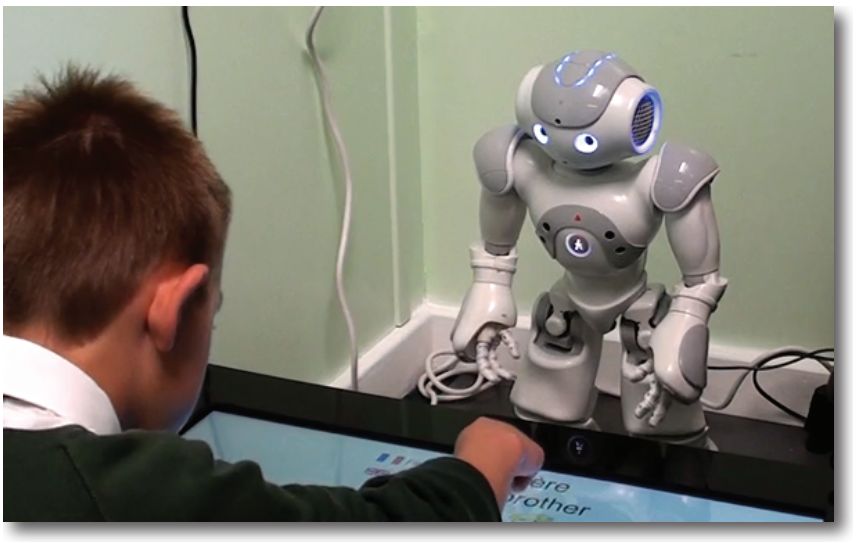

Figure 8.1: A child answering a question on screen during the interaction.

is also to verify children's retention of the material outside of the short-term interaction context (as in Tanaka & Matsuzoe 2012). As such, the hypothesis below will be tested in the short-term, and after a delayed period:

H1: A robot exhibiting more verbal immediacy behaviour will lead to greater child learning gains than a robot without this behaviour.

8.3 Experimental Setup

The study used 2 robot conditions (high vs. low verbal immediacy) in a between-subjects design, with an additional control condition (no intervention) to verify that a practice effect was not introduced from exposure to the learning test. French is commonly taught in English schools, so would have clear relevance for the children. However, it does not receive very much lesson time (the majority of schools offer 30-45 minutes per week at the age used in this study; Board & Tinsley, 2015), so there is plenty of scope to teach new concepts. As such, French was selected as the second language to teach in this study. The learning material was developed in collaboration with an academic researcher in language development, a native French speaker, and a teacher.

The structure of the lesson content was designed based on previous work in which children learnt mathematical concepts, such as Kennedy et al. (2015d), and a pilot study involving a human tutor and children. The aim was for the children to learn that nouns in French have a gender, that this changes which article is used ('le' or 'la'), and that for some words

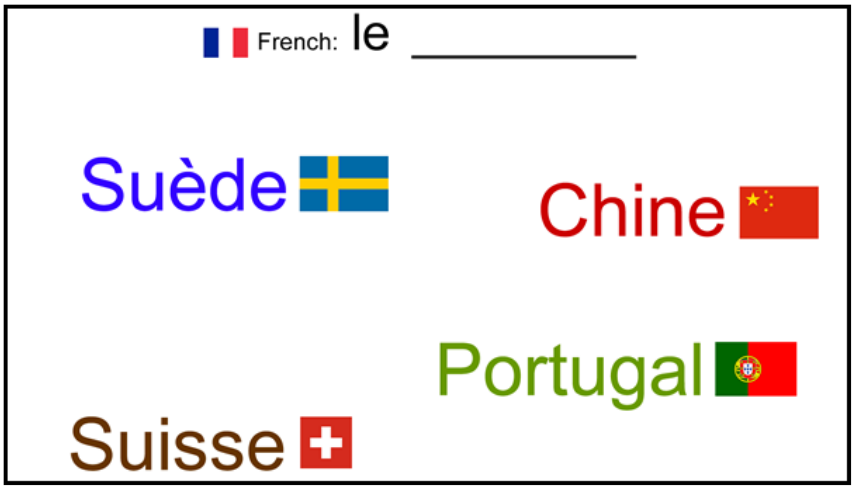

Figure 8.2: Screenshot from the touchscreen showing a question. Children can touch a word, drag it to the blank space and release to answer. Here the correct answer being 'Portugal'.

there are patterns which can be used to help work out which article to use. This is a novel learning topic for HRI.

An Aldebaran NAO robot acted as a tutor, delivering all lessons through speech and moving words on a touchscreen (Figure 8.1). As such, the children were exposed to both the words' pronunciation and orthography. The robot demonstrated how questions could be answered by dragging and dropping the correct answer in the blank space (see Figure 8.2). The robot first explained the concept of words having a gender by using an English example (using 'waiter' for a man, and 'waitress' for a woman). Following this, it explained how the French word for 'the' could be 'le' or 'la' depending on the gender of the noun it precedes. The robot then explained rules for working out whether to use 'le' or 'la'. After explaining each rule, the child's understanding was checked (Figure 8.3).

During the lessons the robot would explain a rule and then use the screen to show an example. The rules used were taken from online French language learning guides[1,2] and were verified by a French native speaker. The rules were as follows: 1) 'le' is used for male people, and 'la' is used for female people, 2) 'la' is used for countries ending in 'e', 3) 'la' is used for fruit or vegetables ending in 'e'. Whilst these are recognised techniques for people learning a second language, it should be made clear that it is unlikely that a native speaker would learn in this way, and that there are a limited number of exceptions to rules 2 and 3 (but these were avoided in the lesson content here). As described in Chapter 2, it is not intended that the best teaching strategy for the concept is determined, but that the

[1] http://goo.gl/JPjmPO
[2] https://goo.gl/WY37z5

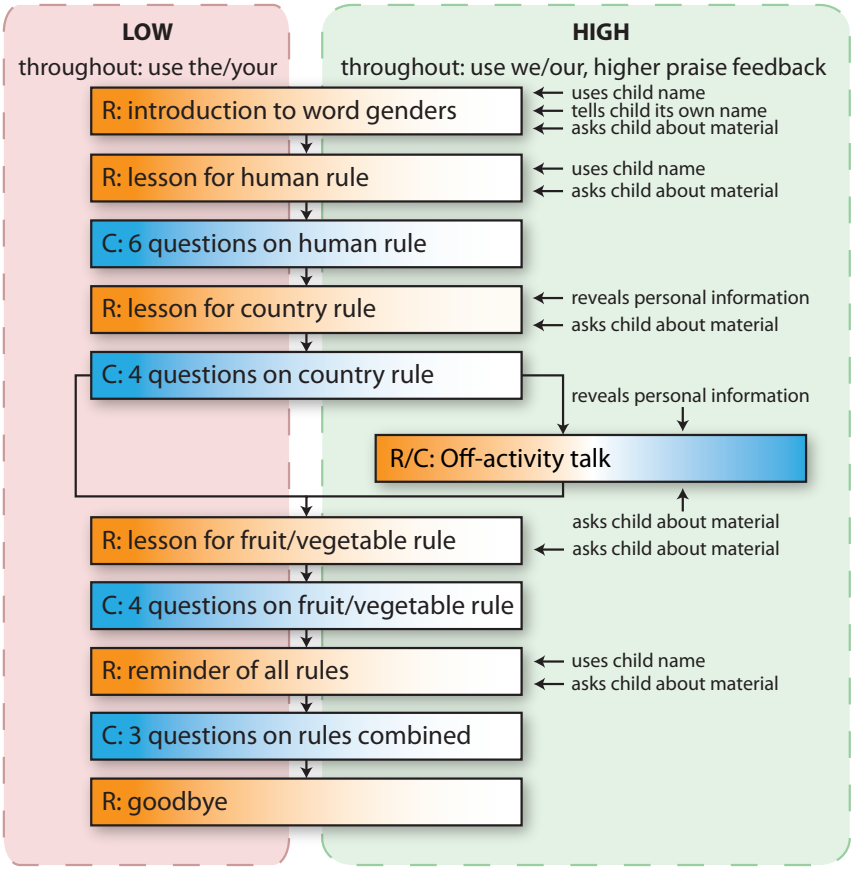

Figure 8.3: Structure of the task. *R* refers to robot explanation sections and *C* refers to child question answering sections. The robot dictates the structure of the interaction through speech and by presenting questions on the touchscreen, informing the child of when it is their turn answer questions on the screen. The HIGH condition includes many manipulations in the verbal behaviour to make it more 'available'.

effect of robot behaviour on learning is ascertained.

Questions were designed to get progressively more complex as the interaction progressed. To start with, English translations and pictorial representations of the words were provided alongside the French. At this stage, the child was only required to select the article 'le' or 'la' to add to the word. Towards the end of the interaction, all English translations were removed so that only the French and the pictures remained. The question structure was also changed in later stages: the child was required to match a noun to the article (Figure 8.2), which requires them to assess several nouns for each question, rather than just one as in the earlier questions.

All feedback was provided verbally by the robot; no feedback was shown on the screen. When providing feedback, the robot's TTS would switch to French so that the child could hear the correct pronunciation. The robot was autonomous throughout, except for some short vocal phrases in one condition, which were triggered by the experimenter (see

Section 8.4.3).

8.4 Evaluation

8.4.1 Participants

A total of 67 children were included in the study after exclusions due to technical issues (1 child) or absence from school during one of the two visiting periods (7 children). All children were native English speakers and were from the same year group (with three class teachers) from a primary school in the U.K. (average age M=8.8, SD=0.4; 30M, 37F). Only one child was fluent in another language (this language was not used in this study). Children were distributed randomly between groups whilst balancing for gender and class teacher. All children had parental/guardian permission and gave their consent to take part in the study.

8.4.2 Measures

Learning was measured through pre-, post- and retention tests, which can be seen in Appendix J. These tests sought to examine various aspects of the children's learning, including their vocabulary acquisition, and their ability to apply each of the 3 rules in isolation and combination with each other. The test consisted of 12 questions: 3 vocabulary-based (1 French-English and 2 English-French), 2 about humans (rule 1), 2 about countries (rule 2), 3 about fruits and vegetables (rule 3), and 2 combined all three rules. Each question had 4 multiple choice answers and used the same formats as questions on the touchscreen. The majority of the test questions used words that the children had not seen in the learning material in order to ensure generalised learning was taking place, rather than memorisation of specific instances; exceptions are discussed in Section 8.6. The pre-, post- and retention-tests were all the same as this was necessary to account for children's prior knowledge (they had learnt some French vocabulary in school before), and to accurately measure their recall. The children were not given any feedback on their tests at any stage.

The child's perception of the robot was measured through a questionnaire combining verbal immediacy and nonverbal immediacy items. This 23 question questionnaire was completed on paper and was multiple choice. The verbal immediacy and nonverbal

immediacy items were based on those used in Wilson & Locker Jr (2007), but were modified such that the language could be understood by children. The final questionnaire used can be seen in Appendix H. Verbal immediacy includes aspects of behaviour such as personalisation, off-activity talk, and student opinion solicitation. Nonverbal immediacy covers overt social behaviours, such as whether gestures are used, whether the robot looks at the child, and so on.

8.4.3 Conditions and Robot Behaviour

In order to address the hypothesis in Section 8.2, three conditions were devised: 1) a robot with high verbal immediacy (HIGH, n=20), 2) a robot with low verbal immediacy (LOW, n=20), 3) a control with no robot and just a pre- and retention test (CTRL, n=27). The robot with low verbal immediacy doesn't have the verbal behaviours which lead to being considered available as defined by verbal immediacy (Figure 8.3). The control condition is used to verify that there are no learning effects due to exposure to the test material.

In both robot conditions, the nonverbal behaviour was kept constant. The behaviour used was designed to be of high nonverbal immediacy, with the robot's gaze randomly moving in the direction of the child, gestures during speech, a slight lean forward of the body, and slight motor noise in the arms to give the impression of being relaxed. The perception of this behaviour as being of high nonverbal immediacy is verified through the questionnaire after the interaction (as described in Section 8.4.2).

The speech of the robot was kept the same in both conditions outside of the experimental manipulations as described below. This ensures that the lesson content is largely unchanged between conditions, although the experimental manipulations require some language adjustments, these should not impact on the coherence or intelligibility of the lessons.

The verbal immediacy questionnaire (Gorham, 1988) introduced in Chapter 3 was used to create the robot conditions with different immediacy levels. In order to generate the behaviour for the conditions, all of the verbal immediacy questionnaire items possible were applied to the speech for the HIGH condition, and were not applied for the LOW condition. The following differences were present in the HIGH condition robot behaviour, but not in the LOW condition:

1. use the child's name (3 times)

2. tell the child its name
3. reveal personal information about itself (twice in addition to its name)
4. ask the child how they felt about the material (e.g., "does everything make sense to you so far?" 6 times)
5. ask the child about their hobbies and continue the discussion for 2 or 3 speech turns
6. use "we/our" work (as opposed to "the/your", throughout)
7. provide higher praise feedback (e.g., "You're doing really well! That was right", as opposed to simply "That was right" in the LOW condition)

Two items of the verbal immediacy questionnaire were not manipulated: humour and feedback provision. Humour was considered to be inappropriate to add given the context of the interaction and difficulties in selecting a comment that would be universally funny. Whether or not feedback was provided was not manipulated between conditions as in this context, the only way of getting feedback was from the robot and missing feedback here would confound any findings related to learning.

To compensate for unreliable speech recognition, a Wizard-of-Oz intervention was used in the HIGH condition to let the robot reply 'that's great' after the children answered a question from the robot about their understanding of the material (children always said they had understood the lesson), and to trigger pre-scripted phrases at the appropriate time for the discussion about the child's hobby. A video figure demonstrating the differences between the conditions can be seen online via the ACM digital library: `http://dl.acm.org/citation.cfm?id=2906873`; follow links to 'Source Materials' and then 'suppl.mov'.

8.4.4 Procedure

The interactions took place in a quiet working space on the school premises familiar to the children. The child sat across from an Aldebaran NAO with a 27 inch touchscreen placed horizontally between them (Figure 8.1). Two video cameras were used to record the interactions. One experimenter sat behind and to the side of the child, out of their view (Figure 8.4). The time children spent interacting with the robot was on average M=11min 26s (SD=1min 11s).

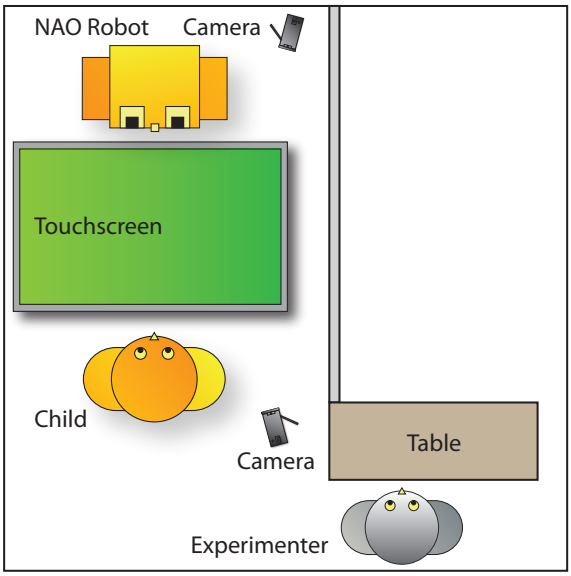

Figure 8.4: Schematic overview of the interactions being investigated in this paper. The child and the Aldebaran NAO robot sit across a touchscreen from one another. An experimenter sits behind and out of view of the child. Two video cameras record the interaction. Figure not to scale.

The experimenter spent a full week in the school, plus one day the following week. On the first Monday of the visit, pre-tests were delivered to all children in their main classrooms. These were completed under the supervision of the experimenter and the class teacher to make sure that children completed them individually. Throughout the week those children interacting with the robot would be taken out of class individually, take part in the interaction, and then complete the post-interaction test and questionnaire on paper, to the side of where the experimenter had been sitting (so they can no longer see the robot or touchscreen). The robot condition was switched between each interaction to ensure a balance throughout the week.

On the Monday of the following week the experimenter returned to deliver the retention test to the children under the same conditions as the pre-test. Children in the control group therefore completed a pre-test and a retention test without any teaching input. The children had not been informed that they would be tested again on the material that they had covered with the robot. After each class had completed the retention test, the experimenter gave an overview of the study and a presentation of social robots to all children. This meant that all children understood the study and had the opportunity to interact with the robot.

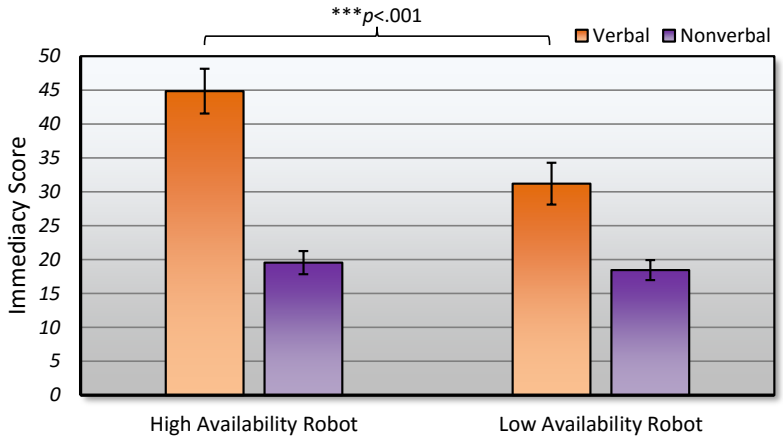

Figure 8.5: Verbal and nonverbal immediacy scores for the high immediacy (HIGH) and low immediacy robot (LOW) conditions. The HIGH condition is perceived to have significantly higher verbal immediacy while having the same nonverbal immediacy. *Error bars* show 95% CI.

8.5 Results

8.5.1 Perception of the Robot

To verify that children perceive differences in the verbal immediacy of the robot, the results of the post-interaction questionnaire were analysed. The questionnaire is broken down into the several parts which measure different constructs, as described in Section 8.4.2. The manipulations were conducted on the verbal immediacy element of the questionnaire. An unpaired t-test is used to compare between the two independent robot condition groups (Kolmogorov-Smirnov test reveals no significant deviation from normality, $p > .05$; Levene's test indicates homogeneity of variances, $p > .05$). The test reveals a significant difference between the average verbal immediacy measure for the LOW condition (M=31.2, 95% CI [28.1,34.3]) and the HIGH condition (M=44.9, 95% CI [41.6,48.2]); $t(38)$=6.322, p<.001. This confirms that the children could indeed perceive the difference between the conditions (despite not having seen the other condition for comparison).

Nonverbal immediacy scores were also compared using the same statistical test; the difference between the nonverbal immediacy score in the LOW condition (M=18.5, 95% CI [17.0,19.9]) was not found to be significantly different to that of the HIGH condition (M=19.6, 95% CI [17.8,21.3]); $t(38)$=1.020, p=.314 (Figure 8.5). This provides some validation for the control of nonverbal behaviour between the conditions.

8.5.2 Learning Gains

Learning gains are measured through scores on the tests conducted before the interaction (pre-test), immediately after the interaction (post-test), and 3-7 days after the interaction (retention test). Questions on the tests are equally weighted, so scores are out of a maximum of 12. Before analysis of the two robot conditions can be conducted there are some potential confounds which must be eliminated as factors: the differences in time between the interaction and retention test, and the impact of exposure to the test (as the same test is used).

It could be expected that children who interacted with the robot at a time closer to the retention test would outperform those who interacted with the robot earlier in the visit. To explore whether this was a factor, the day on which the interaction took place was correlated with the difference between the post-test and the retention test. The correlation is weak and non-significant; $r(36)$=-.079, p=.637, indicating that the time from interaction to retention test can be eliminated as a factor. It is suggested that the absolute number of days does not make a difference to the retention, but the number of days out of school during this period is more important, which was constant for all children (a weekend of 2 days).

The control condition is used to verify whether exposure to the test makes a difference to the findings. It would not be expected that there would be a difference as the children are given no feedback on the tests at any stage, but the control condition allows verification. Equivalency testing is used to demonstrate not only that there is no significant difference, but that there is significant equivalency. For children in the control condition, the pre-test score (M=3.96, 95% CI [3.26,4.66]) and retention test score (M=3.89, 95% CI [3.28,4.49]) can be considered equivalent. Two one-sided t-tests (TOST; Schuirmann, 1987) with a 1 point threshold confirm the test scores are equivalent at the $p<.05$ level: $t(52)=-2.061$, p=.022/$t(52)$=2.391, p=.010. This indicates that exposure to the test is not a confounding factor.

A repeated measures ANOVA was used to explore H1, that the robot condition affects learning, in both the short-term and after a delayed period. The repeated measures ANOVA was used as each test is a matched measure for that participant; 2 levels are used so that the effect of test and condition can be explored across the continuous measure of test score. Figure 8.6 and Table 8.2 show the results for test scores by condition.

Mauchly's Test of Sphericity indicated that the assumption of sphericity had not been violated[3], $\chi^2(2)=1.873$, $p=.392$. No significant interaction was found between test and condition; Wilk's Lambda=.998, $F(2,35)=0.04$, $p=.963$. A main effect was found for test, Wilk's Lambda=.391, $F(2,35)=27.21$, $p<.001$, but not for condition; $F(1,36)=0.08$, $p=.774$. Bonferroni pairwise comparisons find that there is a significant difference between pre-test and post-test, and pre-test and retention test scores (all $p<.001$), but no difference between post-test and retention test ($p=1.00$).

These results do not support H1, as children learn between the pre- and post-tests, and retain their learning in the retention test, but this does not differ by robot condition. Further to this, Weber & Popova paired-samples equivalency tests (Weber & Popova, 2012)[4] show that the post and retention test scores are equivalent in both the HIGH ($t(18)=0.67$, $p=.022$) and LOW ($t(18)=0.73$, $p=.025$) conditions, with Cohen's $d=.50$. Whilst this is an 'intermediate' effect size for demonstrating equivalency, it should be noted that the sample size is relatively small on a per-condition basis, leading to a higher variation in scores, which raises the level at which equivalency can be shown. Combined, these findings suggest that the children learn a significant amount from the pre-test to the post-test, and the post-test and retention test scores can be considered largely equivalent, demonstrating their retention of the learning in both conditions.

The ANOVA results do not support H1 (that higher immediacy will lead to greater learning) as no significant effect was found for robot condition. Nor can a significant difference be seen between the improvement in the LOW condition ($M=3.80$, 95% CI [2.55,5.05]) and the HIGH condition ($M=3.35$, 95% CI [1.78,4.92]); $t(38)=0.470$, $p=.641$. The drop in score from post-test to retention test can also be considered equivalent between conditions; using a Weber & Popova independent-samples equivalance test, $t(36)=0.07$, $p=.004$ with Cohen's $d=.50$. Therefore, Hypothesis H1 must be rejected as there are no significant differences observed between conditions in terms of learning.

Based on the rules taught to the children, one could suggest that learning a very simple rule of: "if the word ends in an 'e', then use *la*, otherwise use *le*" may be adopted as a 'shortcut' and could account for the learning differences. This would then have nothing to

[3]Sphericity is an important assumption for using a repeated-measures ANOVA. Sphericity indicates equal variances of the differences between all possible pairs of groups (levels of the independent variable). If this is not the case, then variance calculations may not be accurate, producing an inflated F-ratio. Further details about Mauchly's test in particular can be seen in Mauchly (1940).

[4]The Weber & Popova equivalency tests consider paired samples, whereas the TOST equivalency technique previously used is appropriate for independent samples.

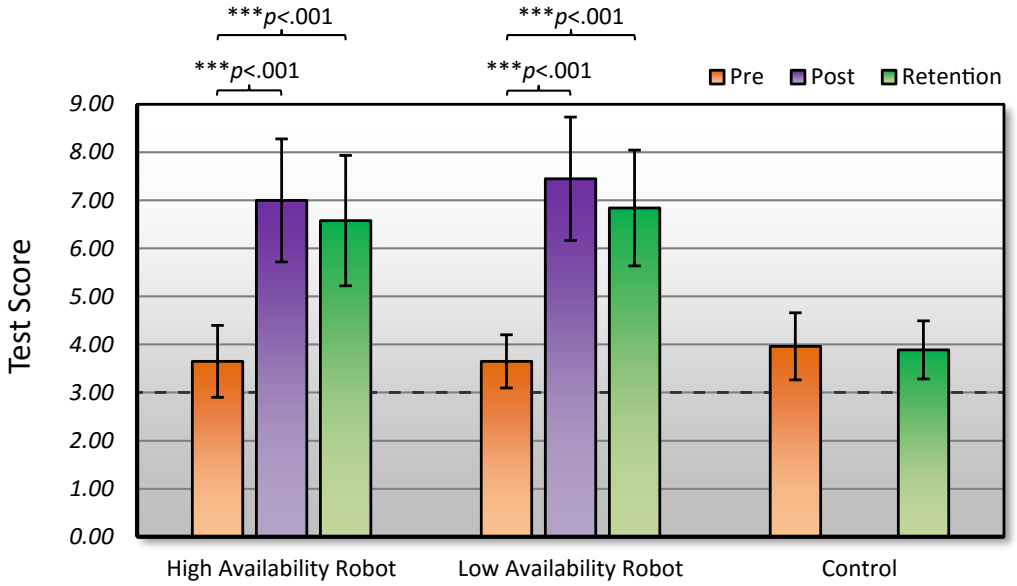

Figure 8.6: Pre-test, post-test and retention test scores by condition (chance score=3; maximum score=12). Children learn a significant amount from the robot between pre- and post-tests; this gain is sustained to the retention test. *Error bars* show 95% CI. Darker dashed line indicates the 'chance' baseline.

