International Journal of Social Robotics manuscript No.
This is an author pre-print and not the final version of the manuscript.
The final publication will be available at Springer via http://dx.doi.org/10.1007/s12369-014-0277-4

Comparing Robot Embodiments in a Guided Discovery Learning Interaction with Children

James Kennedy · Paul Baxter · Tony Belpaeme

Received: 31 March 2014 / Accepted: 13 December 2014

Abstract 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. In this paper, we seek to explore children's behaviour towards a robot tutor for children in a novel guided discovery learning interaction. 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 is assessed. The results demonstrate that children overcome strong incorrect biases in the material to be learned, but with no significant differences between embodiment conditions. However, the data do suggest that the use of real robots carries an advantage in terms of social presence that could provide educational benefits.

Keywords Social Robotics · Embodiment · Human-Robot Interaction · Child-Robot Interaction · Child Learning

1 Introduction

Child education is emerging as one of the many promising application domains for social human-robot interaction (HRI). In this context, robots have the potential to both support and augment existing educational strategies by, for example, increasing children's motivation to learn, or supplementing teacher-led learning with more targeted support for individuals. In this paper, we seek to provide evidence in support of the role of social robots as tutors, facilitating the learning progress of the child.

The use of robots in education has been the subject of some prior exploration, with some prominent studies discussed in [28]. The majority of these studies have been primarily concerned with the social presence of the robot. However, there have been recent attempts to explore responses to a robot's social cues and the impact that this can have on learning. An early example by Kanda *et al.* studied a robot placed in a classroom for a two week period [21]. The results provided an indication that robots could successfully be used to improve children's learning, although most of the effect observed is attributed to the presence of the robot increasing motivation, rather than to specific behaviours of the robot. Regarding the role of social behaviour more specifically, Huang and Mutlu have studied the impact of robot gestures on information recall [19], finding that the types of gestures that a robot uses can influence how much human participants could recall from a presentation.

This paper seeks to assess how the embodiment of a robot tutor influences children's behaviour during an interaction in which they aim to learn novel information. The focus is on single robot–single child interactions, with the robot taking on the role of a tutor. After deriving a suitable robot tutoring behaviour from observations of human tutor–learner interactions in the same interaction context [23], the purpose of the present study is to evaluate whether the resulting robot system can facilitate child learning. 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

Centre for Robotics and Neural Systems
Cognition Institute
Plymouth University, U.K.
E-mail: james.kennedy@plymouth.ac.uk
E-mail: paul.baxter@plymouth.ac.uk
E-mail: tony.belpaeme@plymouth.ac.uk

children learn, and secondly, how the children behave in the interaction in response to the tutoring strategy of the robot.

The remainder of this paper is organised as follows. In reviewing the literature in robot-supported learning, the issue of robotic embodiment versus virtual agents is raised as a primary point of consideration, with an examination of teaching styles showing the benefit of a learner self-guided approach (Section 2). We then describe the hypotheses, experimental setup and methodology used for the present study (Sections 3 and 4). A number of different aspects of the results are examined in detail. Firstly, the overall learning effects are analysed, taking into account an apparent bias in the unfamiliar subject matter (Sections 5 and 5.1). Secondly, the interaction behaviours of the child in response to the robot behaviour is examined in greater detail, demonstrating the differential effect of real robot embodiment (Section 6). Finally, we conclude by examining the support (or otherwise) for the experimental hypotheses (Sections 7 and 8).

2 Embodiment and Tutoring in HRI

Learning has been used as a metric in a multitude of HRI experiments. These experiments are often focussed on the embodiment of the robot, or on comparing the robot with other educational media, such as computers or paper-based resources. This section will review a number of these studies, which are later used as a basis for the experimental design used in the study presented in this paper. This section will also serve as a background to some of the decisions made about how learning is measured in this paper.

2.1 Agent Embodiment in Tutoring Interactions

It remains unclear how a robot's embodiment will impact upon the social interaction which takes place and, ultimately, on learning. Social interaction consists of many different elements; some verbal and some nonverbal [8]. These cues have been shown to influence the impression we have of an interaction partner [42]. Real and virtual robots provide different affordances which mean that they can greatly differ in the way that they provide nonverbal cues.