Condition	**Pre-Test** *M* **[95% CI]**	**Post-Test** *M* **[95% CI]**	**Retention Test** *M* **[95% CI]**
CTRL	3.96 [3.26, 4.66]	N/A	3.89 [3.28, 4.49]
LOW	3.65 [3.10, 4.20]	7.45 [6.17, 8.73]	6.84 [5.64, 8.05]
HIGH	3.65 [2.90, 4.40]	7.00 [5.72, 8.28]	6.58 [5.22, 7.94]

Table 8.2: Test score results for pre-test, post-test and retention tests, by condition.

do with learning aspects of language, but be a basic memory phenomenon. This had been anticipated in the study design, so later questions in the learning material made sure to challenge this approach by including several words ending in 'e' as possible answers, but with those words relating to humans of male gender (therefore requiring 'le', rather than 'la' and violating the shortcut rule). Additionally, a question in the tests used adopted this approach, with several words ending in an 'e', but not all being feminine. This was done to verify whether the shortcut rule had been adopted, or whether the children had really learnt the material as it had been taught, with the ability to discriminate between different types of words. If the children had only learnt the shortcut rule then they would answer this verification question incorrectly, however, it was answered correctly above the average level for the rest of the questions in the test (63% for the verification question, versus 60% for the other questions). This provides some evidence that the children learnt intricacies of the language that was presented to them; further evidence in support of this will be provided in Section 8.6.

8.6 Discussion

The results show that the children perceived the verbal immediacy of the robot conditions as intended, which confirms that the behaviour was designed appropriately to address the research hypothesis. The nonverbal behaviour was kept constant between the two conditions, and this was reflected in the children's questionnaire responses. The children in both robot conditions exhibited significant learning gains between the pre-test and post-test, as well as between the pre-test and retention test, with equivalent scores in the retention test and the post-test. This is a positive result, as it would have been plausible that the children would quickly forget what the robot had taught them once the interaction was over, especially as the children were not aware that they would be re-tested, and so had little motivation to attempt to actively try and retain the information.

The tests which the children had to complete were designed to be challenging. Each answer had four options with no obviously incorrect answers, so the likelihood of a guess being correct would be chance (25%). It was found that children scored slightly above this on the pre-tests as they had done a small amount of French before, so scored closer to 4 than the 3 that would be expected with random guessing. This significantly improved to over 7 out of 12 in the post-tests. Given the difficulty of the tests and the relatively short time the child interacts with the robot learning and practising the material, this is an impressive increase. Indeed, only 6 of the 40 children who interacted with the robot did not improve from pre-test to post-test. Learning of 'le' or 'la' as the article choice could have contributed to part of the increase in scores, however if children had learnt the choice to be le/la then the chance score would go up by 1.5 points from pre-test (chance = 3) to post-test (chance = 4.5). The children actually improve by an average of 3.6 (95% CI [2.6,4.5]), suggesting learning beyond any improvement due to the higher chance score.

Despite the children being able to perceive the difference in verbal aspects of immediacy between the two robot conditions (measured through verbal immediacy), no significant difference was observed in learning in either the post- or retention-test. This finding is surprising given the positive correlation between verbal immediacy and learning in human studies (Gorham, 1988; Witt et al., 2004). The high verbal immediacy robot condition was perceived to have significantly higher verbal immediacy, but this did not translate into additional learning gains as expected. The content of the lessons was the same, but the

difference to learning that social behaviour can make has previously been demonstrated in HRI in spite of the same content between conditions for nonverbal behaviour (Kennedy et al., 2015e; Szafir & Mutlu, 2012). However, the learning material here is different to prior work, so it is possible that the content under consideration interacts with the robot social behaviour, influencing the learning effects.

Aspects of the behaviour manipulated here, such as personalisation (Blanson Henkemans et al., 2013) and off-activity talk (Kruijff-Korbayova et al., 2014), have been studied before in HRI with promising results. However, these studies had too few subjects to make conclusions about learning (Blanson Henkemans et al., 2013), or did not assess learning (Kruijff-Korbayova et al., 2014). In contrast to Kruijff-Korbayova et al. (2014), the children here do perceive differences between the conditions, but in this study the questionnaire is targeted towards specifically measuring the perception of the behaviours which were manipulated, rather than assessing an overall feeling towards the robot. It is possible that despite children perceiving differences in the immediacy of the robot, this did not translate into any difference in feeling towards the robot. If the relationship the child feels towards the robot is no different between conditions then this may go some way to explaining the lack of difference in learning.

The interpretation of the robot character could have been influenced by the TTS voice used by the robot, which would switch when the language changed. These voices were clearly different and this could have impacted how the children perceived the robot. However, the children have no prior experience with the robot, so they may have accepted this as part of the robot's behaviour. As the voices are clearly different, they may also have interpreted this not to be part of the robot's character, but to be the robot playing back other media (akin to a teacher playing recorded French). It is not possible to determine how the children perceived this switch in voice from the data collected, but perceptions of voice switching of multi-lingual robots could be worth explicitly exploring in future work.

Another factor which may have influenced the learning results is novelty. Novelty is often an issue for HRI studies (Leite et al., 2013b; Sung et al., 2009), and it possibly played a role here as the children interact just once with the robot for a brief period of time. Verbal immediacy has been found to consist of four factors, including 'individual friendliness' (Wilson & Locker Jr, 2007). Even if the children were to bond more strongly with the high immediacy robot because of increased friendliness, the short interaction time might not be

enough for differences in the relationship to manifest into learning outcomes. Furthermore, it could be that the behaviour of the more available robot cancels out its own benefits by being so novel as to distract from the learning material. For example, when the robot is conducting off-activity talk during the interaction, this is time when the children are not focussing on the learning task and are possibly forgetting information they have learnt. This doesn't mean that off-activity talk should be avoided for fear of distraction, but that it might only be appropriate in longer, or repeated interactions where novelty is less of an issue. It is hypothesised that given a longer interaction timescale, the learning benefits predicted by the literature of greater immediacy (Gorham, 1988; Witt et al., 2004) would be observed as the novelty wears off (Kanda et al., 2004; Leite et al., 2013b).

In the HHI literature, a lower correlation between verbal immediacy and learning has been found when compared to nonverbal immediacy and learning (Witt et al., 2004). Nonverbal immediacy has previously been found to make a difference to learning in HRI (Kennedy et al., 2015e; Szafir & Mutlu, 2012). This could suggest that verbal behaviour may not be as important for learning (at least in short-term interactions) as overt nonverbal behaviour. It has also been found in humans that the impact of immediacy behaviours is enhanced in line with increases in class size (Gorham, 1988). It could be that the effect of verbal immediacy is simply too far reduced when placed in a one-to-one tutoring context as in this study, rather than the larger classroom setting. The immediacy of the robot would be experienced to some extent in both conditions simply through the nature of the one-to-one interaction.

One interesting finding from the data collected which was not hypothesised was the ability of the children to acquire vocabulary despite the learning material not explicitly requiring them to do so. Three questions of the test were vocabulary based: two requiring translation from English to French, and one French to English. Two of these questions referred to words which the children would have seen on screen and heard the robot say (as they were answers to questions in the learning material). The remaining question was about a word which they would have seen on screen, but the robot did not say (as it was not a correct answer). It is suggested that the two words which were answers in the learning material were more likely to be recalled as the children would have looked at the word for longer and the robot would have said the word. However, a significant increase was found for all 3 of the questions independently, and a repeated measures ANOVA found a significant increase for the average score (out of 3) of children who

correctly translated the words from pre-test to post-test, and from pre-test to retention test. Mauchly's Test of Sphericity indicated that the assumption of sphericity had not been violated, $\chi^2(2)=0.661$, $p=.719$. No significant interaction was found between test and condition; Wilk's Lambda=.968, $F(2,35)=0.58$, $p=.565$. A main effect was found for test, Wilk's Lambda=.595, $F(2,35)=11.94$, $p<.001$, but not for condition; $F(1,36)=0.14$, $p=.710$. Post-hoc Bonferroni pairwise comparisons find that there is a significant difference between pre-test (M=0.8, 95% CI [0.6,1.0]) and post-test (M=1.6, 95% CI [1.3,1.9]), and pre-test and retention test (M=1.4, 95% CI [1.1,1.7]) scores ($p<.001$ and $p=.001$, respectively), but no difference between post-test and retention test scores ($p=.883$).

It is of course possible that the children remembered the words from the pre-test and made an effort to learn these words when they were presented on screen, but this seems unlikely given the time (up to 4 days) between many of the pre-tests and the interactions, and the sheer number of words they were exposed to in the learning content (over 40). For a child to concentrate on learning 3 words from the pre-test, days after having seen it, when being taught a different aspect of language would seem to be highly improbable. As such, this is a promising finding with robots that confirms data from human-human literature whereby children of this age will acquire language through exposure in social interactions (Kuhl, 2010).

8.7 Summary

Children perceived the relative verbal immediacy of the two robot conditions as intended in the design. This confirms that the manipulations made were appropriate to address the question of whether an increase in verbal aspects of immediacy would lead to an increase in learning. As expected, the children did learn elements of a second language from the robot. This was measured immediately after the interaction and also some days later. The retention test scores were slightly lower than the pre-test scores, but can be considered statistically equivalent. However, surprisingly there was a lack of any significant difference between conditions in the immediate post-test score, or the longer-term retention test score. Literature from human-human interaction studies (Gorham, 1988; Witt et al., 2004) and human-robot interaction studies (Blanson Henkemans et al., 2013; Kruijff-Korbayova et al., 2014) would predict an increase in robot verbal immediacy to lead to an increase in learning, but this was not found. It is suggested that in this short-term dyadic interaction

context, additional effort in developing social aspects of a robot's verbal behaviour may not return the desired positive impact on learning gains.

Chapter 9

Synthesis of a Framework for Learning and Social HRI

Key points:

- Previous chapters have made many comparisons between different robot social behaviours and their impact on child learning. This chapter draws these behaviours together through the common metric of nonverbal immediacy.
- A further experimental condition is devised and data is gathered using a human tutor for the prime numbers task to provide a contextual baseline for the robot behaviours.
- Nonverbal immediacy does not fully explain the differences in child learning between robot conditions.
- A framework for child learning as a product of social cues and cue congruency is put forwards to provide an explanation for the data here and to make predictions for future work.

The work presented in this chapter is under review in Kennedy et al. (under review).

The primary tenet of the thesis exists in the relationship between robot social behaviour and child learning. This section will seek to discuss this relationship in detail, drawing on the experimental work conducted throughout this document, alongside additional data collected here. First, the methodology of the additional experimental condition will be outlined. This will be followed by a critical discussion of nonverbal immediacy in relation to the thesis and the research questions identified in Section 1.2.

9.1 Human Prime Tutoring

One question that often arises, particularly from non-academic audiences, is how robots compare to human tutors. The aim of research into robot tutors is rarely to replace human teaching, but to supplement it, so such a comparison is not typically part of experimental hypotheses. Given the link between robot social behaviour and learning (Chapter 7), human behaviour is often used to derive behaviour for robots to provide an upper benchmark of social behaviour that robots can aim for in tutoring. The literature from other fields suggests that human tutoring also provides an upper benchmark in terms of learning gains. VanLehn (2011) finds that human tutoring produces an effect size of d=0.79, while ITS produces an effect of about d=0.76. However, this comparison has not been verified in HRI.

Serholt et al. (2014b) found no significant difference between the performance of children who had been tutored by a humanoid robot compared to a human, but the robot speech was controlled using a Wizard-of-Oz method, introducing additional variability between conditions. The present section reports on data collected from running an additional condition using the prime tutoring task, in which the lesson content was delivered by a human. The purpose of this additional data collection is to provide a benchmark for child learning when a human delivers exactly the same content, but using their natural social cues. The findings will be considered in later sections in combination with robot conditions from this thesis.

9.1.1 Methodology

This study employs one condition: a human tutor teaching the prime numbers task to provide a benchmark comparison for the robot conditions used in previous chapters. The

study employs the same methodology as seen in Chapters 6 and 7. Children aged 8 and 9 engage in a dyadic interaction in their school with a tutor who guides them through a method for prime number identification. The children's learning is measured through a pre-test and a post-test consisting of 12 numbers which need to be categorised as 'prime' or 'not prime' (6 per category). Prior to the interaction, children have not learnt about prime numbers, but the technique relies on their ability to divide by 2, 3, 5 and 7, so this is also tested. The tutor provides hints to help with the division, as well as a lesson about how to identify prime numbers using the Sieve of Eratosthenes technique. Two tests for prime number identification are used in a cross-testing strategy to control for exposure to the tests.

The human was given a word-by-word script to match the lesson content of the robot from Chapter 7, but was not constrained in terms of social behaviour. Due to the script providing precise lesson content (and the study focus on social behaviour and embodiment differences) an expert tutor was not required. 11 children took part in the study (age M=8.8, SD=0.4), with interactions lasting for M=13m10 (SD=3m39).

9.1.2 Results

The pre- and post-test scores are paired values on a continuous measure, distributions did not significantly deviate from normality (Kolmogorov-Smirnov test; $p > .05$). As such, a paired samples t-test is used for analysis. Children improve significantly; the post-test score (M=7.6, 95% CI [5.5,9.8]) is significantly higher than the pre-test score (M=5.2, 95% CI [3.7,6.7]); $t(10)$=2.425, p=.036 (Figure 9.1). The effect size seen here is d=0.89, which is not so dissimilar to that seen in VanLehn (2011), where human tutoring averages d=0.79. As such, this provides some reassurance of the validity of the tutoring undertaken. Although it should be noted that the effect sizes in VanLehn (2011) compare to a no tutoring control, which is not done here since the nature of the task makes learning unlikely without tutoring.

It is also worth noting that the mean was lowered by one instance where the child had clearly learnt the technique, but confused the categories, and so scored 0 on the post-test (i.e., 100%, but incorrect). The child asked for clarification, but as this help would not have been available when the content was delivered by the robot (as in Chapter 7), it was not given by the human at the time.

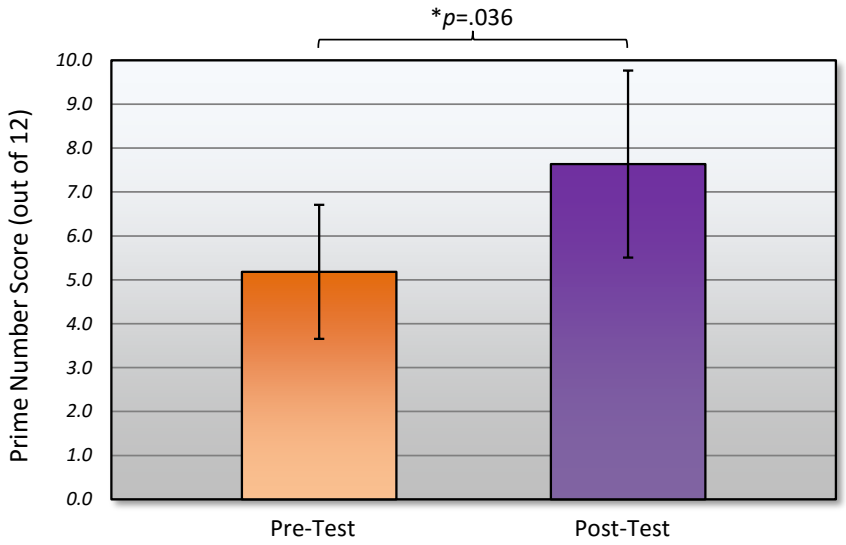

Figure 9.1: Pre-test and post-test scores for prime number learning of children when tutored by a human. The children improve significantly from the pre-test to the post-test. *Error bars* show 95% Confidence Interval.

The specific human used in the tutoring task may have had an impact on the results. The social behaviour was not constrained, which meant that the human could take advantage of some social cues that the robot could not, and could subsequently be more socially adaptive (for example, in mutual gaze) than a robot, which may account for some of the learning differences. A non-expert human was used due to the tightly specified learning content, but an expert tutor may have used different social behaviour, potentially leading to more learning. Overall, the children rated the human behaviour nonverbal immediacy on average M=54.4 (95% CI [52.9,55.9]) using the CNIQ from Appendix F. The findings here will subsequently be further explored in the context of the robot conditions previously used in the prime tutoring task.

9.2 Nonverbal Immediacy and Learning

Five different conditions have been used in the prime number learning task[1]. For three of these conditions, nonverbal immediacy scores were collected directly from the children, and for all five, adult nonverbal immediacy scores are available. Taking nonverbal immediacy as a characterisation of robot social behaviour (Kennedy et al., 2015a), these results can be used to support the central thesis: a robot with tailored social behaviour will improve child learning. The graphs in Figures 9.2 and 9.3 show a clear trend towards grea-

[1]This section refers only to the physically embodied conditions, not the 'no lesson', or 'no robot' control conditions presented in Chapter 6.

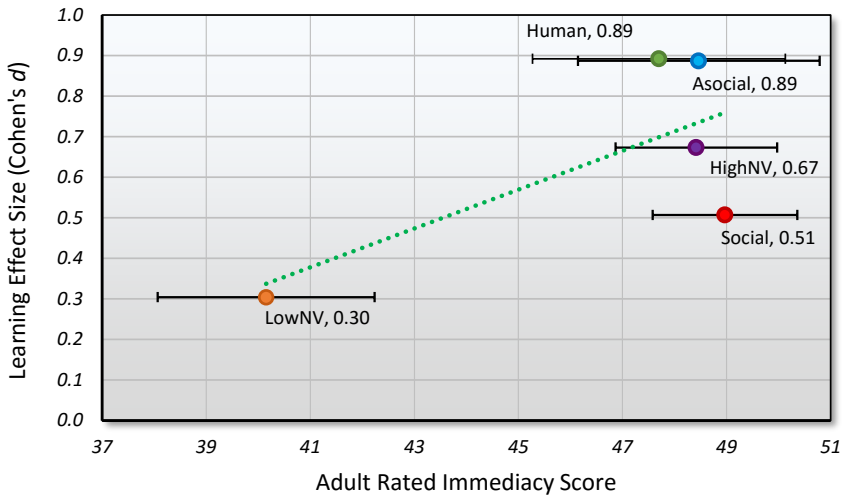

Figure 9.2: Nonverbal immediacy scores as judged by adults and learning effect sizes for the prime number task. The dotted green line indicates a trend towards greater nonverbal immediacy of the tutor leading to increased learning. *Error bars* show 95% Confidence Interval.

ter nonverbal immediacy leading to increased child learning. This is further supported through the strong positive correlation between the learning effect sizes (Cohen's d) and the nonverbal immediacy scores (as judged by adults); $r(3)=0.70$. This provides additional support for the thesis on top of that presented in individual experimental chapters.

Much of this thesis considers social behaviour as characterised through immediacy. It is recognised that immediacy is not a complete measure of social behaviour; indeed it has several shortcomings as such (detailed shortly). Nevertheless, immediacy provides a characterisation of the social behaviour through the overt cues displayed by the robot. Using these overt cues allows more meaningful responses from children than other surveys of child perception typically do (Kennedy et al., 2015f), and Chapter 4 showed that nonverbal immediacy allows transfer from existing HHI literature to cHRI. These findings, combined with the clear relevance of social cues to social behaviour, support the use of immediacy as a proxy for social behaviour.

However, nonverbal immediacy does not account for all of the differences in learning. Three of the conditions have near identical NVI scores as judged by adults, but quite varied learning results (high NVI robot: $M=48.4$ NVI score/$d=0.67$ pre-post test improvement, asocial robot: NVI $M=48.5$/$d=0.89$, social robot: NVI $M=49.0$/$d=0.51$). This partially reflects the slightly mixed picture of immediacy that the HHI literature presents, for example, the disagreement as to whether NVI has a linear (Christensen & Menzel, 1998) or curvilinear (Comstock et al., 1995) relationship with learning. Nonetheless, there are

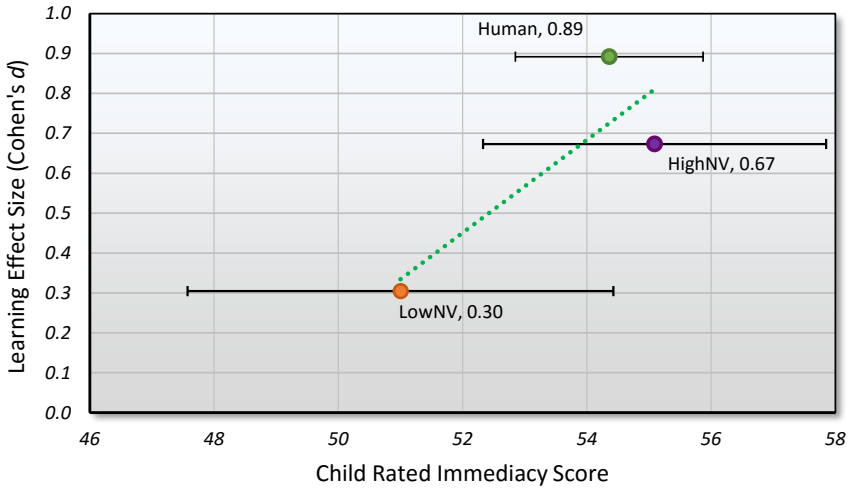

Figure 9.3: Nonverbal immediacy scores as judged by the children in the interaction and learning effect sizes for the prime number task. The dotted green line indicates a trend towards greater perceived nonverbal immediacy of the tutor leading to increased learning. *Error bars* show 95% Confidence Interval.

further factors that may be introduced by the use of a robot that may have had an influence on the results. Nonverbal immediacy only considers overt observed social behaviours, so by design does not cover all possible aspects of effective social behaviour for tutoring. Whilst this seems to be enough in HHI (Witt et al., 2004), it may not be for HRI. Several possible explanations as to why this learning variation is present will now be discussed. From this, a possible model (suggested to be more accurate) of the relationship between social behaviour and learning is proposed. Such a model may be useful in describing (and testing) the relationship between social behaviour and child learning for future research.

9.2.1 Timing of Social Cues

The quantity of social cues used in both the social robot and asocial robot conditions is exactly the same, however the timing is varied. Timing is not considered as part of the NVI metric - the scale measures whether cues have, or have not, been used, rather than whether their timing was appropriate. The cues used in the asocial robot condition were intentionally placed at inappropriate times (for example, waving part-way through the introduction, instead of when saying hello). This is not factored into the NVI measure, but could impact the learning (Nussbaum, 1992).

The timing of social cues in the human condition may also explain why the learning in this condition was higher than the others. The robot conditions are all either non-adaptive, or minimally adaptive (the social robot condition will seek mutual gaze using the Kinect

sensor) in the timing of cues. However, the human is presumably adaptive in both the number of social cues used and the timing of these cues. Again, this would not be directly revealed by the NVI metric, but could account for some of the learning difference. Indeed, the NVI metric comes from HHI studies and has been validated in such environments. In HHI, there is a reasonable assumption that the timing of social cues will be appropriate, and so it may not be necessary to include it as part of a behavioural metric for HHI. However, when applied to social robotics, the assumption of appropriate timing no longer applies and so to fully account for learning differences in HRI, timing may need more explicit incorporation into characterisations of social behaviour.

9.2.2 Adaptation of Social Cues

Personalised adaptation of social cues has previously been suggested to be an important factor in HRI (Dautenhahn, 2004; Tapus et al., 2008). Many researchers have sought to adapt elements of the teaching strategy of a robot to humans (Leyzberg et al., 2014; Gordon & Breazeal, 2015; Westlund & Breazeal, 2015), however this strand of research appears to be less concerned with robot social cues and more with optimal teaching approaches in HRI contexts. Adaptation of robot social cues – focusing on the social behaviour/interaction itself, rather than higher-level teaching strategy – remains relatively underexplored in educational contexts.

Both Szafir & Mutlu (2012) and Brown et al. (2013) have used timing of social cues in response to models of human engagement in order to try and 're-engage' learners in an interaction. Differing degrees of success were achieved in the two studies, but the potential of robot social behaviour for positively influencing learning in educational scenarios is clearly demonstrated. However, these models could be pushed further, with adaptive social behaviour throughout the interaction rather than just in response to specific events such as detected dis-engagement. Such an example would be multimodal behavioural alignment, as discussed in Baxter & Belpaeme (2014), or adaptive personalisation as implemented in Coninx et al. (2016). Adaptation of social cues is, like timing of social cues, not explicitly included in the NVI measure, possibly because it can be assumed in HHI.

9.2.3 Relative Importance of Social Cues

One substantial difference between the robot conditions and the human condition is the possibility of using facial expressions. The robotic platform used for the studies was the Aldebaran NAO. This platform has limited ability to generate facial expressions as none of the elements of the face can move, only the eye colour can be changed. On the other hand, the human has a rich set of facial expressions to draw upon.

Whilst the overall NVI scores for the asocial, social and human conditions are tightly bunched, the make-up of the scores is not. For example, the robot scores (asocial and social combined) are higher for gesturing, averaging M=4.3 (95% CI [4.1,4.5]) out of 5 for the nonverbal immediacy question about gesturing (the robot uses its hands and arms to gesture while talking to you), compared to M=3.1 (95% CI [2.7,3.5]) for the human. However, the human is perceived to smile more (M=2.5, 95% CI [2.1,2.8]) than the robot (M=1.8, 95% CI [1.5,2.0]). Through principle component analysis, Wilson & Locker Jr (2007) found that different elements of nonverbal behaviour do not contribute equally to either the nonverbal immediacy construct, nor to instructor effectiveness. Facial expressions (specifically smiles) have a large impact on both the nonverbal immediacy construct and on instructor effectiveness, whereas gestures do not have such a large effect (although still a meaningful contribution; smiles: .54, gestures: .30 component contribution from Wilson & Locker Jr, 2007).

In the NVI metric, all social cues are given equal weighting. However, this may not always be the most appropriate method for combining the cues given the evidence which suggests that some cues may contribute more than others to various outcomes (McCroskey et al., 1996; Wilson & Locker Jr, 2007). This could be a further explanation as to why several of the conditions under examination here have near identical overall NVI scores, but very different learning outcomes. Considering each of the sub-scales (i.e., social cues) independently would also not account for learning, as they should be judged in context (Zaki, 2013). Nonverbal immediacy is designed to cope with characterising social behaviour across a multitude of contexts, and to weight the importance of cues in a metric could become problematic. When robots are used, due to their different morphological features and capabilities, substantial empirical evidence would need to be gathered for each platform to generate accurate weightings. These would then only apply in the context involving that platform, thus reducing much of the versatility offered by relatively

context-free metrics like immediacy.

9.2.4 Novelty of Character and Behaviour

The novelty of both the character (i.e., robot or human) and of the behaviour itself could have had an impact on the learning results found in the study. Novelty is highlighted as a potential issue in both of the chapters in which the robot conditions were presented for the first time (6 and 7), as well as in other experiments conducted in the field (Kanda et al., 2004; Sung et al., 2009). More long-term studies need to be conducted in order reduce and account for novelty effects.

The novelty of the robot behaviour could override the differences between the conditions and subsequently influence the learning of the child. In the social robot condition here, novel behaviour (such as new gestures) was often introduced when providing lessons to the child. Between humans, this would likely result in a positive effect (Goldin-Meadow et al., 2001), but when done by a robot, the novelty of the behaviour may counter-act the intended positive effect.

There may also be a difference in the novelty effect for the children seeing the robot when compared to the human. Although the human is not one that they are familiar with, they are still 'just' a human, whereas the robot is likely to be more exciting and novel as child interaction with robots is more limited than with humans. The additional novelty of the robot could have been a distraction from the learning, explaining why the learning in the human condition is higher. Alternatively, the manipulations to the robot's immediacy behaviour may not be interpreted in the same way as they would be in a typical classroom, which could influence the outcome of higher immediacy behaviours; this has previously been suggested to be a factor in human-human studies of immediacy (Christensen & Menzel, 1998).

Finally, the novelty may have impacted the NVI scores themselves. It is possible that observers (be they children or adults) score immediacy on a relative scale. It is reasonable to suggest that the immediacy of the characters is judged not as a standalone piece of behaviour, but in the context of an observer's prior experience, or expectations for what that character may be capable of. Clear expectations will likely exist for human behaviour, but not for robot behaviour, which may lead to an over-estimation of robot immediacy. This would impact on the ability of considering the human and robots on the same

nonverbal immediacy scale and drawing correlations with learning, and cannot be ruled out as a factor in the results.

9.2.5 (In)congruency of Social Cues

As previously discussed, the robot is limited in the social cues that it can produce (for example, it cannot produce facial expressions). This meant that the conditions all manipulated the available robot social cues, but if social cues are interpreted as a single percept by the human (as suggested by the literature Zaki, 2013), then this could lead to complications.

In the case of the social robot, many social cues are used to try and maximise the 'sociality' of the robot. This means that there is a lot of gaze from the robot to the child and the robot uses a lot of gestures. However, it still cannot produce facial expressions. This incongruency between the social cues could produce an adverse effect in terms of perception on the part of the child and subsequently diminish the learning outcome. There are clear parallels here with the concept of the Uncanny Valley (Mori et al., 2012), with models for the Uncanny Valley based on category boundaries in perception indicating issues arising from these mismatches (Moore, 2012).

The expectation the child has for the robot social behaviour is suggested to be of great importance (Kennedy et al., 2015a). If their expectations are formed early on through high quantities of gaze and gestures, then there would be a discrepancy when facial expressions do not match this expectation. Again, this expectation discrepancy may lead to adverse effects on learning outcomes, as in the case of perceptual issues due to cue incongruence. These issues may become exacerbated as the overall level of sociality of behaviour of the robot increases as any incongruences then become more pronounced. As stated in Richmond et al. (1987), higher immediacy generally leads to more communication, which can create misperceptions (of liking, or expected behaviour).

As robot social behaviour consists of a multitude of social cues, and the congruence between these cues may play an important role in the overall perception of the behaviour, the need to explore the congruency between cues arises. To achieve this, the nonverbal immediacy scale is returned to, as whilst this is calculated on the basis of a combination of many social cues, the cues are rated on an individual basis. The nonverbal immediacy scale has been rigorously validated (McCroskey et al., 1996; Richmond et al., 2003), and it

is known that it provides a reliable metric for the immediacy construct which it attempts to measure in humans (Cronbach's Alpha is typically between 0.70 and 0.85; McCroskey et al., 1996). Chapter 4 of this document provided a validation of the scale adapted by the author for use with robots (adapted scales are shown in Appendices E and F). When used with adult respondents reporting on robot behaviour, it was found that this adapted scale had a comparably high internal consistency rating ($\alpha = .79$) to the original scale used with humans.

Typically, internal consistency measures of a scale, like Cronbach's alpha, would be used to evaluate the ability of a scale to measure a unidimensional construct. With the assumption (evidenced through the high internal consistency measures from experimental work) that the NVI scale is indeed unidimensional, it becomes feasible to consider that if a robot behaviour is rated and the internal consistency value is low, this is not a reflection on the scale (which is known to be reliable), but instead on the behaviour. Through the nature of the robot's capabilities there are several NVI scale items which are known to be impaired. For example, the robot cannot produce facial expressions. If all of the other social cues that form the scale are manipulated such that they reflect 'high' NVI, then the lack of facial expressions will cause the internal consistency measure to drop. Based on this logic, it is suggested that internal consistency measures could be used as an indicator of the incongruence between social cues within the immediacy scale. The greater the internal consistency, the more congruent the social cues could be, but a lower internal consistency indicates larger variability between scale items, and thus possibly larger incongruency between the cues.

Guttman's λ_6 (or G6) for each condition has been calculated[2], revealing that indeed there are differences in how congruent the cues could be considered to be (Table 9.1; Figure 9.4). The behavioural conditions used here are restricted in such a way that a lower reliability would be expected (as several cues of the scale are not utilised) for some conditions. Indeed, these values fall in line with predictions that could be made based on the social behaviour in each of the conditions. The human reliability score provides a "sanity check"

[2]Cronbach's alpha tends to be the de facto standard for evaluating internal consistency and reliability, however its use as such a measure has been called into question (Revelle & Zinbarg, 2009) – including by its own creator (Cronbach & Shavelson, 2004). Instead G6 is used which considers the amount of variance in each item that can be accounted for by the linear regression of all other items (the squared multiple correlation; Guttman, 1945). Like Cronbach's alpha, G6 normally ranges from 0 to 1, with values closer to 1 indicating higher internal consistency reliability. This provides a lower bound for item communality, becoming a better estimate with increased numbers of items. This would appear to provide a logical (but likely imperfect) indicator for the congruency of cues as required here.

Condition	Learning Effect Size (Cohen's *d*)	Guttman's Lambda 6 (G6)
Asocial robot (Chapter 6)	0.89	0.84
Social robot (Chapter 6)	0.51	0.83
High NVI robot (Chapter 7)	0.67	0.69
Low NVI robot (Chapter 7)	0.30	0.78
Human (Chapter 9)	0.89	0.87

Table 9.1: Guttman's G6 and learning effect size by condition. G6 is used as an indicator of social cue congruency, with a higher value indicating greater congruency between cues.

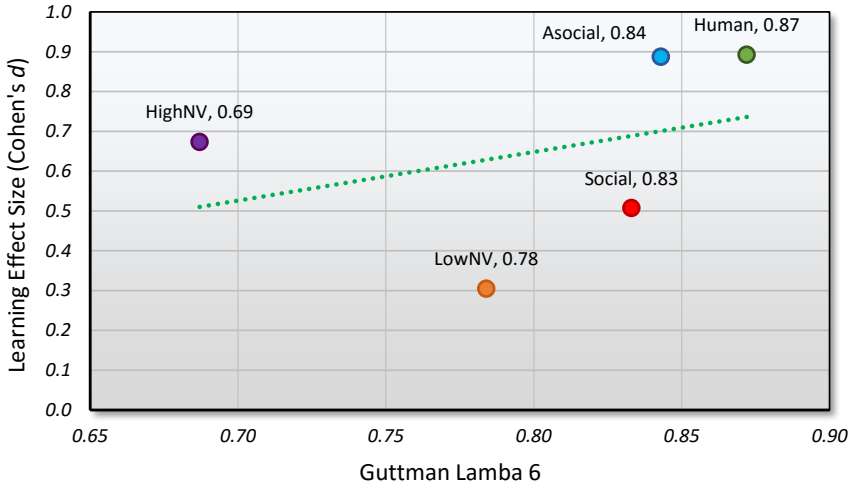

Figure 9.4: Guttman's Lambda 6 against learning effect size for each of the prime tutoring conditions. The dotted line indicates a trend towards greater internal consistency (measured through G6) leading to greater learning.

as it is assumed that human behaviour would have a certain degree of internal consistency between social cues, which is reflected by it having the highest value. Additionally, the low NVI robot condition has intentionally low NVI behaviour, so the lack of smiling or touching (high NVI behaviours) does not cause incongruency (signified by a lower G6 score), whereas the high NVI robot condition has intentionally high NVI behaviour where possible on the robot, so the lack of smiling and touching causes greater overall incongruency, resulting in a considerably lower G6 score.

A Proposed Model Taking Guttman's Lambda 6 to provide a reasonable indication of the congruency of social cues, then it is clear that this alone would not provide a strong predictor of learning (Figure 9.4). However, this data can be combined with the social behaviour (as measured through NVI) to be compared to learning outcomes. In the resulting space, both congruency and social behaviour could have an impact on learning, as hypothesised in the previous section (Figure 9.5).

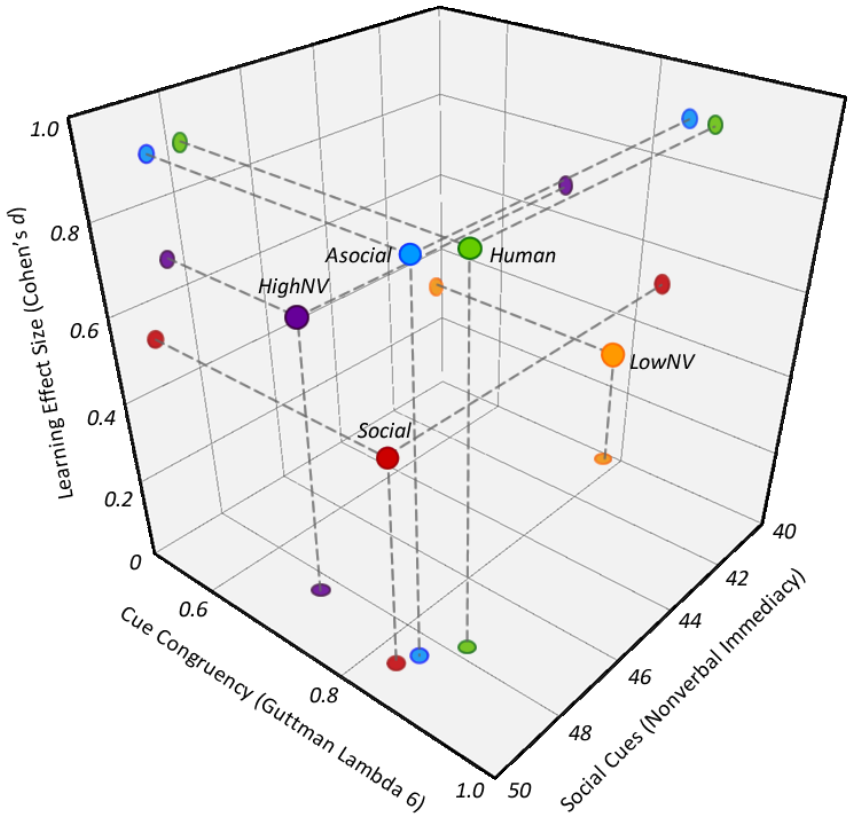

Figure 9.5: Learning, congruency and social behaviour for each of the 5 conditions. Learning is measured in effect size between pre- and post-test for children. Congruency is indicated through Guttman's Lambda 6 of the adult nonverbal immediacy scores. Social behaviour is characterised through nonverbal immediacy ratings from adults. An interactive version of this figure is available online to provide different perspectives of the space: `https://goo.gl/ZNPxc8`.

The data shows that learning is best with human behaviour which is shown to be highly social and congruent. When the social behaviour used is congruent, but not highly social, then the learning drops to a low level. With roughly congruent social behaviour as characterised by nonverbal immediacy (social, asocial, and human conditions), when the congruency of the cues increases (indicated by Guttman's G6), learning also increases. The combination of congruency and social behaviour as characterised by nonverbal immediacy appears to provide a reasonable model for learning predictions, where the combination of high social behaviour and social cue congruency is necessary to maximise potential learning.

Such a model is supported by the view of social cues being perceived as a single percept, as suggested by Zaki (2013). Experimental evidence with perception of emotions would seem to provide additional weight to such a perspective (Nook et al., 2015). This has clear implications for designers of social robot behaviour when human perceptions or outcomes are of any degree of importance. The combination of all social cues in context must be

considered alongside the expectations of the human in order to generate appropriate behaviour. Not only does this give rise to a number of challenges, such as identifying combinatorial contextual expectations for social cues, but it could also have implications for how social cues should be examined experimentally. The isolation of specific social cues in experimental scenarios would not describe the role of that social cue, but the role of that social cue, *given the context of all other cues*. This is an important distinction that leads to a great deal more complexity in 'solving' behavioural design for social robots, but that would also contribute to explanations of why a complex picture is emerging in terms of the effect of robot behaviour on learning, as discussed in Chapter 2.

This model lends itself to generating predictions concerning social behaviour, social cues and learning, as well as providing retrospective characterisation of social behaviour. The following predictions can be derived from the extremities of the space that is presented:

P1. Highly social behaviour of a tutor robot (as characterised by nonverbal immediacy) with high congruency will lead to maximum potential learning.

P2. Low social behaviour of a tutor robot with low congruency will lead to minimal potential learning.

P3. A mismatch in the social behaviour of a tutor robot and the social cue congruency will lead to less than maximum potential learning.

An accurate, or reliable measure for social cue congruency would be desirable in order to properly measure or characterise the social cue congruency dimension (rather than the indicator used here). However, this would not necessarily be something that would be straightforward to achieve due to the potentially complex interactions between large numbers of social cues. Additionally, the data used for the learning axis was collected with relatively few samples (just over 10 per condition) in a specific experimental setup. Ideally, many further samples would be collected in both the short and long-term. The data collected here is over the short term and with children unfamiliar with robots. As longer-term interactions take place, or as robots become more commonplace in society, expectations may change and the model will need revision to account for this. Nevertheless, these predictions support the thesis contained in this document: 'a robot with tailored social behaviour will positively influence the outcomes of tutoring interactions with children and consequently lead to an increase in child learning when compared to a

robot without this social behaviour'.

9.3 Summary

This chapter brought together the findings from the previous chapters in the context of the thesis for this work: that tailored social behaviour of a robot will lead to better child learning. An additional experimental condition was introduced to provide a benchmark and further context for the child learning, and behavioural ratings of the robot conditions. Nonverbal immediacy was used to tie the findings together in support of the thesis. The relationship between child learning and robot social behaviour was then discussed in a broader context, drawing on other literature as well as the data from the work conducted here. A model of the relationship between social behaviour and learning was proposed, with the intention of providing a working hypothesis for future research. Robot social cues and the congruency between the cues are suggested to have a combined effect on learning.

Chapter 10

Discussion

Chapters 3 through 9 presented the experimental approach and evaluations undertaken in seeking to support the thesis for this work:

> *A robot with tailored social behaviour will positively influence the outcomes of tutoring interactions with children and consequently lead to an increase in child learning when compared to a robot without this social behaviour.*

This chapter will present a discussion that draws on the evaluations conducted in the previous chapters and their findings as well as some of the issues introduced in the early chapters of this document. Limitations in the approach adopted here will be outlined, and the discussion will be broadened to show how the work conducted here fits into the larger field of social HRI. Specifically, societal acceptance[1] and ethical considerations will be explored. The chapter will conclude with suggestions for further research that could build on the findings from this thesis.

10.1 Experimental Limitations

This section seeks to acknowledge the limitations in the experimental work conducted throughout this thesis. These limitations stem from various aspects: some theoretical, some practical. These factors must be considered when interpreting the outcomes of this thesis, but they are also used to highlight areas where future work could be performed to further support the findings here (elaborated on in Section 10.6).

[1] some of this section has been published by the author in (Kennedy et al., 2016c)

10.1.1 Ecological Validity and Generalisability

Several aspects of the experimental design limit the potential generalisability of the findings from this thesis. The first of these limitations is due to a necessary compromise between experimental control and perfect ecological validity (Ros et al., 2011). The experiments are all conducted "in the wild", i.e., in the environment of children (their schools). However, the experiments take place outside of the children's normal classroom and instead in other classrooms, or communal spaces. This was done to achieve a certain degree of experimental control (such as preventing other children from overhearing material and therefore being exposed to the material multiple times). This decision may also have led to different behaviour, and subsequently learning outcomes, from the children than would have been observed had the robot been present in the normal classroom with other children. Whilst this design choice may have inhibited the confidence of generalisability of results, it is useful to have evidence for effects in relatively controlled environments so that expectations can be formed for classroom use, and so that robot behaviours can be iteratively designed for this environment.

The subject numbers are relatively low for studies which consider human psychology. Typically, a psychology study aims for at least 30 participants to be considered 'large', although this approach has been the subject of criticism as it over-simplifies a complex decision (Kar & Ramalingam, 2013). A number of studies undertaken here have around 10-15 subjects per condition (Chapters 5, 6, 7 and 9), which is fairly typical for child-robot interaction studies in the educational domain (Leite et al., 2014; Saerbeck et al., 2010; Zaga et al., 2015). Nonetheless, from a psychology perspective, this would mean that they would be judged 'small' studies, and as such, confidence in the accuracy of the statistical analysis would be reduced. To this end, descriptive statistics, such as confidence intervals, have been included for all studies within this document so that the results do not rely purely on significance testing (as recommended in Baxter et al., 2016).

However, the context of the study should also be considered when deciding on and reviewing sample sizes. Throughout this work, sample sizes have typically been determined by school class sizes. In the U.K., classes range from around 20 to 30 students, so this is used as the total number of participants in many of the studies. Variation between schools, and indeed different classes within schools has the potential to introduce confounds. These confounds can be not just prior knowledge, which could be controlled for,

but exposure to different learning styles, or teachers influencing expectations in different ways. Teachers often used interacting with the robot as an incentive to aid classroom management, for example "be quiet, or you won't get to play with the robot". Through doing this, the teacher gives importance to the robot and creates expectations which can shape the interaction (Fischer et al., 2014). Using a single class as a sample means that the children's expectations will have been shaped in a constant way by a single teacher, thereby providing tighter experimental control in spite of lower subject numbers (Baxter et al., 2015).

All of the experiments conducted as part of this thesis involved children who had not previously been exposed to the NAO robot. This was an intentional design decision, to ensure that children's expectations would not have been shaped by prior interactions. This allows a comparison between different conditions without having to consider how previous interactions may have influenced behavioural responses. However, longitudinal studies whereby the children were exposed to the same robot multiple times were not conducted. It is entirely possible that many of the observations were due to a novelty effect, where the children were interested in the robot due to its novelty, as opposed to its behaviour, and subsequently responded in a manner that would not be observed in the long-term. Such effects have been suggested to be responsible for observations in other studies, for example Kanda et al. (2004); Sung et al. (2009). The methodology adopted in Chapter 8 re-tested children on their learning some time after the interaction, thereby verifying that the learning was lasting. The work here goes some way to providing design principles for robot behaviour in educational interactions, but exploration of whether learning effects persist over longer-term interactions with this behaviour remains to be explored. In any case, the evidence here suggests that short-term interactions do cause learning, which could be exploited regardless of whether learning gains are observed in prolonged interactions.

10.1.2 Measures of Learning

As stated in the background for this work (Chapter 2), when measuring a child's application of acquired skills, the measurement may not be of their learning, but their performance in the task (Mikulas, 1977). These are complex concepts to disentangle, and so no distinction was made in the experimental work undertaken here. Consequently,

improvements in task performance may not have been due to learning, but instead because of greater motivation, or children becoming more comfortable over the course of an interaction. However, these factors are constant between conditions, and control conditions with no learning content are also often used in the experiments here. This provides some confidence that what is being measured is in fact child learning.

The measures of child learning focussed on here were cognitive learning gains. This is in part due to the easy collection and good reliability of such data. It is comparatively straightforward to administer pre-tests and post-tests, and to compare scores, as opposed to attempting to collect reliable data on affective learning from children, or to monitor the underlying learning processes of children. Many affective learning measures involve surveys with challenging or irrelevant questions for children of this age, e.g., "likelihood of taking future courses in this area" (McCroskey et al., 1994). Whilst these affective aspects of learning were not within the scope of the research programme undertaken here, it would nonetheless be worthwhile for future research to pursue this topic to broaden studies on the relationship between robot behaviour and child learning.

10.1.3 Robot Platform

The Aldebaran NAO is the only robot platform used in the research conducted throughout this thesis. Whilst the use of just one robotic platform eases comparisons between studies and allows for iterative development of behaviours, it also limits the findings. It is unclear as to whether the observations made here would translate to robots of a different size, appearance, or morphology. In this regard, the results could be somewhat platform specific. However, the findings here do seem to agree with research being conducted in the same domain with different robots.

The NAO robot does not provide the ability to manipulate some of the social cues under consideration in the NVI measure. For example, the robot cannot produce facial expressions, so it cannot smile. This does not mean that people will not perceive the robot to smile, but it does prevent explicit manipulation of this cue. It may be that greater learning effects would be observed if a robot capable of producing facial expressions were to be used, due to the relationship between smiling and learning (Wilson & Locker Jr, 2007), and robot facial expressions and child enjoyment of interactions (Cameron et al., 2015). The model proposed in Chapter 9 proposes that the congruency of social cues is important in the

learning outcomes. As such, it would be preferable to work with a platform capable of social cue manipulation along all items of the nonverbal immediacy measure. Currently, there are a limited number of commercially available robotic platforms that would offer all of the necessary affordances to achieve this. When considering robots that would be an appropriate size to use with children, the NAO remains a strong choice despite the lack of ability to produces facial expressions. This may change as an increasing number of robot manufacturers start to use screens as faces, preventing complications with actuating a face in a confined space.