Whilst the measurements used are somewhat unclear, Han *et al.* suggest that robots can be more effective educators than equivalent web-based instruction or books with audio [16]. These findings have recently been supported by Leyzberg *et al.* who compared an embodied robot tutor with video and voice conditions [29].

Previous child-robot interaction (cHRI) studies comparing robot embodiments have found that children look more often and for longer periods at a real robot than a virtual robot [31]. Although not comparing embodiments, it has also been found that task performance will improve when a real robot provides subjects with more gaze [36]. Additionally, Bartneck found that people enjoyed playing a game no more or less with a virtual or real robot, but that people scored higher with a physical robot present [3]. In a drumming game, children performed better with a real robot than when collaborating with a virtual robot [25].

Conversely, Powers *et al.* found that participants remembered less of their conversation with a robot after communicating with a real robot when compared with a virtual robot [40]. They postulate that this is because people are distracted by the novelty of the physical robot. Similar effects have been seen when comparing virtual agents with simple paper media. Users remembered less about a healthy eating message when it was delivered by a virtual agent than when it was delivered on paper [9].

Real robots would appear to hold some advantages over virtual robots in social interactions, provided that nonverbal cues are used effectively. However, it is unclear whether the real robot improves task performance, or distracts from a task. Indeed, the varied results from previous literature suggest that it is necessary to study the impact of the robot's embodiment in the experiment conducted here.

2.2 Teaching Styles

In many HRI studies there has been a focus on prescriptive tutoring, with the robot providing instructional lessons to subjects; a 'teacher-centered' approach to learning, for example [30]. However, educational literature suggests that a 'learner-centered' 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 [50]. Such an approach is taken in [22], for example, where children undergo collaborative learning with a robot in a variety of group and individual lessons.

The learner-centered approach taken in this study, 'guided discovery learning', has overlaps with the collaborative learning seen in [21], 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 [13]. It has also been suggested that this type of learning can promote a better understanding of a domain when compared to teacher-centered learning [1, 55].

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 [13, 17].

This application of a different teaching style contributes to the novelty of the work conducted here. The teaching style requires the programming of a robot behaviour which does not deliver a complete instructional lesson to participants, but instead guides the child by making them analyse their own approach to solving the problem. To the authors knowledge, the specific application of this teaching context has previously been unexplored in HRI.

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 [22, 30, 39, 45]).

3 Experiment 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. The hypotheses for the study are enumerated below:

1. It is hypothesised that the real robot will attract more gaze than the virtual robot from the children. Other work, such as [31], 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.
2. If the robot behaviour is sufficiently socially contingent then the children will remain engaged with both the robot and the task throughout the interaction.
3. Prior studies with a similar task structure and hardware configuration (robot with a large touchscreen), for example [5], have found that the children will gaze more towards the touchscreen than the robot, but they will still pay attention to, and respond to, the robot's behaviour. The same is expected to occur here.
4. We hypothesise that there will be a difference in learning gains between the two embodiment conditions. This is based on the findings of other studies, which have found that the physical presence of a robot causes an increase in learning gains, for example [29].

4 Methodology

The study design 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 [4]. 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 [5, 23]. 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.

4.1 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, average age=7.9, *SD*=0.31; 15 real robot, 13 virtual).

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, average age=7.9, *SD*=0.33; 14 real robot, 12 virtual).

4.2 Experimental Conditions

A two-condition, between-subject design was employed for the study. The embodiment of the robot was swapped between the two conditions. In one condition, children were presented with the real, physical robot, in the other they were presented with a large monitor displaying the virtual robot. The use of these two embodiment conditions allows evaluation of the experiment hypotheses (Section 3), 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 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 [48]. Although the children were interacting with the robot in only one of the embodiment conditions,

Fig. 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.

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 1).

4.3 Experimental Set-Up

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 devices are outlined in Sections 4.4 and 4.5 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 2).

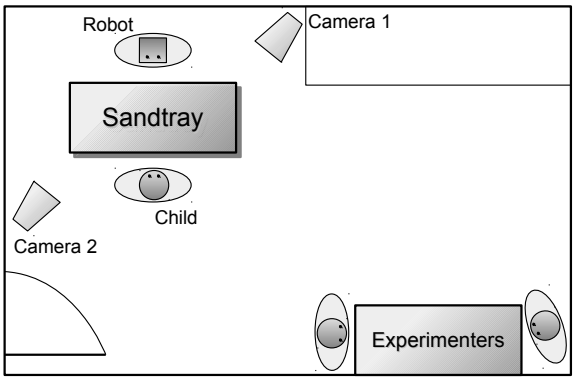

Fig. 2 Schematic overview of the mediation device-centered interactions under investigation in this paper. Two interactants (the child and the robot) face one another over the touchscreen. Two video cameras record the interactants during the studies. The experimenters are also in the room, but out of view of the child. Figure not to scale.

4.4 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 [49]. It has been shown that differing knowledge levels lead to different interpretations of a problem [11] and also require substantially different teaching formats to cope with this [20]. 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.* [32]).

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 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

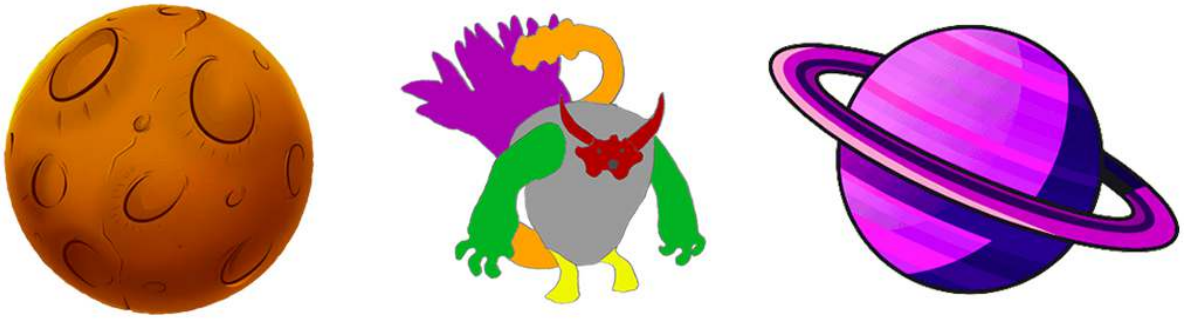

Fig. 3 From left to right: 1. the 'orange planet' category image, 2. an outline of one of the aliens with shading to differentiate the body parts which can be manipulated (an actual image has not been used due to copyright restrictions), 3. the 'purple planet' category image

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 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 to 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.

4.5 Robot Behaviour

A behaviour for the robot was created by analysing the behaviour of a human teacher when guiding a student unknown to them through a sorting task on the Sandtray, as in [23]. The human teacher 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 interactions with different children underwent video analysis in order to get an objective measure of particular movements and vocalisations made in the interaction. 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 A), 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 robot was provided with information about images by the mediator, allowing it 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 [26, 34] and Section 2.2, 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 the robot remained 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 [5, 23].

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 [41]). 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 4.8 for more details of this). The Wizard was one of the experimenters located in the room with the child, as described in Section 4.3.

4.6 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 4.5. 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 [12].

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 [7, 43].

Section 4.4 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.

4.7 Measuring Learning

As is commonly applied in HRI studies examining learning, pre- and post-tests (as described in [14]) 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 4.4. 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.

4.8 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 learning phase. This would then trigger the post-test script at the end of the current library.

The average length of an interaction was 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 308s, *SD*=45s.

4.9 Video Data

All 28 videos were coded by one coder; the tracks coded were as follows:

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

The coding scheme used was as objective as possible, based solely on overt child or robot behaviour. 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 [35], [38] and [51]. The videos 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 [27]. Table 1 shows the agreement for the tracks which will be used for analysis in Section 6 of this paper.

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 1 Inter-coder agreement by track coded.

5 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 4.7, 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. The post-test score (M=9.12, SD=2.44) was significantly higher than the pre-test score (M=7.08, SD=1.83), $t(25)$=3.016, p=0.006. 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 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 5.1.

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 an average of 7.93 (SD=1.82) for Test A and an average of 6.08 (SD=1.31) for Test B. This explains why learning effects measured from Test A to Test B may be hidden, but may be amplified from Test B to Test A.