The NAO robot also has relatively loud motors. This could be a factor in some of the findings throughout the thesis, particularly when considering social responses as a product of embodiment. In Chapter 5, a physical robot was compared with a virtual one on a screen. When the physical robot moves, the motors could attract attention due to the noise, which would not be the case for the virtual robot. Consequently, children may have looked at the robot more due to the noise (although this also could have been due to the forward motion of the physical robot, that would not be present with the virtual on-screen robot). The motor noise could also have played a role in Chapter 7, where the robot in one condition moved more than the other condition. If a robot with less motor noise were to be used, it may be that some of the effects between embodiments are no longer as prominent. On the other hand, learning may actually increase. The robot would often be programmed to perform gestures whilst explaining topics (in an aim to be appropriately social), but the motor noise that this introduces may hinder the children's understanding of the verbal content. Without the motor noise, child understanding, and subsequent learning, may improve.

Due to the current wide availability and common use of the Aldebaran NAO, the findings in this thesis are still directly relevant to many other researchers regardless of whether the results generalise to other platforms. The immediacy construct and measurements also provide a framework enabling other researchers to compare other platforms and behaviours against the results found here. Investigation of social behaviour implemented on other platforms characterised through immediacy would certainly provide an interesting avenue for future work.

10.2 Ethical Questions

Chapter 3 introduced some issues of ethics from a practical perspective. Of course, ethical issues relate not just to practical execution of experiments, but to a wider set of considerations pertaining to whether robots are suitable for use in education at all. Serholt et al. (2016) highlight four areas of concern following focus groups with 77 educators in three countries: child privacy, robot responsibility, long-term impact on the child, and accountability of robot actions. These four areas will therefore form the framework for the discussion in this section.

Privacy is recognised as a major ethical concern in relation to robots; partially because robots are commonly used to actively perform surveillance tasks (Calo, 2011). These concerns change in focus when robots become social companions. Lee et al. (2011) suggest that this is due in part to the tendency of humans to anthropomorphise, and the increased likelihood of robots being shared when compared to other technology. Anthropomorphism could cause humans to reveal more information, whilst the shared aspect presents greater opportunity for the spread of the information. People are not fully comfortable with revealing personal information to a robot that it may store, but do see it as a "necessary evil" provided that there is a benefit to the user (Syrdal et al., 2007b). A reluctance to reveal personal information due to perceptions of sociality could make social robots unsuitable for some applications where the information is critical in forming an appropriate response, or where perceptions of being judged socially may lead to negative behaviours (and indeed where less-social agents may bring about benefits; Howley et al., 2014).

When applied to social robots in education, it is noted by Serholt et al. (2016), that child privacy is a wider concern in education already, with teachers collecting a variety of data about children without their permission. Indeed, the U.K. Department for Education's suggested privacy notice text (Department for Education, 2016) includes collecting pupil data to: support learning, monitor and report progress, provide pastoral care, and assess the quality of teaching services. This can include personal characteristics, as well as learning data. Children are not given a choice as to whether they wish to opt out of this data collection. Robots are unlikely to extract any more information than this, however the data is (currently) more likely to be sent to a third party, whereas data at present is kept internally within schools. Striking a balance between privacy and the robot having

enough information to be effective, along with whom may access this data, will be an important step in establishing how such technology can be used in an ethically responsible manner in schools.

The role, and associated expectations and responsibilities of robots in schools must be carefully managed. There have long been calls for education to keep up with the pace of technological development (arguably before some of the more impactful technological advances had become commonplace in society; Will, 1986). However, there is a distinction between delivering content that educates about technology, and technology that delivers other learning content, although these aspects could of course be intertwined. When technology delivers content, there is a risk that supervising humans do not thoroughly understand the limitations, or the limitations are not immediately apparent. As a consequence, there is a greater chance that it will be used in an unintended manner, with negative cognitive or affective outcomes. While this is in part the fault of the human who employed the technology in that role, the purveyors of the technology must also take some responsibility if limitations are not made clear (Johnson & Powers, 2005). It is argued that the temptation to consider technology as "natural" must be resisted to prevent undesired moral actions being taken with technology (Johnson & Powers, 2005). Clearly in the case of social robots, this temptation is exacerbated due to the precise aim of making the technology more natural in some manner (Breazeal, 2002), combined with the human tendency to anthropomorphise. Limitations of social robots for education should therefore be well documented and portrayed such that the technology is used appropriately.

As part of using social robots in an appropriate manner and to portray limitations accurately, it is necessary to establish the effects not only in terms of learning (as explored in this thesis), but also in terms of social well-being. This should be done both at an individual level, and also at a societal level; current decisions about technology can have an impact on future generations, and these temporally ethical decisions must be respected (Groves, 2006). Future research should strive to identify and address long-term risks and benefits of exposure to social robots in general, as well as the case of educative robots in specific. Inspiration can be taken from research of current pervasive technologies such as mobile phones (Kamibeppu & Sugiura, 2005) and the Internet (Beranuy et al., 2009), but the social aspects and implications need to be explored further due to the social nature of the robots being developed.

The accountability of robot actions is part of an ongoing larger ethical and legal debate surrounding machine autonomy (currently often related to autonomous vehicles as they begin to reach the mass-market; Beiker, 2012). Questions of whether the programmer, or the user, are at fault for negative outcomes are often debated, with further complexities in cases where learning algorithms are employed (Ros et al., 2011). It has been suggested that legal frameworks will need to be created on an application-by-application basis (Bertolini, 2013), which presents an opportunity for research into educative social robotics to contribute to the legal debate. Decisions made in this regard will no doubt shape the future of what research may be possible or encouraged.

10.3 Educator and Societal Acceptability

Whilst the work conducted here was carried out with the support of schools, there are broader questions of whether robots in education are acceptable to teachers, and society at large. This section seeks to explore some of these areas, also highlighting how the work conducted here might contribute to these larger issues.

As this field of research pushes forwards, and if we seek further real-world or mass-market implementation in schools, it is important to understand attitudes towards the technology. For successful adoption of such technologies, it is necessary for both teachers and the general public to be willing participants in increased uptake. Recent findings from the Eurobarometer report (European Commission, 2012) have suggested that whilst there is generally a positive view towards robots in Europe, there is a sizeable contingent (34%) that would see robots banned from use in education. However, the survey administered in this report does not provide a context for many of the questions, calling into contention how well-understood or known the kind of social robots that would be used in this domain were for the respondents.

Research has suggested that there are barriers to adoption and use of technology by teachers. These can be first-order (extrinsic) barriers, or second-order (personal) barriers. While the extrinsic barriers cannot be discounted, it has been found that positive beliefs of teachers about the effectiveness for learning (i.e., personal factors) are a significant predictor of actual technology use (Blackwell et al., 2013). For this reason, it is important to understand (and possibly influence) how teachers feel towards social robots if we intend to see them widely adopted.

Previous pan-European work (Serholt et al., 2014a) found that views of teachers are generally positive, but that there are concerns over fairness to access, the robustness of the technology, and potential disruption to classrooms. Some of these same concerns were observed prior to an experiment in the USA, but after the experiment had been completed, views had changed (Kory Westlund et al., 2016). Teachers expected the robot to be disruptive to the classroom, but found that it was not, although this is partially mitigated as headphones were used so that the possibility of audible disruption would be minimised. A large-scale survey conducted in South Korea (Lee et al., 2008) found that teachers were generally positive about the use of robots in education, but they were more negative than other stakeholders.

When exposed to a highly scripted interaction with a robot, teachers showed fairly positive reactions (Fridin & Belokopytov, 2014), however it was concluded that the interaction here was not related to the educational quality that the robot could offer, and this is where the focus should be. Incorporating the views of teachers in educational technology design has been highlighted as a particularly important aspect of creating a partnership that allows teachers to identify the benefits and shortcomings of technology when related to the curriculum (Okita & Jamalian, 2011).

Revisiting Serholt et al. (2016), the long-term impact of educative robots on children was raised as one of four central issues following focus groups with 77 educators in three countries. Long-term consideration of children's welfare has been raised as an issue in other studies as well. Kennedy et al. (2016c) finds that educators concerns about appropriate long-term social skills for the robots dominate over practical and ethical concerns. These social skills involve the richness of the interaction, the adaptability of the robots to change in response to child behaviours, or the suitability of social robots to develop children's peer-group sociality, all of which present fundamental questions for research in this field. It is therefore suggested that these behavioural considerations must remain central to the research agenda of child-robot interaction (Kennedy et al., 2016c), but that we progress bearing in mind the responsibility we have to children (Serholt et al., 2016).

Due to the technological nature of robots, it is reasonable to hypothesise that they will be seen as a tool for STEM education, rather than for the teaching of humanities. This is reflected in the research being conducted with robots in education: they are commonly

applied in STEM education, with promising outcomes Karim et al. (2015), although research is also prominent in language contexts Hood et al. (2015); Kanda et al. (2004); Kennedy et al. (2016b); Tanaka & Matsuzoe (2012). However, there are comparatively few robots being used to teach art or religious education, for instance (a reference to work in either of these domains could not be identified at the time of writing). Some perceptions based on pre-conceptions may well change with greater exposure to social robots that can do more than be used as a tool for STEM subjects (for example, as recently shown with handwriting learning; Hood et al., 2015). This is potentially where the broader aspects of using a social robot could be beneficial in breaking down some barriers to use. The robot is a technological device, but could be used to teach a variety of subjects with an element of sociality. The use of the robot could stimulate interest in technology, and the social aspects of robot behaviour could be used to create reciprocal interest in those subjects (as has been attempted for some aspects of behaviour; Gordon et al., 2015). This calls for a greater exposure of teachers to our robotic systems, so that they better comprehend the capabilities, current limited performance, and possible future applications of social robots in education. The efficacy of such applications must also be clearly demonstrated.

Successfully addressing the concerns highlighted by educators in the various studies would provide an essential first step towards achieving greater adoption (if indeed this is desirable). Some of the concerns may arguably be alleviated once the teachers and the children familiarise themselves with the robots (the robot being a source of distraction is likely to resolve quickly after novelty goes away) or once the penetration of robots in classrooms increases to a point where dedicated companies could regularly take over training and maintenance issues. Overall, the attitude towards social robots in schools is potentially accepting (Lee et al., 2008; Serholt et al., 2014a). For the educators, concerns about appropriate social skills for the robots are raised in addition to the practical and ethical concerns. The work in this thesis contributes in some part to addressing concerns regarding social capabilities of such robots, but this should remain a focus for child-robot interaction research as more work is required to explore longer-term impacts on children interacting regularly with social robots.

10.4 Sociality and Learning

The thesis put forward here centres around *social behaviour* and *learning*, and in particular, the connection between these concepts. This section seeks to critically discuss the approach to social behaviour and learning in the work carried out, as well as to consider the three predictions generated from the model created in Section 9.2 in a broader context.

As identified in the background (Chapter 2), acquiring data from children regarding their subjective perceptions of an interaction and robot behaviour is particularly challenging. To minimise the problems in acquiring such data, nonverbal immediacy was used, which considers only overt social cues. Chapter 4 demonstrated the validity of this approach, but it is still not without limitations, as highlighted in Chapter 9.

While nonverbal immediacy provides a characterisation of robot behaviour as perceived by children, it does not describe the relationship they feel towards the robot, or their perception of the social aspects of the interaction as a whole. In the studies conducted here, additional questions were commonly added to the end of questionnaires that were administered, but these gave an indication of the relationship felt by the children, rather than exploring it in detail. For example, "For me, I think the robot was like a - ...", as seen in Appendix I. Greater nonverbal immediacy has been correlated with increased friendliness (Wilson & Locker Jr, 2007), but children did not (statistically) consistently report robot conditions with higher nonverbal immediacy to be more like a friend. Of course, increased friendliness would not necessarily lead to children relating to the robot as a 'friend', but given the limited data collected here this would provide the best indicator. It would be desirable to explore the relationship between the child and the robot in greater detail to better understand the interaction from the child's perspective. This could reveal areas for further improvement in the robot social behaviour. However, this is an extremely challenging task, with no straightforward solution to implement when conducting studies with reasonably high numbers of children. Structured interviews could be valuable, but are incredibly time consuming when dealing with tens of young subjects, who are likely nearing their attention limit after interacting with the robot.

Specific aspects of learning were the focus of the research here. Figure 10.1 has been reproduced from Chapter 2, which indicates where the focus of learning was. It is in the cognitive domain, and it is clear that a large area of the taxonomy is not addressed. Nor

is the affective domain addressed – that is not to say it was not influenced in the work here, but measuring and accounting for any change in this domain was not within the scope of the work (Section 1.1). This is not uncommon in the field of HRI research, with many studies exploring 'remember' (Alemi et al., 2014; Szafir & Mutlu, 2012), 'understand' (Kory & Breazeal, 2014; Tanaka & Matsuzoe, 2012), or 'apply' (Leyzberg et al., 2014) cognitive processes. Relatively few studies consider meta-cognitive analysis, evaluation, or creation. This may in part be due to the additional complexity in scenario and knowledge interpretation/representation that would be required either automatically from robot sensors, or to be input by a Wizard.

		Cognitive Process dimension					
		Remember	Understand	Apply	Analyse	Evaluate	Create
Knowledge dimension	Factual	X	X	X			
	Conceptual	X	X	X			
	Procedural	X	X	X			
	Meta-cognitive						

Figure 10.1: The revised educational objectives 'Taxonomy Table' (adapted from Krathwohl, 2002). Crosses indicate the areas focused on in studies throughout the research here, with the red cross signifying the intersection at which performance is most often measured. Reproduced from Chapter 2.

In Chapter 9, a model was proposed that made three predictions for the relationship between sociality and learning. To explore these further, other literature will be considered in the context of the predictions, which will also tie back to the background from Chapter 2 (specifically Section 2.7), where a mixed picture was discovered in the relationship between social behaviour and learning. As a reminder, the predictions of the model were as follows:

P1. Highly social behaviour of a tutor robot (as characterised by nonverbal immediacy) with high congruency will lead to maximum potential learning.

P2. Low social behaviour of a tutor robot with low congruency will lead to minimal potential learning.

P3. A mismatch in the social behaviour of a tutor robot and the social cue congruency will lead to less than maximum potential learning.

Prediction 1 suggests that highly social behaviour with high congruency will lead to maximum potential learning, and prediction 2 is the inverse of this. The former could

be demonstrated through Saerbeck et al. (2010), where a number of social cues were modified in unison, including gestures, verbal utterances and emotional expressions. As such, it could be suggested that the cues are both highly social and congruent, explaining the improvement in learning that is shown when compared to a robot without these behaviours. Herberg et al. (2015) may have experienced unexpected results because the robot did not employ many social cues (there are no gestures, body posture changes, or facial expressions, for example), but then one condition has more animated gaze than the other condition. Through adding this gaze, the cues are not only minimally social, but now also incongruent (as one cue is social whilst the others are not utilised), leading to less learning than the robot without this additional gaze cue (which therefore has low social cue use, but with more congruency), demonstrating prediction 2. Prediction 3 could explain the results of Kanda et al. (2012), where increased social behaviour was operationalised only through verbal utterances and no learning differences were observed. Although the addition of verbal utterances increased the social behaviour, it simultaneously reduced the congruency (as no other cues were made more social), leading to no change in learning.

It is acknowledged that these predictions are made based on the framework of the data generated from this thesis, which as this section has discussed, only considers a sub-set of the expanse of how both learning and social behaviour may be defined. This does not necessarily detract from the model, or the possible validity of the predictions, but instead highlights the current challenges in this relatively new domain and the need for further exploration. Fully understanding how to teach using machines is a long-studied, and ongoing field of research, with the recent addition of robots introducing additional unanswered questions (Timms, 2016). Characterising and understanding social behaviour is also a complex task, with research in neuroscience and psychology still not providing an account for these aspects. It is becoming recognised that models of unified behaviour may be required (Zaki, 2013), with a broad set of social behaviours incorporated (including affective components alongside social cues and social-cognitive functions; Pfeiffer et al., 2013). Restricting the focus of learning and social behaviour therefore becomes a more tractable approach to establish models to build from given the quantity of unknowns in the exploration. However, there is clearly a need to build increasingly comprehensive models for generating and understanding social behaviour for robots once more basic principles have been rigorously established; this is highlighted as future work in the subsequent section.

10.5 Comparisons Between Children and Adults

Much of the work conducted in this thesis has been influenced by findings from human-human literature with adults, in particular the use of immediacy, which is commonly applied in university lecture contexts (Witt et al., 2004). Additionally, the quantity of similar work focussing on learning and social behaviour in the field of human-robot interaction applied to child tutoring, while increasing, is not sufficient to provide a thorough background for the work here. As such, comparisons were often made between results found with children and results found with adults, and experiments would be explicitly designed in a manner that attempted to tie findings between human adults to HRI scenarios with children (for example, Chapter 4). This section seeks to draw on this experience to discuss the merits of such comparisons.

When attempting to apply the concepts and adapt the measures used in nonverbal immediacy for use with children, many issues arose in terms of the language used in the original questionnaires as they had been designed for adults. This was an even greater problem in other questionnaire series, with common use of abstract terms, e.g., Bartneck et al. (2009b). This lead to an adaptation of nonverbal immediacy being used. Challenges then arise because the metric is no longer the one previously validated with adults, and children are known to interpret and complete questionnaires in a different manner to adults in any case (Belpaeme et al., 2013; Borgers et al., 2000, 2004). For this reason, adults were used as a baseline in the initial validation of the adapted scale performed in Chapter 4. However, when comparing the child and adult perceptions according to the scale, whilst the ranking was the same and the correlation was strong, the children were clearly far more tightly bunched in their scores. This could be due to a tendency for children to score highly on questionnaires, or it could be a genuine difference in perceptions. This would not be surprising, given that children are still establishing social behavioural norms and learning to process social information (Bandura & Walters, 1963). Indeed, it may be the case that children focus on different social features to adults while they develop concepts of social behaviour from salient cultural practices (Whiting et al., 1992). For this reason, the comparison between child and adult perceptions may not be particularly worthwhile. Although the adult results provide some context, they do not necessarily help to understand the perspective of the child.

Due to the relative lack of other research in the field of social HRI exploring robots for tutoring, many hypotheses and experiments were derived from results acquired with adults. As an example, Chapter 5 sought to address issues of embodiment and drew on findings from Bartneck (2003) and Leyzberg et al. (2012) to motivate the hypothesis that a physically embodied robot would lead to greater child learning. This was not found to be the case, in agreement with other work conducted with children using a similar methodology (the same robot on a screen vs. physically present; Looije et al., 2012). Consequently, the findings appear to be unclear: they are in agreement with some work, but not others. It could be a product of subject age that this problem occurs; if only the child results are considered, then it may simply be the case that the embodiment effect is not present. Comparing to the adult results may serve to confuse matters in this instance.

Furthermore, there are limited resources for use with children in terms of holistic metrics for social behaviour, an issue that is exacerbated when looking specifically for measures that might work with robots. There are currently also limited publications in the field of HRI using children as subjects (although recently this appears to be changing; Mutlu & Tscheligi, 2017). This creates a challenge when attempting to base experimental work in the context of prior literature. It is suggested here that when the subject group changes from children to adults, comparisons with prior literature has limited utility as age will always serve as a confound. As such, it would seem prudent for those working with children to focus on developing more tools and findings specifically for this age group (or possibly age groups, splitting by developmental stage). Such an approach would provide a more satisfactory context for future work and possibly assist in creating a more coherent body of literature, an issue highlighted in Chapter 2.

10.6 Future Work

The work conducted in this thesis has clearly highlighted the potential for the use of robots as tutors for children, and the educational value of tailoring robot social behaviour with the aim of improving learning outcomes. It should be made clear, however, that there remain many open questions for such applications, either because they fell outside the scope of the research conducted here, or because of current limitations in social robotics. This section picks up from many threads of the discussion above, and from Chapter 9, outlining possible future research directions that may provide further value and insight

for the design of robot tutors for children.

10.6.1 Building on the Social Cue and Congruency Model

Chapter 9 proposed a model for the explanation of child learning as a product of social cues and cue congruency. Future work could seek to explore this model and extend or revise it where necessary. The work in this thesis used nonverbal immediacy as the characterisation of social cues for the model. As discussed in Chapter 9, the indicator of cue congruency, Guttman's G6 across the nonverbal immediacy items is likely an imperfect measure for this purpose. Ideally, to build on the model, a metric for social cue congruency would be developed and validated to provide more confidence in the evaluations for this dimension. It may also be necessary to modify or use an alternate scale for the characterisation of social behaviour as measuring only social cue use might not be sufficient given that the timing of the cues is not evaluated in NVI (as highlighted in Chapter 9).

Regardless of whether the dimensions on the scale are modified or not, the model put forwards in this thesis provides a set of three testable predictions, reproduced below. It would be straightforward to create a set of robot behaviours for each one of these predictions, for example, a robot with high social cue use, and high congruency for P1, and so on. These could then be applied in a variety of learning contexts to validate whether the predictions hold true across contexts given consistent behaviour. The same could be done with a variety of robot hardware platforms to examine the impact of certain cues not being available. Collection of a variety of data in this regard would provide more data points for the proposed model space, and could be used to validate, or revise and improve the model.

P1. Highly social behaviour of a tutor robot (as characterised by nonverbal immediacy) with high congruency will lead to maximum potential learning.

P2. Low social behaviour of a tutor robot with low congruency will lead to minimal potential learning.

P3. A mismatch in the social behaviour of a tutor robot and the social cue congruency will lead to less than maximum potential learning.

10.6.2 Accounting for Individual Differences

Whilst the robot behaviour in the interactions that took place throughout the experimental studies here were often adaptive to the children in some way (for example, behaviours tracking child gaze, or occurring in response to children moving images on the touchscreen), incorporating greater adaptation into robot social behaviour may lead to further educational, and interaction-quality, advantages. This principle has been demonstrated at a task level (Coninx et al., 2016), with work in preparation suggesting that the same could apply to lower level social behaviours (Baxter et al., undergoing revisions). However, in order to generate appropriately adaptive behaviours, we need to establish not only a model of child social (and learning) responses to robot social behaviour, but to also characterise the child through some means in order to inform any adaptation that takes place. This could be done on-line through adapting to the child's social behaviour, or off-line through responses to personality questionnaires or similar. This is being explored by other researchers in terms of modelling and generating behaviour in response to child knowledge (Jones et al., 2015), affective states (Spaulding et al., 2016), and behavioural traits (Baxter et al., undergoing revisions). However, work from neuroscience suggests it may be necessary to draw more of these elements together into larger models of social cognition (Pfeiffer et al., 2013) due to the unified nature of social processing (Zaki, 2013). This is clearly an ambitious path of research due to the scale of the problem, but it presents an opportunity for HRI to contribute to our understanding of human psychology.

10.6.3 Increasing Interactivity

The interactions taking place in this research did not typically involve large elements of verbal interaction. This was due to current limitations in child speech recognition (Kennedy et al., 2014b). It is challenging to use speech recognition with children in a robust manner; often this can be circumvented through using a 'Wizard' to type in speech, or to perform some part of the cognitive processing to do with language on behalf of the robot (Baxter et al., 2016). However, this technique can introduce methodological variations and does not accurately reflect current robot abilities, so was not adopted here. Clearly, as speech recognition for children improves, more interaction can take place in the verbal domain, opening further opportunities for research.

Using a different platform to the one used here (the Aldebaran NAO) would also create a

useful extension to the current findings. As indicated previously, the Aldebaran NAO is limited in terms of certain social cues, such as facial expressions. It is possible that with the addition of other modalities, the relationship and interactivity with the child may also change. The characterisation of social behaviour through immediacy laid out in this research provides an ideal framework within which this extension could be performed to build on the findings here.

10.6.4 The Robot Role

As stated in Chapter 3, the aim here was not to replace human tutors, but instead use robots to offer additional opportunities to supplement current human tutoring provision. Robots can assume a wider variety of roles than a human in tutoring. For example, robots can assist teachers (Alemi et al., 2014), or offer children a chance to teach a less-able peer (Tanaka & Matsuzoe, 2012; Hood et al., 2015). Alternatively, robots could provide personalised support which falls outside of typical lessons or the school environment, such as additional language support for non-native children, as discussed in Belpaeme et al. (2015). These opportunities typically lend themselves to the one-to-one scenario explored here, but focus on the robot adopting a more peer-like role than the robot often did in the research here. These complementary roles are certainly worth further exploration.

The scope of this work focussed on the relationship between the child and the robot in response to social behaviour, and the impact that this has on the cognitive learning that takes place. However, cognitive learning is only one aspect of learning, with affective learning also playing a key role in the pedagogical process (Krathwohl et al., 1964). The affective domain includes consideration of attitudes towards learning and motivation for learning, so takes on increasing importance over the longer-term. This is an area which could initially be exploited by any novelty effects that social robots may introduce to learning, but further research into longer-term maintenance of both affective and cognitive learning effects would be worthwhile; an issue raised by many other researchers, and well summarised in Leite et al. (2013b).

10.6.5 Sustained Use and Adoption

Section 10.3 raised questions as to whether society and educators will accept educative social robots, and indeed Section 10.2 challenges whether this should even be an aim.

In both cases, the long-term well-being of children who interact for sustained periods with social robots appears to be of utmost importance. As such, research should seek to establish the cognitive and socio-emotional effects of long-term interaction with social robots. This must of course be conducted in an ecologically valid manner such that findings can inform decisions made at a societal level as to when and how the robotic technology should be used. It is likely that exploration in this direction will simultaneously require more complex robot behaviour (whether through regular updates to introduce novelty or through models of cognition) to sustain interactions over time. This is a substantial challenge not only for social robots in education, but for social robots more generally (Leite et al., 2013b).

10.7 Summary

This chapter outlined limitations in the approaches adopted in the experimental studies conducted throughout this thesis, highlighting the potential impact on the findings, and future avenues for research, where relevant. Additionally, the discussion was broadened to take a wider view of where the program of research undertaken here fits into the field of HRI and society more generally. The ethical implications of such technology were explored through other literature, as was societal and educator acceptance of the technology. From these considerations, it was suggested that further study into the long-term effect of social robots not only on child learning, but also their more general well-being should be pursued. To achieve this, improved robot social behaviour may be necessary (along with understanding of how to generate effective social behaviour), as well as an expanded record of efficacy over longer time periods for the educational aspects for the investment to be worthwhile. Many of these challenges will align with problems that the field of HRI is facing generally, but adaptations must be made to tailor findings for the education domain, and often also for children.

Chapter 11

Contribution and Conclusion

This chapter seeks to provide an overview of the findings and topics covered in this thesis. The contributions to the field of social HRI are outlined and summarised. Following this, a conclusion is provided to briefly encapsulate the primary outcome of this work.

11.1 Summary

The background to this thesis introduced a variety of work that suggested a difference in the way that humans perceive and interact with a physically present, real robot when compared to a virtual, on-screen robot. Some studies had shown that the embodiment of the real robot led to differences in compliance with robot instructions (Bainbridge et al., 2008), or in task performance (Leyzberg et al., 2012). Some work had suggested that differences were also present in learning outcomes, but learning measures were unclear or strict experimental controls were not in place (Han et al., 2008). As such, this was highlighted as an area that this research could make a contribution to.

Chapter 5 sought to explore the impact of robot embodiment on child behaviour and learning. A real and virtual robot were compared in a sorting task with children. There were no learning differences between the conditions, however this may have been due in part to complications with measuring learning with a biased dataset. Nonetheless, behavioural differences were found between embodiment conditions: children gazed more towards the real robot than the virtual one. This is an encouraging first step towards learning, given the social basis of learning (Kuhl, 2007b) and the important role of gaze in establishing attention (Wu et al., 2010).

These findings were furthered in Chapter 6 where a control condition with the robot removed was used in a prime number sorting task. It was found that when a robot was present, statistically significant learning occurred, but when the same information came from just the touchscreen, the learning was not significant. This observation provides support for the utility of using social robots in child education, with clear benefits in terms of cognitive learning outcomes. Other contemporary work reports similar learning advantages when robots are used in addition to other media. For example, Alemi et al. (2014) show that when using a robot in addition to a human teacher, children learn significantly more. Agreement between these studies, despite differing child ages and learning materials, contribute to a growing body of evidence demonstrating the advantages that real robots can confer in education.

However, Chapter 6 also revealed surprising findings, where a robot behaviour considered to have lower social skills, led to more learning than another with higher social skills. This was in line with the background to this research, which explored the impact of various robot social behaviours on human learning, revealing a complex and apparently inconsistent picture. It was suggested that a measure to characterise social behaviour would be beneficial, as this would aid researchers in the comparison of results in the context of different social behaviours. *Immediacy* was identified as an appropriate measure due to its extensive use in human-human interaction research and ties with learning gains. This was adapted and validated by the author for use with robots and with children in Chapter 4.

The two components of immediacy: nonverbal immediacy and verbal immediacy were then explored in Chapters 7 and 8, respectively. Robot social behaviours were designed (and verified) to be of high or low nonverbal (Chapter 7), or verbal (Chapter 8) immediacy. It was found that higher nonverbal immediacy led to increased cognitive learning, as predicated by the human-human literature, but that verbal immediacy did not affect child learning.

Chapter 9 brought together many of the results from the studies undertaken as part of this research, along with a human condition for comparison. These were tied together through the use of nonverbal immediacy ratings from both children and adults (through online crowdsourcing of observed interaction clips). Limitations in the utility of nonverbal immediacy for characterising interactions (as opposed to social cues) were highlighted

and the data collected was used to propose a model for the design of social behaviour for maximising learning gains. Limitations of the research were then raised and discussed; many of these led to suggestions for future work, which are expanded on in Section 10.6 below.

The synthesis of the experiments in Chapter 9, as well as the discussions that have taken place alongside each experiment in earlier chapters, show that there is data to support the thesis: a robot with tailored social behaviour will lead to greater child learning than a robot without this social behaviour. However, it was also shown that the precise implementation of this social behaviour is not always straightforward. It is necessary to factor in not just the social cues that the robot exhibits, but how congruent social cues are with one another, within the restrictions that certain robotic platforms can impose.

11.2 Contributions

This section will revisit the contributions outlined in the introduction (Chapter 1), with further expansion and explanation. The main contributions of this thesis are as follows:

- **Further evidence for the advantage of physical robots for child social responses**, particularly in gaze towards the robot, when compared to virtual agents (Chapter 5 and Kennedy et al., 2015b). It appears that children respond to a physical robot with greater gaze than they do with a virtual, on-screen robot. This difference in social response is encouraging for the benefits of using a physically present robot for cHRI.
- **Further evidence for the advantage of robots on learning outcomes** with findings to show that children learn more when a robot is present than when only a touchscreen is used, regardless of having the same lesson content (Chapter 6 and Kennedy et al., 2015d). This is a promising finding which supports the use of robots for tutoring children.
- **Adaptation and development of scales for use with children for characterising verbal and nonverbal social behaviour of humans and robots** (Chapters 3, 4, 8, and Kennedy et al., 2016b). These questionnaires are based on the short-form nonverbal immediacy questionnaire (Richmond et al., 2003) and the verbal immediacy questionnaire (Gorham, 1988). The language used was adapted to be suitable for children, and with robots. They can be seen in the appendices: Child Nonverbal

Immediacy Questionnaire (CNIQ), the Robot Nonverbal Immediacy Questionnaire (RNIQ), and the Robot Immediacy Questionnaire (RIQ).

- **Validation of the CNIQ and RNIQ with children and adults for humans and robots** (Chapter 4). The questionnaires were validated with a large sample of children and adults to verify that the questionnaires were suitable for use with children, and also that the concepts mapped to HRI. It was found that the questionnaires were reasonably reliable and learning effects predicted from the HHI literature were observed, providing further reassurance of the applicability of immediacy as a metric for cHRI.

- **Findings to show that a robot with higher nonverbal immediacy leads to greater child learning** (Chapter 7 and Kennedy et al., 2015e). When an increased number of nonverbal immediacy cues are utilised, such as gaze, gestures, and so on, child learning increases. This demonstrates the value of improving robot social behaviour for child learning outcomes.

- In short-term interactions, **higher verbal immediacy of a robot does not seem to improve the learning of children**, nor the retention of this learning (Chapter 8 and Kennedy et al., 2016b). Verbal immediacy centres around personalisation and aspects of 'friendliness' in dialogue, such as using the child's name, soliciting opinions, revealing personal information, and more besides. This does not appear to impact on learning in short interactions, although it is suggested that it may manifest into greater differences in the longer-term.

- **A proposed model of the relationship between robot social behaviour and child learning**. This model incorporates the quantity and congruence of social cues, where a combination of a high number of social cues with high congruency will lead to maximal possible learning (Chapter 9). This model is derived from the data collected throughout the experimental work undertaken here, and can be used to produce further hypotheses for future work.

11.3 Conclusion

The thesis presented here is that a robot with tailored social behaviour will positively influence the outcomes of tutoring interactions with children and consequently lead to

an increase in child learning when compared to a robot without this social behaviour. The work undertaken adds to the evidence that robots hold a social advantage over other technological media, and that this indeed leads to increased learning. Whilst evidence gathered through a series of experiments supports the thesis, it is also found that care must be taken when seeking to tailor social behaviour with the aim of improving interaction outcomes (as measured through learning). By characterising social behaviour using immediacy, it was shown that a greater use of immediacy behaviours generally does tend to lead to increased learning, but a complex picture emerges. If more social cues are used, but they are not congruent with one another, then this can negatively affect the learning outcome of the interaction. Merely the addition of more social behaviour is insufficient to increase learning; it is found that a balance should be struck between the addition of social cues, and the congruency of these cues.

Glossary

immediacy Describes both a set of concepts for characterising social behaviour, and a scale used to measure this construct. Please see verbal immediacy and nonverbal immediacy for further details. 9, 28, 33–35, 41, 43–45, 49, 52, 55, 57, 58, 65, 67, 119

learning For a full discussion of definitions of learning, please see Section 2.6. Throughout the thesis, learning will often refer specifically to cognitive learning gains that require the application of newly acquired knowledge to a novel problem, i.e., more than mere recall of information. i, xii, xiii, 2–5, 7–10, 34, 69, 71, 72, 100–103, 105–107, 109–117, 119–121, 125, 127–132, 134, 135, 138–141, 144–151, 153–162, 164–167, 169–172, 174, 175, 178–181, 183–189

nonverbal immediacy Addresses the nonverbal aspects of the immediacy construct, or can refer to the scale for these aspects. These include a number of nonverbal social cues, such as: gestures, gaze, vocal prosody, touch, facial expressions, and proximity. Please see Section 3.6 for an introduction to the scale used to measure nonverbal immediacy. ix–xii, 7, 9, 10, 20, 33, 34, 43, 45–49, 51–55, 57–67, 96, 97, 99, 116, 117, 119–121, 125, 127, 129–132, 139, 140, 143, 149, 153, 154, 156–158, 160–162, 164–166, 178, 179, 186–188, 191

tailored Used in reference to the thesis of this work, 'tailored' describes the intentional design of social behaviours such that they are adapted specifically for the application domain of educational interactions. i, 4, 8, 166, 169, 187, 188

verbal immediacy Addresses the verbal aspects of the immediacy construct, or can refer to the scale for these aspects. These include a variety of personalisation elements, such as revealing personal information in an interaction, as well as encouraging thoughts from students. A human-based verbal immediacy questionnaire can be

seen in Gorham (1988). xii, 7, 9, 33, 34, 41, 44–47, 131, 132, 135, 136, 139–141, 143, 147, 149, 150, 186, 188, 191

Acronyms

AIED Artificial Intelligence in Education. 41

ALIZ-E Adaptive Strategies for Sustainable Long-Term Social Interaction (European FP7 project). 4, 8, 42, 69, 99

cHRI Child-Robot Interaction. 31, 71, 157, 187, 188

CNIQ Child Nonverbal Immediacy Questionnaire. 6, 156, 188

DREAM Development of Robot-Enhanced Therapy for Children with Autism Spectrum Disorders (European FP7 project). 4, 8

HCI Human-Computer Interaction. 21, 23

HHI Human-Human Interaction. 5, 12, 15–17, 19–21, 23, 24, 30–32, 40, 43, 44, 62, 67, 100, 121, 128, 129, 157–159, 181, 188

HRI Human-Robot Interaction. x, xiii, 1–3, 5, 8, 11–13, 16–25, 28, 29, 32–34, 37, 40, 41, 43, 52, 61–65, 67, 70, 77, 78, 94, 97, 98, 100, 105, 116, 117, 121, 128, 132, 133, 135, 137, 154, 157, 159, 169, 179, 182, 184, 185, 188

ITS Intelligent Tutoring Systems. 12, 41, 154

L2TOR Second Language Tutoring using Social Robots (European Horizon 2020 project). 8

NVI Nonverbal Immediacy. xi, xiii, 120, 121, 123–129, 157–161, 163, 164, 172

RIQ Robot Immediacy Questionnaire. 6, 188

RNIQ Robot Nonverbal Immediacy Questionnaire. 6, 126, 188

WoZ Wizard-of-Oz. 77

Appendices

Appendix A

Example Video Coding Manual

Video Coding Manual for Sandtray Experiments

James Kennedy, Paul Baxter
WP1, ALIZ-E
Plymouth University, U.K.

October 2013

Abstract

This document describes the coding process for the video data obtained from child-robot interaction with the Sandtray setup. The initial purpose of this coding scheme is to characterise the robot- versus touchscreen-directed overt behaviours of the child to assess the social interaction.

Contents

1 Purpose of this Behaviour Coding **2**

2 Tools and Files **2**

2.1 ANVIL 2

2.2 Video Files 3

3 Behaviour Coding Categories **3**

3.1 Group: Interaction 3

3.2 Group: Child Behaviour 4

3.3 Group: Robot Behaviour 5

3.4 Group: Other 5

4 Coding Procedure **6**

4.1 ANVIL Issues 7

5 Validation **7**

A Robot Scripted Phrases **8**

1 Purpose of this Behaviour Coding

The purpose of video coding is to characterise the events that occur in a video stream, according to some defined set of measures/characteristics. This document outlines one such coding scheme, and provides guidelines for its application to a video dataset.

This particular coding scheme is concerned with characterising the video data obtained from the Sandtray study run at Salisbury Road Primary School, July 2013. The coding scheme is roughly based on the information provided in [1, 2], and the coding categories inspired by a range of HRI studies, including [3].

The initial purpose of this preliminary coding scheme is to characterise the robot- versus touchscreen-directed overt behaviours of the child to assess the social behaviours taking place.

2 Tools and Files

2.1 ANVIL

ANVIL is a freeware/opensource video annotation tool: it was chosen for its cross-platform support, its various import/export options, its prior application and validation on a number of datasets, and the prior experience that certain people within the ALIZ-E project have with it. Information and installation instructions can be found here: `http://www.anvil-software.org/`, and a brief description here: `http://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4492/tanmartin.pdf?sequence=1`.

Regarding the use of ANVIL for video annotation, it is useful to distinguish between three levels of organisation, which are used in the coding description below:

- *Element*: is an attribute/value/characterisation of a particular type, has 'start', 'end' and 'value' attributes.
- *Track*: is a group of elements for one type of information. Within a track, only one element can be attributed at any one time, elements within a track are non-overlapping.
- *Group*: is a collection of tracks related to a type of information.

The definition of the coding scheme is contained within a "Specification file" (an XML document, **.xml*). This is provided for the coding scheme described below. For compatibility, and to enable use of the various statistical methods built into ANVIL, *the provided specification file must remain unchanged* over the course of coding and analysis for this study.

Each coder should create a separate "Annotation file" (**.anvil*) for each of the videos to be coded: use the same naming convention as for the videos (described in next section), but append the initials of the coder (e.g. *d1i01-JK.anvil*).

2.2 Video Files

For compatibility with ANVIL, the videos have the following characteristics: **.mov* container, using the *Apple Cinepak codec*, with *uncompressed audio*, and a *70% quality* of the original video file (to reduce the file size - ANVIL recommends to only deal with video sizes of less than 1Gb).

Split video file names have the following format:

dDiXXcine.MOV

Where:

- "*D*" is the day; either 1 or 2.
- "*XX*" is the interaction number on that day.

There are a total of 28 video files to code, but for second coding, only a proportion of these will be done. The files to code will be:

- d1i08mov.cine
- d1i11mov.cine
- d1i13mov.cine
- d2i06mov.cine
- d2i13mov.cine

3 Behaviour Coding Categories

The coding relates only to overt behaviours of the robot and child, and not to perceived/inferred internal states or intentions. This is to keep the study an objective one, and which should also lead to a higher inter-rater reliability.

3.1 Group: Interaction

- *Track: Interaction stage.* Composed of the following elements: *Please try to assign one of these elements to the entire timeline.*
  - Introduction - from the moment the video starts until immediately before the robot starts to wave.
  - Robot Instructions - from the start of the robot wave, until it finishes saying 'now you can start'.
  - Pre-Test - from the end of the robot instructions speech until the robot moves to change the library when there are no images on the touchscreen.

- Main Activity - from the proceeding session until the robot introduces the post-test 'right, we'll do just one more set of aliens...'
- Post-Test - from the start of the 'right, we'll do just one more set of aliens...' speech until the child has categorised all touchscreen images.
- Robot Goodbye - from the proceeding category until the robot has finished waving.
- Debriefing - from the proceeding category until the child leaves the room (normally when the video ends).

3.2 Group: Child Behaviour

- *Track: Child gaze.* Based on eye direction, where the child's attention is directed, composed of the following elements. *Please try to assign one of these elements to the entire timeline.*
  - Towards robot
  - Towards touchscreen
  - Towards another person
  - Other: including occluded gaze
- *Track: Child gestures. No need to provide complete coverage of the timeline.*
  - Preparation/Moving: moving towards/away from screen with hand or moving between images on screen
  - Touching screen: whether moving or not
  - Pointing/Gesturing towards robot
  - Pointing/Gesturing towards another person
  - Other: primarily for annotating interesting looking behaviours that are not part of the defined coding scheme - fist pumps, pulling at clothes, repetitive scratching e.t.c.
- *Track: Child vocalisations.* Where distinguishable from the video. Just code for the length of the speech; there is no need to annotate the words which are spoken.
  - Towards robot
  - Towards another person
  - Towards self
  - Other
- *Track: Other. No need to provide complete coverage of the timeline.*
  - Undefined: no aspect of child behaviour is distinguishable, cannot be characterised using above track definitions, or is another behaviour of interest not covered above.

3.3 Group: Robot Behaviour

- *Track: Robot gaze.* Based on head direction (height of gaze must be inferred due to camera angle), with the following elements (bear in mind that in the current study, the robot gaze direction is randomly defined, except where a touchscreen directed movement is taking place). *Please try to assign one of these elements to the entire timeline.*
  - Towards Child
  - Towards Touchscreen
  - Other: including occluded gaze and gaze which is at neither the child or touchscreen, e.g. around the room. Please include a short comment if other.
- *Track: Robot gestures.* Bear in mind that the default behaviour of the robot is to shift its weight randomly, this behaviour need not be coded. *No need to provide complete coverage of the timeline.*
  - Highlight screen item - from the moment the robot starts to raise its arm to move a screen item, until the moment it has finished lowering.
  - Change screen library - again from the start until the end of movement, performed when screen items have run out.
  - Speech gesture - from the start of the movement, until the end of the gesture. Do not include the slow return to the 'default' position.
  - Other: including motions errors, etc.
- *Track: Robot Vocalisations.* These should cover the length of each phrase while the robot is talking. There is no need to annotate the words the robot is saying, just the length.
  - Unscripted/Feedback to child - these are used after the child does a move, or the robot is highlighting a move for the child
  - Scripted - at certain points in the interaction, the robot will follow a script. The scripted items are listed in the appendix of this document and each phrase should be coded as one element.
  - Other: not present in this case
- *Track: Other.* This track is primarily for annotations of interesting behaviour that is not covered in the defined coding scheme.

3.4 Group: Other

- *Track: Undefined.* This is a string field that may be used to either annotate situations that the camera is occluded, other people walk into the room, or other features not captured by the defined coding scheme. *No need to provide complete coverage of the timeline.*

4 Coding Procedure

To begin a new annotation, locate and open the desired video, select the "sals-sandtray-spec.xml" specification file, and "UTF-16" encoding (no particular reason for this encoding choice - but stick to it for consistency).

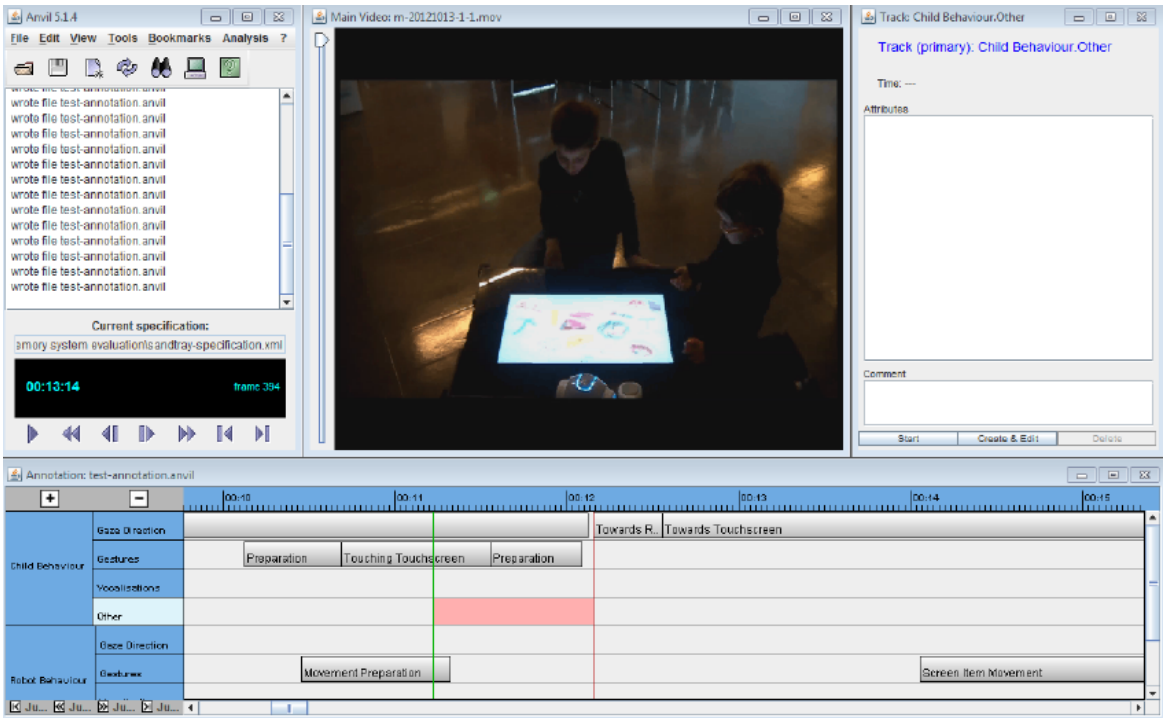

Figure 1: Screenshot of the ANVIL interface (In Windows), with a sample video and associated annotations.

In Figure 1, four panels of ANVIL are shown: (top left) main control panel with video playback control, open/save annotation files here; (top centre) the video playback panel (the SPACE key can be used to start/pause video play); (top-right) the track property panel; (bottom) the annotation timeline panel.

The various tracks and associated elements can be seen on the left hand side of this lower panel. Two coloured vertical lines can be seen here: the green one can be moved by double-clicking on the timeline (or press F1 on Windows; 1 on Mac) and indicates the start of an annotation object (element), the red one indicates the current play location in the video, and the end of an annotation object if desired. To add an annotation, right-click on the highlighted area (shown in pink in the lower panel of figure 1) and select '*Create and Edit*' (or press F3 on Windows; 3 on Mac). This brings up the set of allowed annotation options for that element. Choose the appropriate one, and add a clarification note if necessary. Click OK to, or Ctrl + Enter to save the element.

Recommended coding procedure: Making multiple passes through the video and coding only

one element per pass works best (rather than stepping through the video). When coding tracks which can change at a very quick rate, such as child gaze, it is recommended to use the arrow keys (left and right) to move forwards and backwards through the video on a frame-by-frame basis.

4.1 ANVIL Issues

Occasionally, when deleting or editing an item, ANVIL will appear to freeze. What has happened is the confirmation box has opened off-screen! To fix this (at least in Windows):

1. Open the Task Manager (ctrl + shift + esc)
2. Select the 'Warning' entry
3. Click 'End Task'

This should allow ANVIL to be used again as before. Your edit/delete will not have been completed, so you will have to try it again.

5 Validation

Validation of the video coding: two coders per video, use of built-in ANVIL tool to calculate inter-rater reliability (Cohen's corrected kappa, κ). Based on the characterisation in (Landis & Koch, 1977), we may consider that $0.41 < \kappa < 0.6$ to be 'moderate' agreement, $0.61 < \kappa < 0.8$ to be 'substantial' agreement, and $0.81 < \kappa < 1.0$ to be 'almost perfect' agreement.