Condition A	Condition B	*t*-test used	A mean (*n*, *SD*)	B mean (*n*, *SD*)	*p* value	*t* (*df*)
Pre-test A	Post-test B	Two tailed, paired	7.93 (14, 1.82)	8.43 (14, 2.93)	0.627	$t(13)=0.498$
Pre-test B	Post-test A	Two tailed, paired	6.08 (12, 1.31)	9.92 (12, 1.44)	* <0.001	$t(11)=6.823$
Virtual gain	Real gain	Two tailed, unpaired	2.42 (12, 2.78)	1.71 (14, 4.01)	0.614	$t(24)=0.510$
Male gain	Female gain	Two tailed, unpaired	1.45 (11, 3.11)	2.47 (15, 3.72)	0.471	$t(24)=0.733$
Pre-test A	Post-test A	Two tailed, unpaired	7.93 (14, 1.82)	9.92 (12, 1.44)	* 0.005	$t(24)=3.051$
Pre-test B	Post-test B	Two tailed, unpaired	6.08 (12, 1.31)	8.43 (14, 2.93)	* 0.017	$t(24)=2.558$

Table 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 0.05 level.

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 2).

5.1 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 to 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 [33]. Utilising 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 1 was formulated as a measure of bias.

$$bias = \%P - \%O \tag{1}$$

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 male and 30 females completed the tests, on average, the children were 7.2 years old, *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 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 0.761 between bias value and percentage categorised as purple, and -0.761 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 4):

- bias >0.2 = 9 images biased towards purple
- –0.2 <bias <0.2 = 9 minimally biased images
- bias <–0.2 = 6 images biased towards orange

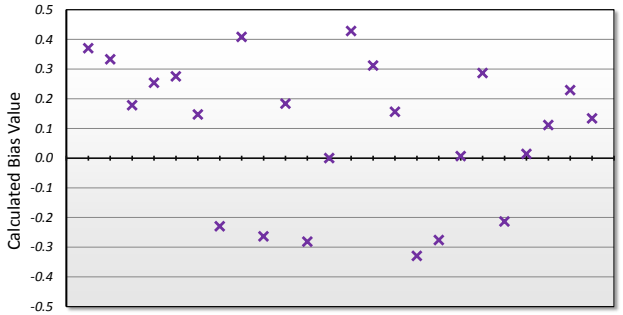

Fig. 4 Calculated bias value for each of the images used in the experimental pre- and post-tests using Equation 1.

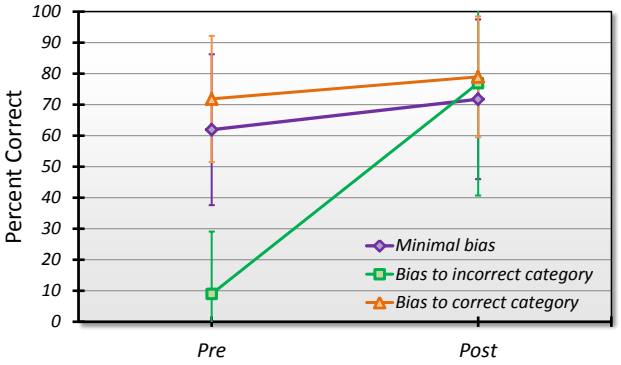

Fig. 5 Pre-test and post-test percentage of correct answers for images, grouped by bias type and direction. Error bars show the standard deviation.

5.2 Accounting for Bias

Table 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 5.1. 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 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, Table 3). The bias groupings were considered across the two embodiment conditions, but no significant differences were found.

6 Behaviour and Embodiment

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 stated in Section 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.

6.1 Compliance

The children clearly complied with the robot's suggestions for moves, as the percentages of responses below show. 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 [24].

- 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

There were no significant differences between the two embodiment conditions for the number of moves taken immediately (p=0.129, $t(26)$=1.568), although the real robot had a slightly higher average of 89.5% (SD=21.3), compared to 77.7% (SD=17.9) for the virtual robot. The lack of significant difference here is not surprising as the robot suggestions

Bias group	Pre-test (*SD*)	Post-test (*SD*)	*p* value	*t* (*df*)
Minimal	62% (24.3)	72% (25.7)	0.195	$t(25)=1.330$
Bias to correct	72% (20.3)	79% (19.4)	0.216	$t(25)=1.268$
Bias to incorrect	9% (20.1)	77% (36.2)	* <0.001	$t(25)=9.657$

Table 3 Learning effect *t*-test results, grouped by bias. All *t*-tests are two tailed, unpaired tests. Average percent correct is shown for both pre- and post-tests. * indicates a significant *p* value at the 0.05 level.

are simple and not unusual [2]. No 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 = -0.1369.