References

[1] Peter H. Kahn, Batya Friedman, Nathan G. Freier, and Rachel Severson. *Coding Manual for Childrens Interactions with AIBO, the Robotic Dog: the Preschool Study.* 2003.

[2] Jean-marc Colletta, Olga Capirci, Carla Cristilli, Susan Goldin-meadow, Michle Guidetti, and Susan Levine. *Coding Manual.* 2009.

[3] Marcel Heerink, Marta Daz, Jordi Albo-Canals, Cecilio Angulo, Alex Barco, Judit Casacuberta, and Carles Garriga. A field study with primary school children on perception of social presence and interactive behavior with a pet robot. In *RoMan 2012*. IEEE Press, 2012.

A Robot Scripted Phrases

Below is a list of the robot scripted phrases and where they occur in the interaction. These should be coded as 'scripted' under robot vocalisations. Each phrase should be coded as one element.

- Robot Instructions
  - Hello! I'm Pop/Crackle.
  - Right, what we are going to be doing today is sorting out some aliens.
  - We have two species of aliens that are lost in space and we have to return them to their home planet. Okay?.
  - So here we have our different types of aliens and our two planets, the purple and the orange.
  - We need to sort them into their two different groups.
  - I'd like you to see if you can guess which planets the aliens are from.
  - You can touch an alien and you drag it to the planet you think it's from, and it'll tell you whether you are right or not.
  - I won't help you on your first go. Let's see how well you can do on your own!
  - Now you can start.
- Main Activity
  - Lovely, well done.
  - Now I'll give you a clue, the aliens from the purple planet all have something in common.
- Post-Test
  - Right, we'll do just one more set of aliens.
  - Using the practice we've just done, let's see how well you can do.
  - I won't help you this time.
  - Have a go.
- Robot Goodbye
  - Well done. thank you very much.
  - Thank you for helping me out today.
  - You can go back to your class.
  - Goodbye!

Appendix B

Crowdsourced Immediacy Results

Below is the table of data collected as described in Chapter 3, from adults rating the robot behaviour in several of the experimental conditions.

Table B.1: Adult crowdsourced nonverbal immediacy ratings for robot behaviours, including the chapter that the behaviour is evaluated in, and basic demographic details.

Chapter	**Condition**	**Adult *N***	**Age *M* (*SD*)**	**Gender**	**NVI Score *M* (95% *CI*)**
5	'Real' robot	37	35.8 (*13.2*)	24M/13F	51.9 (*50.5,53.3*)
5	'Virtual' robot	35	35.6 (*13.1*)	20M/15F	50.2 (*48.5,51.9*)
6	Social robot	33	29.0 (*10.4*)	21M/12F	49.0 (*47.6,50.4*)
6	Asocial robot	30	39.0 (*12.2*)	14M/16F	48.5 (*46.1,50.8*)
7	Low NVI robot	33	31.5 (*12.2*)	19M/14F	40.2 (*38.1,42.2*)
7	High NVI robot	31	35.6 (*11.7*)	18M/13F	48.4 (*46.9,50.0*)
9	Human	30	32.9 (*12.3*)	18M/12F	47.7 (*45.3,50.1*)

Appendix C

Chapter 4 Short Story Script

The following is the short story script as used in Chapter 4. The story is largely based on one from the following website: http://freestoriesforkids.com/children/stories-and-tales/robot-virus (produced here with permission from the author).

Hello, I'm Charlie. Today I'm going to tell you one of my favourite robot stories. It is about a boy, his name is Ricky, and his robot helper, Johnny. Ricky lived in a lovely futuristic house, which had everything you could ever want. Though he didn't help much around the house, Ricky was still as pleased as punch when his parents bought him the latest model of helper robot. As soon as it arrived, off it went; cooking, cleaning, ironing, and - most importantly - gathering up old clothes from Ricky's bedroom floor, which Ricky didn't like having to walk on.

On that first day, when Ricky went to sleep, he had left his bedroom in a truly disastrous state. When he woke up the next morning, everything was perfectly clean and tidy. In fact, it was actually too clean. Ricky could not find his favourite blue skateboard. However much he searched, it did not reappear, and the same was starting to happen with other things. Ricky looked with suspicion at the gleaming helper robot. He hatched a plan to spy on the robot, and began following it around the house.

Finally he caught it red-handed. It was picking up a toy to hide it. Off he went, running to his parents, to tell them that the helper was broken and badly programmed. Ricky asked them to have it changed. But his parents said absolutely not; it was impossible, they were delighted with the new helper, and that it was the best cleaner they had ever met. So Ricky needed to get some kind of proof; maybe take some hidden photos. He kept nagging his parents for 3 whole weeks about how much good stuff the robot was hiding. Ricky argued that this was not worth the clean house because toys are more important.

One day the robot was whirring past, and heard the boy's complaints. The robot returned with five of his toys, and some clothes for him."Here sire, I did not know it was bothering you", said the helper, with its metallic voice. "How could it not you thief?! You've been nicking my stuff for weeks", the boy answered, furiously. The robot replied, "the objects were left on the floor. I therefore calculated that you did not like them. I am programmed to collect all that is not wanted, and at night I send it to places other humans can use it. I am a maximum efficiency machine. Did you not know?".

Ricky started feeling ashamed. He had spent all his life treating things as though they were useless. He looked after nothing. Yet it was true that many other people would be delighted to treat those things with all the care in the world. And he understood that the robot was neither broken nor badly programmed, rather, it had been programmed extremely well! Since then, Ricky decided to become a Maximum Efficiency Boy, and he put real care into how he treated his things. He kept them tidy, and made sure that he didn't have more than was necessary. And, often, he would buy things, and take them along with his good friend, the robot, to help out those other people who needed them.

The end... I hope you enjoyed the story. Goodbye!

Appendix D

Chapter 4 Short Story Recall Questionnaire

Name_____

Please put a circle around your choice for each question.

1. What is the name of the boy in the story?

Ricky	Mickey	Harry	Jeff

2. What is the name of the robot in the story?

Rupert	John	Johnny	George

3. What was the most important thing for the robot to pick up from the floor of the boy's bedroom?

Clothes	Food	Toys	T-Shirts

4. What did the boy think about doing to get proof of the robot taking his things?

Taking photos	Shouting at it	Taking video	Telling his parents

5. What toy couldn't the boy find the first day after the robot had tidied?

Orange skateboard	Games console	Blue skateboard	Blue doll

6. How many toys did the robot give back to the boy after he complained?

Eight (8)	Five (5)	Three (3)	Six (6)

7. How long did the boy complain to his parents for?

Three (3) weeks	Eight (8) days	Three (3) days	Four (4) weeks

8. What type of boy did he decide to be at the end of the story?

Maximum efficiency	Tidy	Minimum efficiency	Messy

9. What type of robot is the one in the story?

Angry	Purple	Helper	Flying

10. What is the robot in the story especially good at?

Ironing	Swimming	Jumping	Cleaning

11. What was the moral of the story?

_____

_____

Appendix E

Robot Nonverbal Immediacy Questionnaire

Name_____

Please put a circle around your choice for each question.

1. The robot uses its hands and arms to gesture while talking to you

Never	Rarely	Sometimes	Often	Very Often

2. The robot uses a dull voice while talking to you

Never	Rarely	Sometimes	Often	Very Often

3. The robot looks at you while talking to you

Never	Rarely	Sometimes	Often	Very Often

4. The robot frowns while talking to you

Never	Rarely	Sometimes	Often	Very Often

5. The robot has a very tense body position while talking to you

Never	Rarely	Sometimes	Often	Very Often

6. The robot moves away from you while talking to you

Never	Rarely	Sometimes	Often	Very Often

7. The robot changes how it speaks while talking to you

Never	Rarely	Sometimes	Often	Very Often

8. The robot touches you on the shoulder or arm while talking to you

Never	Rarely	Sometimes	Often	Very Often

9. The robot smiles while talking to you

Never	Rarely	Sometimes	Often	Very Often

10. The robot looks away from you while talking to you

Never	Rarely	Sometimes	Often	Very Often

11. The robot has a relaxed body position while talking to you

Never	Rarely	Sometimes	Often	Very Often

12. The robot stays still while talking to you

Never	Rarely	Sometimes	Often	Very Often

13. The robot avoids touching you while talking to you

Never	Rarely	Sometimes	Often	Very Often

14. The robot moves closer to you while talking to you

Never	Rarely	Sometimes	Often	Very Often

15. The robot looks keen while talking to you

Never	Rarely	Sometimes	Often	Very Often

16. The robot is bored while talking to you

Never	Rarely	Sometimes	Often	Very Often

Appendix F

Child Nonverbal Immediacy Questionnaire

Name_____

Please put a circle around your choice for each question.

1. The man uses his hands and arms to gesture while talking to you

Never	Rarely	Sometimes	Often	Very Often

2. The man uses a dull voice while talking to you

Never	Rarely	Sometimes	Often	Very Often

3. The man looks at you while talking to you

Never	Rarely	Sometimes	Often	Very Often

4. The man frowns while talking to you

Never	Rarely	Sometimes	Often	Very Often

5. The man has a very tense body position while talking to you

Never	Rarely	Sometimes	Often	Very Often

6. The man moves away from you while talking to you

Never	Rarely	Sometimes	Often	Very Often

7. The man changes how he speaks while talking to you

Never	Rarely	Sometimes	Often	Very Often

8. The man touches you on the shoulder or arm while talking to you

Never	Rarely	Sometimes	Often	Very Often

9. The man smiles while talking to you

Never	Rarely	Sometimes	Often	Very Often

10. The man looks away from you while talking to you

Never	Rarely	Sometimes	Often	Very Often

11. The man has a relaxed body position while talking to you

Never	Rarely	Sometimes	Often	Very Often

12. The man stays still while talking to you

Never	Rarely	Sometimes	Often	Very Often

13. The man avoids touching you while talking to you

Never	Rarely	Sometimes	Often	Very Often

14. The man moves closer to you while talking to you

Never	Rarely	Sometimes	Often	Very Often

15. The man looks keen while talking to you

Never	Rarely	Sometimes	Often	Very Often

16. The man is bored while talking to you

Never	Rarely	Sometimes	Often	Very Often

Appendix G

Chapter 5 Robot Script

Robot Instructions

"Hello! I'm Pop/Crackle. Right, what we are going to be doing today is sorting out some aliens. We have two species of aliens that are lost in space and we have to return them to their home planet. Okay? So here we have our different types of aliens and our two planets, the purple and the orange. We need to sort them into their two different groups. I'd like you to see if you can guess which planets the aliens are from. You can touch an alien and you drag it to the planet you think it's from, and it'll tell you whether you are right or not. I won't help you on your first go. Let's see how well you can do on your own! Now you can start."

Prior to Guided Discovery Phase

"Lovely, well done. Now I'll give you a clue, the aliens from the purple planet all have something in common."

Prior to Post-Test

"Right, we'll do just one more set of aliens. Using the practice we've just done, let's see how well you can do. I won't help you this time. Have a go."

Robot Goodbye

"Well done. thank you very much. Thank you for helping me out today. You can go back to your class. Goodbye!"

Appendix H

Robot Immediacy Questionnaire

Name_____ Age _____

Please put a circle around your choice for each question.

1. The robot talked about itself

Never	Rarely	Sometimes	Often	Very Often

2. The robot asked me questions or encouraged me to talk

Never	Rarely	Sometimes	Often	Very Often

3. The robot talked about things I brought up even when it wasn't part of the lesson

Never	Rarely	Sometimes	Often	Very Often

4. The robot was funny

Never	Rarely	Sometimes	Often	Very Often

5. The robot used my name

Never	Rarely	Sometimes	Often	Very Often

6. The robot had a chat with me before, during, or after the lesson

Never	Rarely	Sometimes	Often	Very Often

7. The robot talked about "our" work, or what "we" are doing

Never	Rarely	Sometimes	Often	Very Often

8. The robot gave me feedback about my work

Never	Rarely	Sometimes	Often	Very Often

9. The robot asked me questions even if I didn't show that I wanted to talk

Never	Rarely	Sometimes	Often	Very Often

10. The robot asked how I felt about the lesson

Never	Rarely	Sometimes	Often	Very Often

11. The robot asked for my opinion

Never	Rarely	Sometimes	Often	Very Often

12. The robot thought that what I did was good

Never	Rarely	Sometimes	Often	Very Often

13. The robot talked about things that had nothing to do with the lesson

Never	Rarely	Sometimes	Often	Very Often

14. The robot invited me to use its name

Never	Rarely	Sometimes	Often	Very Often

15. The robot gestured while talking to me

Never	Rarely	Sometimes	Often	Very Often

16. The robot used a dull voice while talking to me

Never	Rarely	Sometimes	Often	Very Often

17. The robot looked at me while talking to me

Never	Rarely	Sometimes	Often	Very Often

18. The robot smiled while talking to me

Never	Rarely	Sometimes	Often	Very Often

19. The robot looked tense while talking to me

Never	Rarely	Sometimes	Often	Very Often

20. The robot moved around while talking to me

Never	Rarely	Sometimes	Often	Very Often

21. I think I did pretty well at this activity, compared to other classmates

Not at all	Not really	A bit	Yes	Yes very much

22. I was pretty skilled at this activity

Not at all	Not really	A bit	Yes	Yes very much

23. For me, I think the robot was like a - (*please circle only one*)

Brother or sister	Stranger	Classmate	Relative (e.g. cousin or aunt)
Friend	Parent	Teacher	Neighbour

Appendix I

Robot Relationship Questionnaire

Name……

For each question, please answer by circling your choice. Only circle <u>one</u> box per question.

For me, I think the robot was like a –

Brother or sister	Classmate
Stranger	Relative (e.g. cousin or aunt)
Friend	Parent
Teacher	Neighbour

For me, I think playing with the robot was like playing with -

A toy	A games console
A friend	A pet
Other _____	

Appendix J

Chapter 8 French Language Test

Name _____ Age _____

Please circle your choice to fill in the blank for each question

1. French: _____
   English: brother

homme	hermano	frère	bruder

2. French: _____
   English: Switzerland

Souriant	Suisse	Suiza	Schweiz

3. French: champignon
   English: _____

lemon	radish	mushroom	raspberry

4. French: ___ père
   English: the father

lu	la	lo	le

5. French: ___ tante
   English: the aunt

la	lo	le	lu

6. French: ___ Brésil
   English: Brazil

le	la	lo	lu

Name _____ Age _____

7. French: ___ Belgique

English: Belgium

lo	le	la	lu

8. French: ___ banane

English: the banana

la	lo	lu	le

9. French: ___ cassis

English: the blackcurrant

lo	la	lu	le

10. French: la _____

raisin	melon	tomate	brocoli

11. French: le _____

fraise	Congo	Pologne	fille

12. French: la _____

femme	prince	Togo	poivron

Bibliography

Admoni, H., Bank, C., Tan, J., Toneva, M., & Scassellati, B. (2011). Robot gaze does not reflexively cue human attention. In *Processings of the 33rd Annual Conference of the Cognitive Science Society (CogSci 2011)*, (pp. 1983–1988).

Aiello, J. R., & Aiello, T. D. C. (1974). The development of personal space: Proxemic behavior of children 6 through 16. *Human Ecology*, *2*(3), 177–189.

Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (2012). Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems*, (pp. 114–130). Springer.

Alemi, M., Meghdari, A., & Ghazisaedy, M. (2014). Employing Humanoid Robots for Teaching English Language in Iranian Junior High-Schools. *International Journal of Humanoid Robotics*, *11*(3).

Alemi, M., Meghdari, A., & Ghazisaedy, M. (2015). The impact of social robotics on l2 learners' anxiety and attitude in english vocabulary acquisition. *International Journal of Social Robotics*, (pp. 1–13).

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*(1), 1.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of personality and social psychology*, *64*(3), 431.

Anderson, L. W. (1975). Student involvement in learning and school achievement. *California Journal of Educational Research*, *26*(2), 53–62.

Andrist, S., Spannan, E., & Mutlu, B. (2013). Rhetorical robots: making robots more effective speakers using linguistic cues of expertise. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, (pp. 341–348). IEEE Press.

Askew, M., Brown, M., Rhodes, V., Johnson, D., & Wiliam, D. (1997). *Effective teachers of numeracy*. London: Kings College.

Atkinson, R. K., Mayer, R. E., & Merrill, M. M. (2005). Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology*, *30*(1), 117–139.

Bailenson, J., Blascovich, J., Beall, A., & Loomis, J. (2001). Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence*, *10*(6), 583–598.

Bailenson, J. N., Beall, A. C., Loomis, J., Blascovich, J., & Turk, M. (2005). Transformed social interaction, augmented gaze, and social influence in immersive virtual environments. *Human Communication Research*, *31*(4), 511–537.

Bailenson, J. N., Blascovich, J., Beall, A. C., & Loomis, J. M. (2003). Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin*, *29*(7), 819–833.

Baillie, J.-C., Demaille, A., Hocquet, Q., Nottale, M., & Tardieu, S. (2008). The Urbi Universal Platform for Robotics. In *Workshop Proceedings of the International Conference on Simulation, Modeling and Programming for Autonomous Robots*, SIMPAR 2008, (pp. 580–591).

Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008). The effect of presence on human-robot interaction. In *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN 2008, (pp. 701–706). IEEE.

Bainbridge, W. A., Hart, J. W., Kim, E. S., & Scassellati, B. (2011). The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, *3*(1), 41–52.

Bandura, A., & McClelland, D. C. (1977). *Social learning theory*. Prentice-Hall Englewood Cliffs, NJ.

Bandura, A., & Walters, R. H. (1963). *Social learning and personality development*, vol. 14. JSTOR.

Bartneck, C. (2003). Interacting with an embodied emotional character. In *Proceedings of the 2003 international conference on Designing pleasurable products and interfaces*, (pp. 55–60). ACM.

Bartneck, C., Kanda, T., Mubin, O., & Al Mahmud, A. (2009a). Does the design of a robot influence its animacy and perceived intelligence? *International Journal of Social Robotics*, *1*(2), 195–204.

Bartneck, C., KuliÄĞ, D., Croft, E., & Zoghbi, S. (2009b). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*(1), 71–81.

Baxter, P., Ashurst, E., Kennedy, J., Senft, E., Lemaignan, S., & Belpaeme, T. (2015). The wider supportive role of social robots in the classroom for teachers. In *Proceedings of the 1st International Workshop on Educational Robotics, at ICSR'15*.

Baxter, P., Ashurst, E., Read, R., Kennedy, J., & Belpaeme, T. (undergoing revisions). Robot Education Peers in a Situated Primary School Study: Personalisation Promotes Child Learning. *PLOS ONE*.

Baxter, P., & Belpaeme, T. (2014). Pervasive Memory: the Future of Long-Term Social HRI Lies in the Past. In *Third International Symposium on New Frontiers in Human-Robot Interaction at AISB*.

Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., & Belpaeme, T. (2016). From characterising three years of hri to methodology and reporting recommendations. In *Proceedings of the 11th ACM/IEEE Conference on Human-Robot Interaction (alt.HRI)*.

Baxter, P., Wood, R., Baroni, I., Kennedy, J., Nalin, M., & Belpaeme, T. (2013). Emergence of turn-taking in unstructured child-robot social interactions. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 77–78). IEEE Press.

Baxter, P., Wood, R., & Belpaeme, T. (2012). A touchscreen-based 'sandtray' to facilitate, mediate and contextualise human-robot social interaction. In *Proceedings of the 7th annual ACM/IEEE international conference on Human-Robot Interaction*, (pp. 105–106). ACM.

Baxter, P., et al. (2014). Tracking gaze over time in HRI as a proxy for engagement and attribution of social agency. In *Proc. of the 9th ACM/IEEE Int. Conf. on HRI*, (pp. 126–127).

Baylor, A., Ryu, J., & Shen, E. (2003). The effects of pedagogical agent voice and animation on learning, motivation and perceived persona. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, (pp. 452–458).

Baylor, A. L., & Kim, Y. (2004). Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. In *Intelligent Tutoring Systems*, (pp. 592–603). Springer.

Becker-Asano, C., Stahl, P., Ragni, M., Courgeon, M., Martin, J.-C., & Nebel, B. (2013). An affective virtual agent providing embodied feedback in the paired associate task: system design and evaluation. In *Intelligent Virtual Agents*, (pp. 406–415). Springer.

Beebe, B., Jaffe, J., & Lachmann, F. (1992). A dyadic systems view of communication. In N. Skolnick, & S. Warshaw (Eds.) *Relational Perspectives in Psychoanalysis*. Analytic Press.

Begg, I., Armour, V., & Kerr, T. (1985). On believing what we remember. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *17*(3), 199.

Beiker, S. A. (2012). Legal aspects of autonomous driving. *Santa Clara L. Rev.*, *52*, 1145.

Belpaeme, T., Baxter, P., De Greeff, J., Kennedy, J., Read, R., Looije, R., Neerincx, M., Baroni, I., & Zelati, M. C. (2013). Child-robot interaction: Perspectives and challenges. In *Proceedings of the 5th International Conference on Social Robotics*, ICSR '13, (pp. 452–459). Springer.

Belpaeme, T., Baxter, P., Read, R., Wood, R., Cuayáhuitl, H., Kiefer, B., Racioppa, S., Kruijff-Korbayová, I., Athanasopoulos, G., Enescu, V., Looije, R., Neerincx, M., Demiris, Y., Ros-Espinoza, R., Beck, A., Cañamero, L., Hiolle, A., Lewis, M., Baroni, I., Nalin, M., Cosi, P., Paci, G., Tesser, F., Sommavilla, G., & Humbert, R. (2012). Multimodal Child-Robot Interaction: Building Social Bonds. *Journal of Human-Robot Interaction*, *1*(2), 33–53.

Belpaeme, T., Kennedy, J., Baxter, P., Vogt, P., Krahmer, E. E., Kopp, S., Bergmann, K., Leseman, P., Küntay, A. C., Göksun, T., Pandey, A. K., Gelin, R., Koudelkova, P., & Deblieck, T. (2015). L2TOR - Second Language Learning Tutoring using Social Robots. In *Proceedings of the First International Workshop on Educational Robots at ICSR'15*.

Beranuy, M., Oberst, U., Carbonell, X., & Chamarro, A. (2009). Problematic internet and mobile phone use and clinical symptoms in college students: The role of emotional intelligence. *Computers in human behavior*, *25*(5), 1182–1187.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American psychologist*, *37*(3), 245.

Bertolini, A. (2013). Robots as products: the case for a realistic analysis of robotic applications and liability rules. *Law, Innovation and Technology*, *5*(2), 214–247.

Biocca, F., Harms, C., & Burgoon, J. K. (2003). Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators & Virtual Environments*, *12*(5), 456–480.

Birch, S. A., Vauthier, S. A., & Bloom, P. (2008). Three-and four-year-olds spontaneously use others' past performance to guide their learning. *Cognition*, *107*(3), 1018–1034.

Blackwell, C. K., Lauricella, A. R., Wartella, E., Robb, M., & Schomburg, R. (2013). Adoption and use of technology in early education: The interplay of extrinsic barriers and teacher attitudes. *Computers & Education*, *69*, 310–319.

Blancas, M., Vouloutsi, V., Grechuta, K., & Verschure, P. F. (2015). Effects of the robot's role on human-robot interaction in an educational scenario. In *Biomimetic and Biohybrid Systems*, (pp. 391–402). Springer.

Blanson Henkemans, O. A., Bierman, B. P., Janssen, J., Neerincx, M. A., Rosemarijn, van der Bosch, H., & van der Giessen, J. A. (2013). Using a robot to personalise health education for children with diabetes type 1: A pilot study. *Patient Education and Counseling*, *92*(2), 174–181.

Bloom, B., Engelhart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. Donald McKay.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational researcher*, (pp. 4–16).

Board, K., & Tinsley, T. (2015). *Language Trends 2014/15: The state of language learning in primary and secondary schools in England*. CfBT Education Trust.

Borgers, N., De Leeuw, E., & Hox, J. (2000). Children as respondents in survey research: Cognitive development and response quality 1. *Bulletin de methodologie Sociologique*, *66*(1), 60–75.

Borgers, N., Sikkel, D., & Hox, J. (2004). Response Effects in Surveys on Children and Adolescents: The Effect of Number of Response Options, Negative Wording, and Neutral Mid-Point. *Quality and Quantity*, *38*(1), 17–33.

Boucher, J.-D., Ventre-Dominey, J., Dominey, P. F., Fagel, S., & Bailly, G. (2010). Facilitative effects of communicative gaze and speech in human-robot cooperation. In *Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments*, AFFINE '10, (pp. 71–74). New York, NY, USA: ACM.

Breazeal, C. (2002). *Designing Sociable Robots*. Cambridge, MA, USA: MIT Press.

Breazeal, C. (2003). Toward sociable robots. *Robotics and autonomous systems*, *42*(3), 167–175.

Breazeal, C. (2004). Social interactions in HRI: the robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *34*(2), 181–186.

Broekens, J., Heerink, M., & Rosendal, H. (2009). Assistive social robots in elderly care: a review. *Gerontechnology (International Journal on the Fundamental Aspects of Technology to Serve the Ageing Society)*, *8*(2), 94–103.

Brown, L., Kerwin, R., & Howard, A. M. (2013). Applying behavioral strategies for student engagement using a robotic educational agent. In *2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, (pp. 4360–4365). IEEE.

Bull, R., & Gibson-Robinson, E. (1981). The influences of eye-gaze, style of dress, and locality on the amounts of money donated to a charity. *Human Relations*, *34*(10), 895–905.

Butland, M. J., & Beebe, S. A. (1992). Teacher immediacy and power in the classroom: The application of implicit communication theory. In *Proceedings of the Speech Communication Association Conference*. ERIC.

Byrd, C. E., McNeil, N., D'Mello, S., & Cook, S. W. (2014). Gesturing may not always make learning last. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, (pp. 1982–1987).

Calo, M. R. (2011). Robots and privacy. In P. Lin, K. Abney, & G. A. Bekey (Eds.) *Robot ethics: the ethical and social implications of robotics*. MIT press.

Cameron, D., Fernando, S., Collins, E., Millings, A., Moore, R., Sharkey, A., Evers, V., & Prescott, T. (2015). Presence of Life-Like Robot Expressions Influences Children's Enjoyment of Human-Robot Interactions in the Field. In *Proceedings of the 4th International Symposium on New Frontiers in Human-Robot Interaction at AISB 2015*.

Castellano, G., Paiva, A., Kappas, A., Aylett, R., Hastie, H., Barendregt, W., Nabais, F., & Bull, S. (2013). Towards empathic virtual and robotic tutors. In *Artificial Intelligence in Education*, (pp. 733–736). Springer.

Charisi, V., Davison, D., Wijnen, F., van der Meij, J., Reidsma, D., Prescott, T., van Joolingen, W., & Evers, V. (2015). Towards a Child-Robot Symbiotic Co-Development: a Theoretical Approach. In *Proceedings of the 4th International Symposium on New Frontiers in HRI at AISB 2015*.

Chen, T. L., King, C.-H., Thomaz, A. L., & Kemp, C. C. (2011). Touched by a robot: An investigation of subjective responses to robot-initiated touch. In *Proceedings of the 6th international conference on Human-robot interaction*, (pp. 457–464). ACM.

Chesebro, J. L. (2003). Effects of teacher clarity and nonverbal immediacy on student learning, receiver apprehension, and affect. *Communication Education*, *52*(2), 135–147.

Chesebro, J. L., & McCroskey, J. C. (1998). The relationship of teacher clarity and teacher immediacy with students' experiences of state receiver apprehension. *Communication quarterly*, *46*(4), 446–456.

Chesebro, J. L., & McCroskey, J. C. (2000). The relationship between students' reports of learning and their actual recall of lecture material: A validity test. *Communication Education*, *49*(3), 297–301.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices*. *Cognitive science*, *5*(2), 121–152.

Chidambaram, V., Chiang, Y.-H., & Mutlu, B. (2012). Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, (pp. 293–300). ACM.

Christensen, L. J., & Menzel, K. E. (1998). The linear relationship between student reports of teacher immediacy behaviors and perceptions of state motivation, and of cognitive, affective, and behavioral learning. *Communication Education*, *47*(1), 82–90.

Christophel, D. M. (1990). The relationships among teacher immediacy behaviors, student motivation, and learning. *Communication Education*, *39*(4), 323–340.

Coe, R., Aloisi, C., Higgns, S., & Major, L. E. (2014). What makes great teaching? Review of the underpinning research. Tech. rep., Sutton Trust.

Coleridge, S. T. (1817). *Biographia Literaria*. Project Gutenberg. Republished in 2004 by Project Gutenberg.

Comstock, J., Rowell, E., & Bowers, J. W. (1995). Food for thought: Teacher nonverbal immediacy, student learning, and curvilinearity. *Communication Education*, *44*(3), 251–266.

Coninx, A., Baxter, P., Oleari, E., Bellini, S., Bierman, B., Blanson-Henkemans, O., Canamero, L., Cosi, P., Enescu, V., Ros, R., Hiolle, A., Humbert, R., Kiefer, B., Kruijff-Korbayova, I., Looije, R., Mosconi, M., Neerincx, M., Paci, G., Patsis, G., Pozzi, C., Sacchitelli, F., Sahli, H., Sanna, A., Sommavilla, G., Tesser, F., Demiris, Y., & Belpaeme, T. (2016). Towards long-term social child-robot interaction: Using multi-activity switching to engage young users. *Journal of Human-Robot Interaction*, *5*(1), 32–67.

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, *106*(2), 1047–1058.

Cramer, H., Kemper, N., Amin, A., Wielinga, B., & Evers, V. (2009). 'give me a hug': The effects of touch and autonomy on people's responses to embodied social agents. *Computer Animation and Virtual Worlds*, *20*(2-3), 437–445.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and psychological measurement*, *64*(3), 391–418.

Crone, E. A., Jennings, J. R., & Van der Molen, M. W. (2004). Developmental change in feedback processing as reflected by phasic heart rate changes. *Developmental psychology*, *40*(6), 1228.

Dalzel-Job, O., Oberlander, J., & Smith, T. J. (2011). Don't look now: The relationship between mutual gaze, task performance and staring in second life. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, (pp. 832–837).

Dautenhahn, K. (2004). Robots We Like To Live With?!-A Developmental Perspective on a Personalized, Life-Long Robot Companion. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication*, RO-MAN 2004, (pp. 17–22). IEEE.

De Jong, T., & Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of educational research*, *68*(2), 179–201.

Dede, C. (2009). Immersive interfaces for engagement and learning. *science*, *323*(5910), 66–69.

Delaunay, F., De Greeff, J., & Belpaeme, T. (2009). Towards retro-projected robot faces: an alternative to mechatronic and android faces. In *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication*, (pp. 306–311). IEEE.

Department for Education (2016). Privacy notice: suggested text for pupils. Tech. rep., Department for Education.

Destephe, M., Brandao, M., Kishi, T., Zecca, M., Hashimoto, K., & Takanishi, A. (2015). Walking in the uncanny valley: importance of the attractiveness on the acceptance of a robot as a working partner. *Frontiers in psychology*, *6*.

Dimitrov, D. M., & Rumrill, P. D., Jr (2003). Pretest-posttest designs and measurement of change. *Work: A Journal of Prevention, Assessment and Rehabilitation*, *20*(2), 159–165.

DiPiro, J. (2010). Student learning: Perception versus reality. *American Journal of Pharmaceutical Education*, *74*(4), 63.

Diyas, Y., Brakk, D., Aimambetov, Y., & Sandygulova, A. (2016). Evaluating Peer Versus Teacher Robot within Educational Scenario of Programming Learning. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, (pp. 425–426).

Duffy, B. R., & Zawieska, K. (2012). Suspension of disbelief in social robotics. In *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN 2012, (pp. 484–489). IEEE.

Dunbar, R. I. (2002). The social brain hypothesis. *Foundations in social neuroscience*, *5*(71), 69.

European Commission (2012). Special Eurobarometer 382: Public Attitudes Towards Robots. Tech. rep., European Commission.
URL http://ec.europa.eu/public_opinion/archives/ebs/ebs_382_en.pdf

Fischer, K., Soto, B., Pantofaru, C., & Takayama, L. (2014). Initiating interactions in order to get help: Effects of social framing on people's responses to robots' requests for assistance. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN 2014, (pp. 999–1005). IEEE.

Fisher, J. D., Rytting, M., & Heslin, R. (1976). Hands touching hands: Affective and evaluative effects of an interpersonal touch. *Sociometry*, (pp. 416–421).

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, *42*(3), 143–166.

Freedman, R., Ali, S. S., & McRoy, S. (2000). What is an intelligent tutoring system? *Intelligence*, *11*(3), 15–16.

Fridin, M., & Belokopytov, M. (2014). Acceptance of socially assistive humanoid robot by preschool and elementary school teachers. *Computers in Human Behavior*, *33*, 23–31.

Fukuda, H., Shiomi, M., Nakagawa, K., & Ueda, K. (2012). 'midas touch' in human-robot interaction: evidence from event-related potentials during the ultimatum game. In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 131–132). ACM.

Fulford, C. P. (1992). Systematically designed text enhanced with compressed speech audio. In *Proceedings of Selected Research and Development Presentations at the Convention of the Association for Educational Communications and Technology*.

Garner, P. W. (2010). Emotional competence and its influences on teaching and learning. *Educational Psychology Review*, *22*(3), 297–321.

Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A. C., et al. (2005). Designing Robots for Long-Term Social Interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IROS '05, (pp. 1338–1343). IEEE.

Goldin-Meadow, S., Kim, S., & Singer, M. (1999). What the teacher's hands tell the student's mind about math. *Journal of Educational Psychology*, *91*(4), 720–730.

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12*(6), 516–522.

Goldin-Meadow, S., & Wagner, S. M. (2005). How our hands help us learn. *Trends in Cognitive Sciences*, *9*(5), 234–241.

Goldin-Meadow, S., Wein, D., & Chang, C. (1992). Assessing knowledge through gesture: Using children's hands to read their minds. *Cognition and Instruction*, *9*(3), 201–219.

Goodboy, A. K., Weber, K., & Bolkan, S. (2009). The effects of nonverbal and verbal immediacy on recall and multiple student learning indicators. *The Journal of Classroom Interaction*, *44*(1), 4–12.

Goodrich, M. A., & Schultz, A. C. (2007). Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, *1*(3), 203–275.

Goodwyn, S. W., & Acredolo, L. P. (1998). Encouraging symbolic gestures: A new perspective on the relationship between gesture and speech. *New Directions for Child and Adolescent Development*, *1998*(79), 61–73.

Gordon, G., & Breazeal, C. (2015). Bayesian Active Learning-based Robot Tutor for Children's Word-Reading Skills. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, AAAI-15.

Gordon, G., Breazeal, C., & Engel, S. (2015). Can Children Catch Curiosity from a Social Robot? In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 91–98). ACM.

Gorham, J. (1988). The relationship between verbal teacher immediacy behaviors and student learning. *Communication Education*, *37*(1), 40–53.

Green, J., & Weade, R. (1985). Reading between the words: Social cues to lesson participation. *Theory Into Practice*, *24*(1), 14–21.

Greene, L. R. (1977). Effects of verbal evaluation feedback and interpersonal distance on behavioral compliance. *Journal of Counseling Psychology*, *24*(1), 10.

Groves, C. (2006). Technological futures and non-reciprocal responsibility. *The International Journal of the Humanities*, *4*(2), 57–62.

Guéguen, N. (2002). Touch, awareness of touch, and compliance with a request. *Perceptual and motor skills*, *95*(2), 355–360.

Gullberg, M., & Holmqvist, K. (2002). Visual attention towards gestures in face-to-face interaction vs. on screen. In *Gesture and Sign Language in Human-Computer Interaction*, (pp. 206–214). Springer.

Gurung, R. A., & Vespia, K. (2007). Looking good, teaching well? linking liking, looks, and learning. *Teaching of Psychology*, *34*(1), 5–10.

Guskey, T. R. (2003). How classroom assessments improve learning. *Educational Leadership*, *60*(5), 6–11.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282.

Ham, J., Bokhorst, R., & Cabibihan, J. (2011). The influence of gazing and gestures of a storytelling robot on its persuasive power. In *International Conference on Social Robotics*.

Han, J., Jo, M., Park, S., & Kim, S. (2005). The educational use of home robots for children. In *Proceedings of the IEEE International Symposium on Robots and Human Interactive Communications*, RO-MAN 2005, (pp. 378–383). IEEE.

Han, J.-H., Jo, M.-H., Jones, V., & Jo, J.-H. (2008). Comparative study on the educational use of home robots for children. *Journal of Information Processing Systems*, *4*(4), 159–168.

Harvel, C. (2010). Guided discovery learning. In H. Lee (Ed.) *Faith-Based Education That Constructs: A Creative Dialogue between Contructivism and Faith-Based Education*, (pp. 169–172). Wipf and Stock Publishers.

Hasegawa, D., Cassell, J., & Araki, K. (2010). The role of embodiment and perspective in direction-giving systems. In *Proceedings of the AAAI Fall Workshop on Dialog with Robots*.

Herberg, J., Feller, S., Yengin, I., & Saerbeck, M. (2015). Robot watchfulness hinders learning performance. In *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN 2015, (pp. 153–160).

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: the cultural intelligence hypothesis. *science*, *317*(5843), 1360–1366.

Hewett, T. T., Baecker, R., Card, S., Carey, T., Gasen, J., Mantei, M., Perlman, G., Strong, G., & Verplank, W. (1992). *ACM SIGCHI Curricula for Human-Computer Interaction*. ACM.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371–406.

Hiroi, Y., & Ito, A. (2011). Influence of the size factor of a mobile robot moving toward a human on subjective acceptable distance. *Mobile Robots âĂŤ Current Trends*, (pp. 177–190).

Holekamp, K. E. (2007). Questioning the social intelligence hypothesis. *Trends in cognitive sciences*, *11*(2), 65–69.

Hood, D., Lemaignan, S., & Dillenbourg, P. (2015). When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 83–90). ACM.

Houston-Price, C., Plunkett, K., & Duffy, H. (2006). The use of social and salience cues in early word learning. *Journal of Experimental Child Psychology*, *95*(1), 27–55.

Howley, I., Kanda, T., Hayashi, K., & Rosé, C. (2014). Effects of social presence and social role on help-seeking and learning. In *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 415–422). ACM.

Huang, C.-M., & Mutlu, B. (2013). Modeling and evaluating narrative gestures for human-like robots. In *Proceedings of the Robotics: Science and Systems Conference*, RSS '13.

Huettenrauch, H., Severinson Eklundh, K., Green, A., & Topp, E. (2006). Investigating spatial relationships in human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (pp. 5052–5059).

Hulme, C., & Tordoff, V. (1989). Working memory development: The effects of speech rate, word length, and acoustic similarity on serial recall. *Journal of Experimental Child Psychology*, *47*(1), 72–87.

Hyun, E., Kim, S., Jang, S., & Park, S. (2008). Comparative study of effects of language instruction program using intelligence robot and multimedia on linguistic ability of young children. In *Proceedings of the 17th IEEE International Symposium on Robots and Human Interactive Communications*, RO-MAN 2008, (pp. 187–192). IEEE.

Jaffe, J., Beebe, B., Feldstein, S., Crown, C. L., Jasnow, M. D., Rochat, P., & Stern, D. N. (2001). Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the society for research in child development*, *66*(2), i–149.

Jeong, S., Gu, J., & Shin, D.-H. (2015). I am interested in what you are saying: Role of nonverbal immediacy cues in listening. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, (pp. 129–130). ACM.

Johnson, D. G., & Powers, T. M. (2005). Computer systems and responsibility: A normative look at technological complexity. *Ethics and Information Technology*, *7*(2), 99–107.

Jones, A., Bull, S., & Castellano, G. (2015). Open learner modelling with a robotic tutor. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, HRI'15 Extended Abstracts, (pp. 237–238). ACM.

Joosse, M., Sardar, A., & Evers, V. (2011). BEHAVE: A Set of Measures to Assess Users' Attitudinal and Non-verbal Behavioral Responses to a Robot's Social Behaviors. In *Proceedings of the 3rd International Conference on Social Robotics*, ICSR'11, (pp. 84–94). Springer.

Joosse, M., Sardar, A., Lohse, M., & Evers, V. (2013). BEHAVE-II: The Revised Set of Measures to Assess Users' Attitudinal and Behavioral Responses to a Social Robot. *International Journal of Social Robotics*, *5*(3), 379–388.

Jung, Y., & Lee, K. M. (2004). Effects of physical embodiment on social presence of social robots. *Proceedings of Presence*, (pp. 80–87).

Kahn Jr, P. H., Freier, N. G., Friedman, B., Severson, R. L., & Feldman, E. N. (2004). Social and moral relationships with robotic others? In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication*, (pp. 545–550). IEEE.

Kahn Jr, P. H., Friedman, B., Freier, N., & Severson, R. (2003). Coding manual for children's interactions with aibo, the robotic dog–the preschool study. *University of Washington CSE Technical Report 03-04*, *3*.

Kalyuga, S. (2008). Relative effectiveness of animated and static diagrams: An effect of learner prior knowledge. *Computers in Human Behavior*, *24*(3), 852–861.

Kamibeppu, K., & Sugiura, H. (2005). Impact of the mobile phone on junior high-school students' friendships in the tokyo metropolitan area. *Cyberpsychology & Behavior*, *8*(2), 121–130.

Kanda, T., Hirano, T., Eaton, D., & Ishiguro, H. (2004). Interactive Robots as Social Partners and Peer Tutors for Children: A Field Trial. *Human-Computer Interaction*, *19*(1), 61–84.

Kanda, T., Shimada, M., & Koizumi, S. (2012). Children learning with a social robot. In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 351–358). ACM.

Kar, S. S., & Ramalingam, A. (2013). Is 30 the magic number? issues in sample size estimation. *National Journal of Community Medicine*, *4*(1).

Karim, M. E., Lemaignan, S., & Mondada, F. (2015). A review: Can robots reshape k-12 stem education? In *Proceedings of the 2015 IEEE International Workshop on Advanced Robotics and its SOcial impacts*, EPFL-CONF-209219.

Kelley, D. H., & Gorham, J. (1988). Effects of immediacy on recall of information. *Communication Education*, *37*(3), 198–207.

Kelly, S. D., Manning, S. M., & Rodak, S. (2008). Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education. *Language and Linguistics Compass*, *2*(4), 569–588.

Kennedy, J., Baxter, P., & Belpaeme, T. (2013). Constraining Content in Mediated Unstructured Social Interactions: Studies in the Wild. In *Proceedings of the 5th International Workshop on Affective Interaction in Naturalistic Environments (AFFINE'13), at ACII'13*, (pp. 728–733). IEEE Computer Society.

Kennedy, J., Baxter, P., & Belpaeme, T. (2014a). Children Comply with a Robot's Indirect Requests. In *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 198–199).

Kennedy, J., Baxter, P., & Belpaeme, T. (2015a). Can Less be More? The Impact of Robot Social Behaviour on Human Learning. In *Proceedings of the 4th International Symposium on New Frontiers in HRI at AISB 2015*.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015b). Comparing Robot Embodiments in a Guided Discovery Learning Interaction with Children. *International Journal of Social Robotics*, *7*(2), 293–308.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015c). Head Pose Estimation is an Inadequate Replacement for Eye Gaze in Child-Robot Interaction. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, (pp. 35–36). ACM.

Kennedy, J., Baxter, P., & Belpaeme, T. (2015d). The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 67–74). ACM.

Kennedy, J., Baxter, P., & Belpaeme, T. (2017). Nonverbal Immediacy as a Characterisation of Social Behaviour for Human-Robot Interaction. *International Journal of Social Robotics*, *9*(1), 109–128.

Kennedy, J., Baxter, P., & Belpaeme, T. (under review). The Impact of Robot Tutor Nonverbal Social Behaviour on Child Learning. *Frontiers in ICT: Human-Media Interaction*.

Kennedy, J., Baxter, P., Senft, E., & Belpaeme, T. (2015e). Higher Nonverbal Immediacy Leads to Greater Learning Gains in Child-Robot Tutoring Interactions. In *Proceedings of the International Conference on Social Robotics*, (pp. 327–336). Springer.

Kennedy, J., Baxter, P., Senft, E., & Belpaeme, T. (2015f). Using Immediacy to Characterise Robot Social Behaviour in Child-Robot Interactions. In *Proceedings of the 1st Workshop on Evaluating Child-Robot Interaction at ICSR 2015*.

Kennedy, J., Baxter, P., Senft, E., & Belpaeme, T. (2016a). Heart vs Hard Drive: Children Learn More From a Human Tutor Than a Social Robot. In *Proceedings of the 11th ACM/IEEE Conference on Human-Robot Interaction*, (pp. 451–452). ACM.

Kennedy, J., Baxter, P., Senft, E., & Belpaeme, T. (2016b). Social Robot Tutoring for Child Second Language Learning. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 67–74). ACM.

Kennedy, J., de Greeff, J., Read, R., Baxter, P., & Belpaeme, T. (2014b). The chatbot strikes back. In *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '14, (pp. 103–103). ACM.

Kennedy, J., Lemaignan, S., & Belpaeme, T. (2016c). The Cautious Attitude of Teachers Towards Social Robots in Schools. In *Proceedings of the Robots 4 Learning Workshop at RO-MAN 2016*.

Kidd, C. D. (2008). *Designing for Long-term Human-robot Interaction and Application to Weight Loss*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA. AAI0819995.

Kim, Y., & Mutlu, B. (2014). How social distance shapes humanâĂŞrobot interaction. *International Journal of Human-Computer Studies*, *72*(12), 783–795.

Kim, Y., & Sundar, S. S. (2012). Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, *28*(1), 241–250.

Kipp, M. (2001). Anvil - a generic annotation tool for multimodal dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, (pp. 1367–1370).

Kory, J., & Breazeal, C. (2014). Storytelling with robots: Learning companions for preschool children's language development. In *Proceedings of the 23rd IEEE International Symposium on Robots and Human Interactive Communications*, RO-MAN 2014, (pp. 643–648). IEEE.

Kory Westlund, J., Dickens, L., Jeong, S., Harris, P., DeSteno, D., & Breazeal, C. (2015). A comparison of children learning new words from robots, tablets, & people. In *Proceedings of the 1st International Conference on Social Robots in Therapy and Education (New Friends)*.

Kory Westlund, J., Gordon, G., Spaulding, S., Lee, J. J., Plummer, L., Martinez, M., Das, M., & Breazeal, C. (2016). Lessons From Teachers on Performing HRI Studies with Young Children in Schools. In *Proceedings of the 11th ACM/IEEE International Conference on Human Robot Interation*, (pp. 383–390). IEEE Press.

Kose-Bagci, H., Ferrari, E., Dautenhahn, K., Syrdal, D. S., & Nehaniv, C. L. (2009). Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. *Advanced Robotics*, *23*(14), 1951–1996.

Krathwohl, D., Bloom, B., & Masia, B. (1964). *Taxonomy of educational objectives: The classification of educational goals. Handbook II: the affective domain*. Donald McKay.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice*, *41*(4), 212–218.

Kreijns, K., Kirschner, P. A., & Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Computers in human behavior*, *19*(3), 335–353.

Kruijff-Korbayova, I., Oleari, E., Baroni, I., Kiefer, B., Zelati, M. C., Pozzi, C., & Sanna, A. (2014). Effects of Off-Activity Talk in Human-Robot Interaction with Diabetic Children. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN 2014, (pp. 649–654).

Kuhl, P. K. (2007a). Cracking the speech code: How infants learn language. *Acoustical Science and Technology*, *28*(2), 71–83.

Kuhl, P. K. (2007b). Is speech learning 'gated' by the social brain? *Developmental science*, *10*(1), 110–120.

Kuhl, P. K. (2010). Brain mechanisms in early language acquisition. *Neuron*, *67*(5), 713–727.

Kuhlthau, C., Maniotes, L., & Caspari, A. (2007). *Guided inquiry: Learning in the 21st century*. Greenwood Publishing Group.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Landsberger, H. A. (1958). *Hawthorne Revisited: Management and the Worker, Its Critics, and Developments in Human Relations in Industry*. N.Y.S. School of Industrial and Labor Relations.

Langton, S. R. (2000). The mutual influence of gaze and head orientation in the analysis of social attention direction. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(3), 825–845.

Langton, S. R., & Bruce, V. (2000). You must see the point: Automatic processing of cues to the direction of social attention. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 747.

Lee, E., Lee, Y., Kye, B., & Ko, B. (2008). Elementary and middle school teachers', students' and parents' perception of robot-aided education in Korea. In *Proceedings of the World conference on educational multimedia, hypermedia and telecommunications*, (pp. 175–183).

Lee, M. K., Tang, K. P., Forlizzi, J., & Kiesler, S. (2011). Understanding users' perception of privacy in human-robot interaction. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 181–182). IEEE.