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.

6.2 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. Children interacting with the real robot (M=5.19, SD=1.29) make significantly more gazes towards the robot per minute than those in the virtual robot condition (M=4.13, SD=1.12), $t(26)=2.296$, p=0.030. The length of each individual gaze is similar between conditions, so those in the real robot condition (M=9.40, SD=1.88) spend significantly more seconds per minute of interaction gazing towards the robot than those interacting with the virtual robot (M=7.53, SD=1.93), $t(26)=2.586$, p= 0.016. This result confirms findings from [31] in a new context, and also supports Hypothesis 1.

A one-way between subjects ANOVA was conducted to compare the effect of interaction time on the child's gaze towards the *virtual* robot. There was no significant effect of interaction segment on child gaze towards the robot at the p <.05 level for the three segments [$F(2,36)=2.445$, p=0.101]. A one-way between subjects ANOVA was conducted to compare the effect of interaction segment on the child's gaze towards 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 [$F(2,42)=5.676$, p=0.007]. Post-hoc comparisons using the Tukey HSD test indicate that the mean score for the first segment (M=6.64, SD=3.26) was significantly different to the second segment (M=4.26, SD=1.06) and to the third (M=4.42, SD=1.50), with p values of 0.012 and 0.020 respectively. The second and third segments had no significant difference, p=0.976. 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 6.

It is suggested that the drop in gaze for the virtual robot is not significant 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 at 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 [6]. When novel social behaviour is re-introduced for the post-test, the engagement level then rises again, in agreement with [52].

6.3 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 4.8, 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" [6]. 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 7). 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 7.2.

Of particular interest is the behaviour of the child during the main learning phase. A one-way between subjects

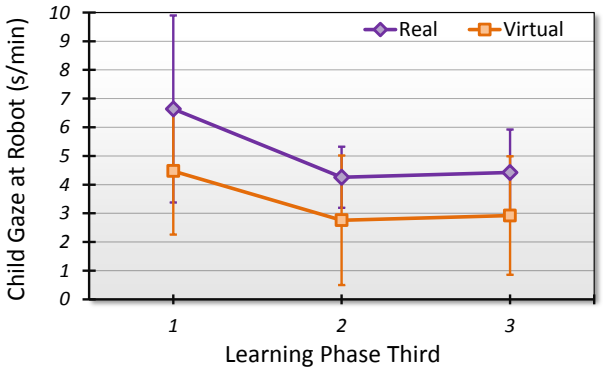

Fig. 6 Seconds per minute that the child spends gazing towards the robot, split by learning phase third, comparing embodiment conditions. Error bars show standard deviation.

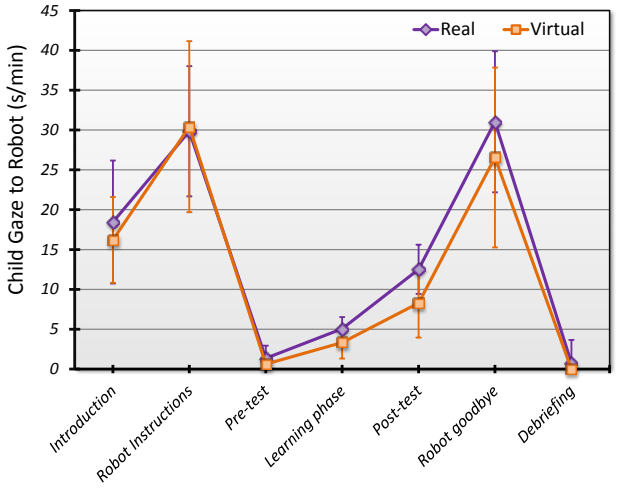

Fig. 7 Seconds per minute that the child spends gazing towards the real and virtual robots, split by interaction segment. Error bars show standard deviation.

ANOVA was conducted to compare the effect of interaction time on the child's gaze towards the robot (both conditions combined). The learning phase was split into thirds for comparison [6]. There was a significant effect of interaction segment on child gaze towards the robot at the $p < .05$ level for the three thirds [$F(2,81)=6.968$, $p=0.002$]. Post-hoc comparisons using the Tukey HSD test indicate that the mean score for the first segment ($M=5.63$, $SD=2.98$) was significantly different to the second segment ($M=3.56$, $SD=1.85$) and the third segment ($M=3.73$, $SD=1.91$), with p values of 0.003 and 0.008 respectively. No significant difference was found between the second and third segments, $p=0.961$. Therefore, children look at the robot significantly more in the first third of the learning phase, before dropping for the rest of the interaction.