Leite, I., Castellano, G., Pereira, A., Martinho, C., & Paiva, A. (2014). Empathic Robots for Long-term Interaction. *International Journal of Social Robotics*, *6*(3), 329–341.

Leite, I., Hajishirzi, H., Andrist, S., & Lehman, J. (2013a). Managing chaos: models of turn-taking in character-multichild interactions. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction*, (pp. 43–50). ACM.

Leite, I., Martinho, C., & Paiva, A. (2013b). Social robots for long-term interaction: A survey. *International Journal of Social Robotics*, *5*(2), 291–308.

Leite, I., McCoy, M., Ullman, D., Salomons, N., & Scassellati, B. (2015). Comparing models of disengagement in individual and group interactions. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 99–105). ACM.

Leite, I., et al. (2012). Modelling Empathic Behaviour in a Robotic Game Companion for Children: an Ethnographic Study in Real-World Settings. In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 367–374). ACM.

Lemaignan, S., Garcia, F., Jacq, A., & Dillenbourg, P. (2016). From real-time attention assessment to "with-me-ness" in human-robot interaction. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE.

Leyzberg, D., Spaulding, S., & Scassellati, B. (2014). Personalizing robot tutors to individual learning differences. In *Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction*.

Leyzberg, D., Spaulding, S., Toneva, M., & Scassellati, B. (2012). The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, CogSci 2012, (pp. 1882–1887).

Li, J. (2015). The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, *77*, 23–37.

Lohan, K. S., Rohlfing, K., Saunders, J., Nehaniv, C., & Wrede, B. (2012). Contingency scaffolds language learning. In *IEEE International Conference on Development and Learning and Epigenetic Robotics*, ICDL, (pp. 1–6).

Looije, R., van der Zalm, A., Neerincx, M. A., & Beun, R.-J. (2012). Help, I need some body the effect of embodiment on playful learning. In *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN 2012, (pp. 718–724). IEEE.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1083.

Macario, J. F. (1991). Young children's use of color in classification: Foods and canonically colored objects. *Cognitive Development*, *6*(1), 17–46.

Macedonia, M., & von Kriegstein, K. (2012). Gestures enhance foreign language learning. *Biolinguistics*, *6*(3-4), 393–416.

Mauchly, J. W. (1940). Significance test for sphericity of a normal n-variate distribution. *The Annals of Mathematical Statistics*, *11*(2), 204–209.

Mayer, R. E. (2002). Rote versus meaningful learning. *Theory into practice*, *41*(4), 226–232.

Mayer, R. E., Fennell, S., Farmer, L., & Campbell, J. (2004). A personalization effect in multimedia learning: Students learn better when words are in conversational style rather than formal style. *Journal of Educational Psychology*, *96*(2), 389.

Mayer, R. E., Sobko, K., & Mautone, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology*, *95*(2), 419–425.

McCroskey, J., Morreale, S., & Brooks, M. (1994). Assessment of affect toward communication and affect toward instruction in communication. In *1994 SCA summer conference proceedings and prepared remarks: Assessing college student competence in speech communication*. Speech Communication Association.

McCroskey, J. C., Sallinen, A., Fayer, J. M., Richmond, V. P., & Barraclough, R. A. (1996). Nonverbal immediacy and cognitive learning: A cross-cultural investigation. *Communication Education*, *45*(3), 200–211.

Mehrabian, A. (1968). Some Referents and Measures of Nonverbal Behavior. *Behavior Research Methods & Instrumentation*, *1*(6), 203–207.

Merrill, D. C., Reiser, B. J., Merrill, S. K., & Landes, S. (1995). Tutoring: Guided learning by doing. *Cognition and instruction*, *13*(3), 315–372.

Mikulas, W. L. (1977). *Introduction: The Nature of Learning*. Nelson Hall.

Moore, R. K. (2012). A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Nature Scientific Reports*, *2*(864).

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *Robotics & Automation Magazine, IEEE*, *19*(2), 98–100.

Morse, A. F., Benitez, V. L., Belpaeme, T., Cangelosi, A., & Smith, L. B. (2015). Posture affects how robots and infants map words to objects. *PloS one*, *10*(3), e0116012.

Morse, A. F., Herrera, C., Clowes, R., Montebelli, A., & Ziemke, T. (2011). The role of robotic modelling in cognitive science. *New ideas in psychology*, *29*(3), 312–324.

Moshkina, L., Trickett, S., & Trafton, J. G. (2014). Social engagement in public places: a tale of one robot. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, (pp. 382–389). ACM.

Mumm, J., & Mutlu, B. (2011). Human-robot proxemics: Physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th International Conference on Human-robot Interaction*, HRI '11, (pp. 331–338). ACM.

Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education (IJAIED)*, *10*, 98–129.

Mutlu, B., & Tscheligi, M. (2017). *HRI '17: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. New York, USA: ACM.

Myers, C. B., & Myers, S. M. (2007). Assessing assessment: The effects of two exam formats on course achievement and evaluation. *Innovative Higher Education*, *31*(4), 227–236.

Nalin, M., Baroni, I., Kruijff-Korbayová, I., Canamero, L., Lewis, M., Beck, A., Cuayáhuitl, H., & Sanna, A. (2012). Children's adaptation in multi-session interaction with a humanoid robot. In *RO-MAN, 2012 IEEE*, (pp. 351–357). IEEE.

Nicol, D., Minty, I., & Sinclair, C. (2003). The social dimensions of online learning. *Innovations in education and Teaching International*, *40*(3), 270–280.

Nook, E. C., Lindquist, K. A., & Zaki, J. (2015). A new look at emotion perception: Concepts speed and shape facial emotion recognition. *Emotion*, *15*(5), 569–578.

Nussbaum, J. F. (1992). Effective teacher behaviors. *Communication Education*, *41*(2), 167–180.

Okita, S. Y., & Jamalian, A. (2011). Current challenges in integrating educational technology into elementary and middle school mathematics education. *Journal of Mathematics Education at Teachers College*, *2*(2), 49–58.

Okumura, Y., Kanakogi, Y., Kanda, T., Ishiguro, H., & Itakura, S. (2013). The power of human gaze on infant learning. *Cognition*, *128*(2), 127–133.

O'Neill, M. E. (2009). The genuine sieve of eratosthenes. *Journal of Functional Programming*, *19*(01), 95–106.

Otteson, J. P., & Otteson, C. R. (1979). Effect of teacher's gaze on children's story recall. *Perceptual and Motor Skills*, *50*(1), 35–42.

Park, H. W., Gelsomini, M., Lee, J. J., & Breazeal, C. (2017). Telling stories to robots: The effect of backchanneling on a childâĂŹs storytelling. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 100–108). ACM.

Park, H. W., & Howard, A. (2013). Providing tablets as collaborative-task workspace for human-robot interaction. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, (pp. 207–208). IEEE Press.

Pereira, A., Martinho, C., Leite, I., & Paiva, A. (2008). iCat, the chess player: the influence of embodiment in the enjoyment of a game. In *Proceedings of the 7th international conference on Autonomous agents and multiagent systems*, vol. 3, (pp. 1253–1256). International Foundation for Autonomous Agents and Multiagent Systems.

Perry, R. P., & Smart, J. C. (2007). *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective: An Evidence-based Perspective*. Springer Science & Business Media.

Peters, P. (2007). Gaining compliance through non-verbal communication. *Pepperdine Dispute Resolution Law Journal*, *7*(1), 87–112.

Pfeiffer, U. J., Vogeley, K., & Schilbach, L. (2013). From gaze cueing to dual eye-tracking: novel approaches to investigate the neural correlates of gaze in social interaction. *Neuroscience & Biobehavioral Reviews*, *37*(10), 2516–2528.

Pickett, C. L., Gardner, W. L., & Knowles, M. (2004). Getting a cue: The need to belong and enhanced sensitivity to social cues. *Personality and Social Psychology Bulletin*, *30*(9), 1095–1107.

Powell, K. L., Roberts, G., & Nettle, D. (2012). Eye images increase charitable donations: Evidence from an opportunistic field experiment in a supermarket. *Ethology*, *118*(11), 1096–1101.

Powers, A., Kiesler, S., Fussell, S., & Torrey, C. (2007). Comparing a computer agent with a humanoid robot. In *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 145–152). IEEE.

Rae, I., Takayama, L., & Mutlu, B. (2013). The influence of height in robot-mediated communication. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, (pp. 1–8). IEEE Press.

Ramachandran, A., Litoiu, A., & Scassellati, B. (2016). Shaping productive help-seeking behavior during robot-child tutoring interactions. In *Proceedings of the 11th ACM/IEEE Conference on Human-Robot Interaction*, (pp. 247–254). IEEE.

Reeves, B., & Nass, C. (1996). *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press.

Remland, M. S., & Jones, T. S. (1994). The Influence of Vocal Intensity and Touch on Compliance Gaining. *Journal of Social Psychology*, *134*(1), 89–97.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on sijtsma. *Psychometrika*, *74*(1), 145–154.

Richmond, V., McCroskey, J., & Payne, S. (1987). *Nonverbal Behavior in Interpersonal Relations*. Prentice-Hall.

Richmond, V. P., & McCroskey, J. C. (1998). *Nonverbal Communication in Interpersonal Relationships*. Allyn and Bacon, 3 ed.

Richmond, V. P., McCroskey, J. C., & Johnson, A. D. (2003). Development of the Nonverbal Immediacy Scale (NIS): Measures of Self- and Other-Perceived Nonverbal Immediacy. *Communication Quarterly*, *51*(4), 504–517.

Riek, L. D. (2012). Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, *1*(1).

Riether, N., Hegel, F., Wrede, B., & Horstmann, G. (2012). Social facilitation with social robots? In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 41–47). IEEE.

Robinson, R. Y., & Richmond, V. P. (1995). Validity of the verbal immediacy scale. *Communication Research Reports*, *12*(1), 80–84.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context*. Oxford University Press.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2012). How teacher turnover harms student achievement. *American Educational Research Journal*, *50*(1), 4–36.

Ros, R., Nalin, M., Wood, R., Baxter, P., Looije, R., Demiris, Y., Belpaeme, T., Giusti, A., & Pozzi, C. (2011). Child-robot interaction in the wild: Advice to the aspiring experimenter. In *Proceedings of the 13th international conference on multimodal interfaces*, (pp. 335–342). ACM.

Roth, W.-M. (2001). Gestures: Their role in teaching and learning. *Review of Educational Research*, *71*(3), 365–392.

Roth, W.-M., & Lawless, D. V. (2002). When up is down and down is up: Body orientation, proximity, and gestures as resources. *Language in Society*, *31*(01), 1–28.

Rumme, P., Saito, H., Ito, H., Oi, M., & Lepe, A. (2008). Gestures as effective teaching tools: Are students getting the point? In *Japanese Cognitive Science Society Meeting 2008*.

Saerbeck, M., Schut, T., Bartneck, C., & Janse, M. D. (2010). Expressive Robots in Education: Varying the Degree of Social Supportive Behavior of a Robotic Tutor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'10, (pp. 1613–1622). New York, NY, USA: ACM.

Salter, T., Dautenhahn, K., & te Boekhorst, R. (2006). Learning about natural human–robot interaction styles. *Robotics and Autonomous Systems*, *54*(2), 127–134.

Salter, T., Werry, I., & Michaud, F. (2008). Going into the wild in child–robot interaction studies: issues in social robotic development. *Intelligent Service Robotics*, *1*(2), 93–108.

Sariçoban, A. (2013). Pre-service elt teachers' attitudes towards computer use: A turkish survey. *Eurasian Journal of Educational Research*, *53*, 59–78.

Schermerhorn, P., & Scheutz, M. (2011). Disentangling the effects of robot affect, embodiment, and autonomy on human team members in a mixed-initiative task. In *ACHI 2011, The Fourth International Conference on Advances in Computer-Human Interactions*, (pp. 236–241).

Schermerhorn, P., Scheutz, M., & Crowell, C. R. (2008). Robot social presence and gender: Do females view robots differently than males? In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, (pp. 263–270). ACM.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics*, *15*(6), 657–680.

Scott, S. D., Sheelagh, M., Carpendale, T., & Inkpen, K. M. (2004). Territoriality in collaborative tabletop workspaces. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, (pp. 294–303). ACM.

Segrin, C. (1993). The effects of nonverbal behavior on outcomes of compliance gaining attempts. *Communication Studies*, *44*(3-4), 169–187.

Segura, E. M., Cramer, H., Gomes, P. F., Nylander, S., & Paiva, A. (2012). Revive!: reactions to migration between different embodiments when playing with robotic pets. In *Proceedings of the 11th International Conference on Interaction Design and Children*, (pp. 88–97). New York, NY, USA: ACM.

Serholt, S., Barendregt, W., Leite, I., Hastie, H., Jones, A., Paiva, A., Vasalou, A., & Castellano, G. (2014a). Teachers' views on the use of empathic robotic tutors in the classroom. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN 2014, (pp. 955–960). IEEE.

Serholt, S., Barendregt, W., Vasalou, A., Alves-Oliveira, P., Jones, A., Petisca, S., & Paiva, A. (2016). The case of classroom robots: teachers' deliberations on the ethical tensions. *AI & Society*, (pp. 1–19).
URL `http://dx.doi.org/10.1007/s00146-016-0667-2`

Serholt, S., Basedow, C. A., Barendregt, W., & Obaid, M. (2014b). Comparing a humanoid tutor to a human tutor delivering an instructional task to children. In *Proceedings of 2014 IEEE-RAS International Conference on Humanoid Robots*.

Setapen, A. A. M. (2012). *Creating robotic characters for long-term interaction*. Ph.D. thesis, Massachusetts Institute of Technology.

Shamsuddin, S., Yussof, H., Ismail, L. I., Mohamed, S., Hanapiah, F. A., & Zahari, N. I. (2012). Initial response in hri-a case study on evaluation of child with autism spectrum disorders interacting with a humanoid robot nao. *Procedia Engineering*, *41*, 1448–1455.

Sharkey, A. (2015). Robot teachers: The very idea! *Behavioral and Brain Sciences*, *38*.

Sharkey, A. J. (2016). Should we welcome robot teachers? *Ethics and Information Technology*, (pp. 1–15).

Sharma, M., Hildebrandt, D., Newman, G., Young, J. E., & Eskicioglu, R. (2013). Communicating affect via flight path: Exploring use of the laban effort system for designing affective locomotion paths. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*, HRI '13, (pp. 293–300).

Sherwood, J. V. (1987). Facilitative effects of gaze upon learning. *Perceptual and Motor Skills*, *64*(3c), 1275–1278.

Shiomi, M., Nakagawa, K., Shinozawa, K., Matsumura, R., Ishiguro, H., & Hagita, N. (2016). Does a robot's touch encourage human effort? *International Journal of Social Robotics*, (pp. 1–11).

Short, E., Swift-Spong, K., Greczek, J., Ramachandran, A., Litoiu, A., Grigore, E. C., Feil-Seifer, D., Shuster, S., Lee, J. J., Huang, S., Levonisova, S., Litz, S., Li, J., Ragusa, G., Spruijt-Metz, D., Matarić, M., & Scassellati, B. (2014). How to Train Your DragonBot: Socially Assistive Robots for Teaching Children About Nutrition Through Play. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*, RO-MAN 2014, (pp. 924–929). IEEE.

Siegler, R. S. (2007). Cognitive variability. *Developmental science*, *10*(1), 104–109.

Simonds, B. K., Meyer, K. R., Quinlan, M. M., & Hunt, S. K. (2006). Effects of instructor speech rate on student affective learning, recall, and perceptions of nonverbal immediacy, credibility, and clarity. *Communication Research Reports*, *23*(3), 187–197.

Sipilä, K. (2010). The impact of laptop provision on teacher attitudes towards ict. *Technology, Pedagogy and Education*, *19*(1), 3–16.

Smith III, J. P., Disessa, A. A., & Roschelle, J. (1994). Misconceptions reconceived: A constructivist analysis of knowledge in transition. *The journal of the learning sciences*, *3*(2), 115–163.

Spaulding, S., Gordon, G., & Breazeal, C. (2016). Affect-aware student models for robot tutors. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, (pp. 864–872). International Foundation for Autonomous Agents and Multiagent Systems.

Spencer, J. A., & Jordan, R. K. (1999). Learner centred approaches in medical education. *Bmj*, *318*(7193), 1280–1283.

Stanton, C. M., Kahn, P. H., Severson, R. L., Ruckert, J. H., & Gill, B. T. (2008). Robotic animals might aid in the social development of children with autism. In *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*, (pp. 271–278). IEEE.

Sung, J., Christensen, H. I., & Grinter, R. E. (2009). Robots in the wild: understanding long-term use. In *Prcoeedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 45–52). IEEE.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, *4*(4), 295–312.

Syrdal, D. S., Koay, K. L., Walters, M. L., & Dautenhahn, K. (2007a). A personalized robot companion?-the role of individual differences on spatial preferences in hri scenarios. In *Proceedings of the 16th IEEE International Symposium on Robot and Human interactive Communication*, RO-MAN 2007, (pp. 1143–1148). IEEE.

Syrdal, D. S., Walters, M. L., Otero, N., Koay, K. L., & Dautenhahn, K. (2007b). "He Knows When You Are Sleeping" – Privacy and the Personal Robot Companion. In *Proceedings of the Workshop on Human Implications of Human–Robot Interaction, at Association for the Advancement of Artificial Intelligence (AAAI'07)*, (pp. 28–33).

Szafir, D., & Mutlu, B. (2012). Pay Attention!: Designing Adaptive Agents that Monitor and Improve User Engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'12, (pp. 11–20). New York, NY, USA: ACM.

Takayama, L., & Pantofaru, C. (2009). Influences on proxemic behaviors in human-robot interaction. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (pp. 5495–5502).

Tanaka, F., Cicourel, A., & Movellan, J. R. (2007). Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences*, *104*(46), 17954–17958.

Tanaka, F., Isshiki, K., Takahashi, F., Uekusa, M., Sei, R., & Hayashi, K. (2015). Pepper learns together with children: Development of an educational application. In *IEEE-RAS 15th International Conference on Humanoid Robots*, HUMANOIDS 2015, (pp. 270–275). IEEE.

Tanaka, F., & Matsuzoe, S. (2012). Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning. *Journal of Human-Robot Interaction*, *1*(1), 78–95.

Tang, A., Tory, M., Po, B., Neumann, P., & Carpendale, S. (2006). Collaborative coupling over tabletop displays. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, (pp. 1181–1190). ACM.

Tapus, A., Ţăpuş, C., & Matarić, M. J. (2008). User–Robot Personality Matching and Assistive Robot Behavior Adaptation for Post-Stroke Rehabilitation Therapy. *Intelligent Service Robotics*, *1*(2), 169–183.

Teo, T. (2006). Attitudes toward computers: A study of post-secondary students in singapore. *Interactive Learning Environments*, *14*(1), 17–24.

Teo, T. (2008). Pre-service teachers' attitudes towards computer use: A singapore survey. *Australasian Journal of Educational Technology*, *24*(4).

Tessmer, M., & Richey, R. C. (1997). The role of context in learning and instructional design. *Educational technology research and development*, *45*(2), 85–115.

Thweatt, K. S., & McCroskey, J. C. (1998). The impact of teacher immediacy and misbehaviors on teacher credibility. *Communication Education*, *47*(4), 348–358.

Timms, M. J. (2016). Letting artificial intelligence in education out of the box: Educational cobots and smart classrooms. *International Journal of Artificial Intelligence in Education*, (pp. 1–12).

Triplett, N. (1898). The dynamogenic factors in pacemaking and competition. *The American journal of psychology*, *9*(4), 507–533.

Tronick, E. Z. (1989). Emotions and emotional communication in infants. *American psychologist*, *44*(2), 112.

Uziel, L. (2007). Individual differences in the social facilitation effect: A review and meta-analysis. *Journal of Research in Personality*, *41*(3), 579–601.

Valenzeno, L., Alibali, M. W., & Klatzky, R. (2003). Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology*, *28*(2), 187–204.

Van Joolingen, W. (1998). Cognitive tools for discovery learning. *International Journal Of Artificial Intelligence In Education (IJAIED)*, *10*, 385–397.

Vanlehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*(3), 227–265.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*(4), 197–221.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, *21*(3), 209–249.

Velez, J. J., & Cano, J. (2008). The relationship between teacher immediacy and student motivation. *Journal of Agricultural Education*, *49*(3), 76–86.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard university press.

Wainer, J., Feil-Seifer, D. J., Shell, D. A., & Mataric, M. J. (2007). Embodiment and human-robot interaction: A task-based perspective. In *Proceedings of the 16th IEEE International Symposium on Robot and Human interactive Communication*, RO-MAN 2007, (pp. 872–877). IEEE.

Walters, M. L., Dautenhahn, K., Te Boekhorst, R., Koay, K. L., Kaouri, C., Woods, S., Nehaniv, C., Lee, D., & Werry, I. (2005a). The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment. In *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, (pp. 347–352). IEEE.

Walters, M. L., Woods, S., Koay, K. L., & Dautenhahn, K. (2005b). Practical and methodological challenges in designing and conducting human-robot interaction studies. In *Proceedings of the AISB*, vol. 5, (pp. 110–119).

Wang, N., Johnson, W. L., & Gratch, J. (2010). Facial expressions and politeness effect in foreign language training system. In *Intelligent Tutoring Systems*, (pp. 165–173). Springer.

Wang, X., Williams, M.-A., Gardenfors, P., Vitale, J., Abidi, S., Johnston, B., Kuipers, B., & Huang, A. (2014). Directing human attention with pointing. In *Proceedings of the 23rd IEEE International Symposium onRobot and Human Interactive Communication*, RO-MAN 2014, (pp. 174–179). IEEE.

Weber, R., & Popova, L. (2012). Testing equivalence in communication research: Theory and application. *Communication Methods and Measures*, *6*(3), 190–213.

Westlund, J. K., & Breazeal, C. (2015). The Interplay of Robot Language Level with Children's Language Learning during Storytelling. In *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, (pp. 65–66). ACM.

Whiting, B. B., Edwards, C. P., & Edwards, C. P. (1992). *Children of different worlds: The formation of social behavior*. Harvard University Press.

Will, M. C. (1986). Educating children with learning problems: A shared responsibility. *Exceptional children*, *52*(5), 411–415.

Wilson, J. H., & Locker Jr, L. (2007). Immediacy scale represents four factors: Nonverbal and verbal components predict student outcomes. *The Journal of Classroom Interaction*, *42*(2), 4–10.

Witkin, H. A., Moore, C. A., Goodenough, D. R., & Cox, P. W. (1977). Field-dependent and field-independent cognitive styles and their educational implications. *Review of educational research*, (pp. 1–64).

Witt, P. L., & Schrodt, P. (2006). The influence of instructional technology use and teacher immediacy on student affect for teacher and course. *Communication Reports*, *19*(1), 1–15.

Witt, P. L., & Wheeless, L. R. (2001). An experimental study of teachers' verbal and nonverbal immediacy and students' affective and cognitive learning. *Communication Education*, *50*(4), 327–342.

Witt, P. L., Wheeless, L. R., & Allen, M. (2004). A Meta-Analytical Review of the Relationship Between Teacher Immediacy and Student Learning. *Communication Monographs*, *71*(2), 184–207.

Wu, R., Gopnik, A., Richardson, D. C., & Kirkham, N. Z. (2010). Social cues support learning about objects from statistics in infancy. In *Proceedings of the 32nd annual conference of the cognitive science society*, (pp. 1228–1233).

Wu, R., & Kirkham, N. Z. (2010). No two cues are alike: Depth of learning during infancy is dependent on what orients attention. *Journal of Experimental Child Psychology*, *107*(2), 118–136.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13), 2149–2165.

Zaga, C., Lohse, M., Truong, K. P., & Evers, V. (2015). The effect of a robot's social character on children's task engagement: Peer versus tutor. In *International Conference on Social Robotics*, (pp. 704–713). Springer.

Zajonc, R. B. (1965). Social facilitation. *Science*, *149*(3681), 269–274.

Zaki, J. (2013). Cue integration a common framework for social cognition and physical perception. *Perspectives on Psychological Science*, *8*(3), 296–312.

Zarbatany, L., & Lamb, M. E. (1985). Social referencing as a function of information source: Mothers versus strangers. *Infant Behavior and Development*, *8*(1), 25–33.

Ziafati, P., RodrÃ■guez Lera, F. J., Pinto Costa, A., Nazarikhorram, A., & Van Der Torre, L. (2017). Procrob architecture for personalized social robotics. In *Proceedings of the Robots 4 Learning Workshop at HRI 2017*.