7 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 3.

7.1 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 [19, 29]. In this case these effects were not found; thereby not supporting Hypothesis 4. However, this is in line with some work, for example [18]. Given the context of the interaction, we suggest 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 contingent (Section 6.3). If the robot could exhibit richer social behaviours, more contingent on the behaviour of the child than at present, 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 contingent 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 [44] and [56].

7.2 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 [31], and the increased mobility of the real robot has been suggested as an explanation, as in [10].

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 [52]. 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, SD=4.2) appears to gaze at the child significantly more than the real robot (M=10.5, SD=4.2) when normalised to s/min, $t(26)$=3.029, p=0.005. 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 socially contingent 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 its 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 authors would still hypothesise 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 oriented, or required more 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 was previously highlighted as an interesting difference to discuss. Figure 7 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 1.01 seconds per minute (*s/m*) gazing towards the robot (SD=1.27); this rises very slightly to 1.31 s/m (SD=1.36) 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 24.99 s/m (SD=9.93) gazing towards the robot. This is almost half of the time and further supports the point made in Section 6.3 suggesting that the children gaze more towards the robot when it is directly addressing them and exhibiting novel behaviour.

Schermerhorn and Scheutz have demonstrated the complex interactions which occur between embodiment and other elements of robot behaviour [46]. Whilst the analysis of an

integrated system is always desirable, we would suggest that first varying just one of the dimensions at a time (either embodiment or behaviour) 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 and Mutlu adopt a multivariate analysis approach to study the impact of a specific behavioural variable, in this case gesture, on knowledge recall [19]. 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, due to the way that social cues may be perceived [57].

7.3 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 5.1. 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 [15, 37].

Furthermore, the unit of measure for learning and how learning is defined is important. Completion time of a puzzle has been used, as in [30]. 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. [47], [53]. 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 [32]. However, a side of 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 [54]. 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. In future work it will be ensured that any novel dataset created will be tested and validated to remove, or at least balance, any possible bias. Moreover, it is more likely that future work 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. It would be preferable to find 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.

8 Conclusion

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, a 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 contingent. This drop-off in apparent engagement signifies that the robot behaviour needs improvement, highlighted by Hypothesis 2 from Section 3 (that a socially contingent robot behaviour will keep the children engaged throughout the interaction).

In support of Hypothesis 3, regarding gaze and attention, children's compliance with the robot's requests (as shown in Section 6) 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., [4], [23].

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, 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 paper 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. This learner-centered, 'guided discovery learning' approach requires a very different model of robot behaviour when compared with the teacher-centered approaches previously applied and investigated in HRI. 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 socially contingent behaviour for the robot to maintain the initially high levels of attention from the child throughout the interaction. Nevertheless, the results provided initial support for the effective application of the guided discovery learning methodology to cHRI.

Acknowledgements This research was partially funded by the EU FP7 ALIZ-E project (FP7-ICT-248116), the FP7 DREAM project (FP7-ICT-611391) and the School of Computing and Maths, Plymouth University. The authors would like to thank Salisbury Road Primary School in Plymouth, U.K. for hosting the study.

Gratitude also goes to Robin Read who provided assistance conducting the experiment, and to John Radnor and Marina Khalil for second coding of the videos.

References

1. Alfieri L, Brooks PJ, Aldrich NJ, Tenenbaum HR (2011) Does discovery-based instruction enhance learning? Journal of Educational Psychology 103(1):1
2. Bainbridge WA, Hart J, Kim ES, 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, IEEE, pp 701–706, DOI 10.1109/ROMAN.2008.4600749
3. Bartneck C (2003) Interacting with an embodied emotional character. In: Proceedings of the 2003 International Conference on Designing Pleasurable Products and Interfaces, ACM, pp 55–60
4. 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 ACM/IEEE International Conference on Human-Robot Interaction, ACM, pp 105–106, DOI 10.1145/2157689.2157707
5. 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, IEEE, pp 77–78
6. Baxter P, Kennedy J, Vollmer AL, de Greeff J, Belpaeme T (2014) Tracking gaze over time in hri as a proxy for engagement and attribution of social agency. In: Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction, ACM, pp 126–127
7. 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
8. 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
9. Berry D, Butler L, de Rosis F (2005) Evaluating a realistic agent in an advice-giving task. International Journal of Human-Computer Studies 63(3):304–327
10. 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, DOI 10.1109/TSMCC.2004.826268
11. Chi MT, Feltovich PJ, Glaser R (1981) Categorization and representation of physics problems by experts and novices*. Cognitive science 5(2):121–152
12. Crone EA, Jennings JR, Van der Molen MW (2004) Developmental change in feedback processing as reflected

by phasic heart rate changes. Developmental psychology 40(6):1228
13. De Jong T, Van Joolingen WR (1998) Scientific discovery learning with computer simulations of conceptual domains. Review of Educational Research 68(2):179–201
14. Dimitrov DM, Rumrill PD Jr (2003) Pretest-posttest designs and measurement of change. Work: A Journal of Prevention, Assessment and Rehabilitation 20(2):159–165
15. Guskey TR (2003) How classroom assessments improve learning. Educational Leadership 60(5):6–11
16. Han J, Jo M, Park S, Kim S (2005) The educational use of home robots for children. In: Proceedings of the 14th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) 2005, IEEE, pp 378–383
17. Harvel C (2010) Guided discovery learning. In: Lee H (ed) Faith-Based Education That Constructs: A Creative Dialogue between Contructivism and Faith-Based Education, Wipf and Stock Publishers, pp 169–172
18. 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
19. Huang CM, Mutlu B (2013) Modeling and evaluating narrative gestures for humanlike robots. In: Proceedings of the Robotics: Science and Systems Conference, RSS 2013, pp 26–32
20. Kalyuga S (2008) Relative effectiveness of animated and static diagrams: An effect of learner prior knowledge. Computers in Human Behavior 24(3):852–861
21. 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
22. 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, ACM, pp 351–358
23. 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, IEEE Computer Society, pp 728–733
24. Kennedy J, Baxter P, Belpaeme T (2014) Children comply with a robot's indirect requests. In: Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction, ACM, pp 198–199, DOI 10.1145/2559636.2559820
25. Kose-Bagci H, Ferrari E, Dautenhahn K, Syrdal DS, Nehaniv CL (2009) Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. Advanced Robotics 23(14):1951–1996
26. Kuhlthau C, Maniotes L, Caspari A (2007) Guided inquiry: Learning in the 21st century. Greenwood Publishing Group
27. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159–174
28. Leite I, Martinho C, Paiva A (2013) Social robots for long-term interaction: A survey. International Journal of Social Robotics 5(2):291–308, DOI 10.1007/s12369-013-0178-y
29. 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
30. 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, ACM, pp 423–430
31. Looije R, van der Zalm A, Neerincx MA, Beun RJ (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, IEEE, pp 718–724, DOI 10.1109/ROMAN.2012.6343836
32. Lupyan G, Rakison DH, McClelland JL (2007) Language is not just for talking: Redundant labels facilitate learning of novel categories. Psychological Science 18(12):1077–1083
33. Macario JF (1991) Young children's use of color in classification: Foods and canonically colored objects. Cognitive Development 6(1):17–46
34. Merrill DC, Reiser BJ, Merrill SK, Landes S (1995) Tutoring: Guided learning by doing. Cognition and Instruction 13(3):315–372
35. Moshkina L, Trickett S, Trafton JG (2014) Social engagement in public places: a tale of one robot. In: Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction, ACM, pp 382–389
36. Mutlu B, Forlizzi J, Hodgins J (2006) A storytelling robot: Modeling and evaluation of human-like gaze behavior. In: Proceedings of the 6th IEEE-RAS International Conference on Humanoid Robots 2006, IEEE, pp 518–523
37. Myers CB, Myers SM (2007) Assessing assessment: The effects of two exam formats on course achievement and evaluation. Innovative Higher Education 31(4):227–236
38. 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: Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive

Communication (RO-MAN) 2012, IEEE, pp 351–357
39. 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 Joint Conference on Autonomous agents and multiagent systems, International Foundation for Autonomous Agents and Multiagent Systems - Volume 3, pp 1253–1256
40. 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, IEEE, pp 145–152
41. Riek LD (2012) Wizard of oz studies in hri: a systematic review and new reporting guidelines. Journal of Human-Robot Interaction 1(1)
42. Riggio RE, Friedman HS (1986) Impression formation: The role of expressive behavior. Journal of Personality and Social Psychology 50(2):421–427
43. Roediger HL, Karpicke JD (2006) The power of testing memory: Basic research and implications for educational practice. Perspectives on Psychological Science 1(3):181–210
44. 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, ACM, pp 335–342
45. Saerbeck M, Schut T, Bartneck C, Janse MD (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, ACM, New York, NY, USA, CHI '10, pp 1613–1622, DOI 10.1145/1753326.1753567, URL http://doi.acm.org/10.1145/1753326.1753567
46. 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
47. Scott SD, Sheelagh M, Carpendale T, Inkpen KM (2004) Territoriality in collaborative tabletop workspaces. In: Proceedings of the 2004 ACM conference on Computer Supported Cooperative Work, ACM, pp 294–303
48. Segura EM, Cramer H, Gomes PF, 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, ACM, New York, NY, USA, pp 88–97, DOI 10.1145/2307096.2307107
49. Smith III JP, Disessa AA, Roschelle J (1994) Misconceptions reconceived: A constructivist analysis of knowledge in transition. The Journal of the Learning Sciences 3(2):115–163, DOI 10.1207/s15327809jls0302_1
50. Spencer JA, Jordan RK (1999) Learner centred approaches in medical education. BMJ: British Medical Journal 318(7193):1280–1283
51. Stanton CM, Kahn PH, Severson RL, Ruckert JH, Gill BT (2008) Robotic animals might aid in the social development of children with autism. In: Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction, IEEE, pp 271–278
52. 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, ACM, New York, NY, USA, CHI'12, pp 11–20, DOI 10.1145/2207676.2207679, URL http://doi.acm.org/10.1145/2207676.2207679
53. 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, ACM, pp 1181–1190
54. Tessmer M, Richey RC (1997) The role of context in learning and instructional design. Educational Technology Research and Development 45(2):85–115
55. Van Joolingen W (1998) Cognitive tools for discovery learning. International Journal Of Artificial Intelligence In Education (IJAIED) 10:385–397
56. Walters ML, Woods S, Koay KL, Dautenhahn K (2005) Practical and methodological challenges in designing and conducting human-robot interaction studies. In: Proceedings of the AISB, vol 5, pp 110–119
57. Zaki J (2013) Cue integration a common framework for social cognition and physical perception. Perspectives on Psychological Science 8(3):296–312

Appendix A: Robot Script

Below is a list of the robot scripted phrases and where they occur in the interaction.

- 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!

James Kennedy received a B.Sc. (Hons) in Music Systems Engineering and an M.Sc. in Information Technology from the University of the West of England (U.K.) in 2010 and 2012, respectively. He is currently studying for a PhD in Human-Robot Interaction at Plymouth University (U.K.). His research interests centre around social companion robots, particularly for use in educational interactions with children. He has previously been involved with the EU FP7 ALIZ-E project and is currently working alongside the EU FP7 DREAM project.

Paul Baxter is a Research Fellow in the Centre for Robotics and Neural Systems at Plymouth University (U.K.). He completed his PhD at the University of Reading (U.K., 2010) in the domain of developmental cognitive robotics, emphasising a memory-centred perspective on cognition that has formed the basis of his subsequent work in Human-Robot Interaction. His current work is part of the EU FP7 DREAM project, on the cognitive and behavioural aspects of a robot for the supervised-autonomy robot-enhanced therapy of autistic children. He previously worked on the EU FP7 ALIZ-E project, developing a distributed memory approach to support long-term social human-robot interaction.

Tony Belpaeme received his PhD in Computer Science from the Vrije Universiteit Brussel in 2002 and is currently Professor in Robotics and Cognitive Systems at Plymouth University (United Kingdom) where he leads a research lab in the Centre for Robotics and Neural Systems. Starting from the premise that cognition is rooted in social interaction, Belpaeme and team try to further the science and technology behind artificial intelligence and social robots. This results in a spectrum of findings, from theoretical insights to practical applications. He coordinated the FP7 ALIZ-E project, and collaborated on the ROBOT-ERA, DREAM and ITALK projects.