doi:10.20944/preprints202012.0061.v1

Article

Robot Tutoring Multiplications: Over one-third Learning Gain for Most, Learning Loss for Some

Johan F. Hoorn [1,*], Ivy S. Huang [3], Elly A. Konijn [2] and Lars van Buuren [2]

[1] The Hong Kong Polytechnic University, Hong Kong SAR; johan.f.hoorn@polyu.edu.hk
[2] Vrije Universiteit, Amsterdam; e.a.konijn@vu.nl
[3] The Education University of Hong Kong, Hong Kong SAR; isihuang@eduhk.hk
* Correspondence: johan.f.hoorn@polyu.edu.hk; Tel.: +528-2766-4509 (J.F.H.)

Abstract: In the design of educational robots, it seems undecided whether robots should show social behaviors and look human-like or that such cues are indifferent to learning. We conducted an experiment with different designs of social robots, rehearsing the multiplication tables with primary school children in Hong Kong. Results show that affective bonding tendencies may occur but did not significantly contribute to the learning progress of these children, perhaps due to the short interaction period. Nonetheless, 5 minutes of robot tutoring improved their scores with about 30% and only for a few challenged children, performance dropped. We discuss that topics such as teaching language skills may be fostered by human likeness in appearance and behaviors but that for STEM-related subjects, the social aspects of robots hardly matter.

Keywords: **r**obot-tutelage; social robots; multiplications; experience design

1. Introduction

With the current pandemic of COVID-19, learners world-wide rely on online teaching and media applications for their education. Nonetheless, the United Nations fear for knowledge deficits, learning losses, and gaps in the learning process from a lack of face-to-face interaction (UN, 2020, p. 4, p. 23). Therefore, the UN plead for different ways of content delivery: Hybrid learning that is flexible and quasi-individualized (UN, 2020, p. 25): “We should seize the opportunity to find new ways to address the learning crisis and bring about a set of solutions previously considered difficult or impossible to implement” (UN, 2020, p. 4). If every child had a robot tutor at home, would this to some extent make up for missing out on human interaction?

Whereas a few years ago, robot teachers were mere science fiction, today a number of schools include some form of robot education. This varies from educational programs such as Science, Technology, Engineering, and Mathematics (STEM) in which young children learn to build and program robots (e.g., Gomoll, Šabanović, Tolar, Hmelo-Silver, Francisco, & Lawlor, 2017; STEMex, 2019) to humanoids that teach children mathematics or language (e.g., Chang, Lee, Chao, Wang, & Chen, 2010; Nuse, 2017). Multiple studies show that robots can be beneficial for learning outcomes. A recent review points out that appearance, behavior, and different kinds of social roles of the robot may positively affect learning outcomes but sometimes also negatively (Belpaeme, Kennedy, Ramachandran, Scassellati, & Tanaka, 2018).

It seems that people go better by instructions forwarded by a social robot than by a tablet with the same programs and voice (e.g., Mann, MacDonald, Kuo, Li, & Broadbent, 2015). Pupils apparently learn significantly more from their robotic tutors than from a tablet or no robot at all (VanLehn, 2011; Brown, Kerwin, & Howard, 2013).

Common understanding has it that in human-human teaching, warm, social, and personal teachers are more successful in advancing their pupils’ level of study performance (e.g., Hattie, 2015; Tiberius & Billson, 1991; Saerbeck, Schut, Bartneck, & Janse, 2010). In human teacher-student relationships, a teacher should not just offer theoretical instruction and correct mistakes but also support students personally while creating a healthy relationship (e.g., Frymier & Houser, 2000;

Hattie, 2015; Skinner & Belmont, 1993). Hamre and Pianta (2001) emphasized that a positive relationship with a teacher makes a child more willing to take on an academic challenge or work on its social-emotional development.

Many researchers expect to find that robots that show more personalized, pro-social behaviors also render better learning results (e.g., Kory-Westlund & Breazeal, 2019; Alves-Oliveira, Tullio, Ribeiro, & Paiva, 2014; Atkinson, Mayer, & Merrill, 2005; Boucher et al., 2012; Esposito, 2011). However, robot researchers have attempted various forms of social interaction and communicative behaviors but as a result obtained a blend of advantageous and unfavorable effects on learning (e.g., Belpaeme et al., 2018; Konijn, Smakman, & Van den Berghe, 2020). It seems that individual differences such as educational ability levels are sensitive to the level of a robot's social behaviors: Certain students seem to flourish with a more a neutral approach (Konijn & Hoorn, 2020b; Konijn, Smakman, & Van den Berghe, 2020, p. 6).

Another point for the mixed results so far may be the topic that is taught. Robots (as tutors) are employed more frequently in non-STEM subjects such as language (e.g., Kennedy, Baxter, Senft, & Belpaeme, 2016; Vogt, Haas, Jong, Baxter, & Krahmer, 2017). In language-related topics such as vocabulary learning or remembering the story line, social behaviors seem to be more beneficial to learning than neutral styles of teaching. For reading aloud a narration from a picture book that featured fictional characters, for example, facial expressions were important to bring the characters to life so the children performed better on story recall and target vocabulary (Kory-Westlund, Jeong, Park Hae, Ronfard, Adhikari, Harris, DeSteno, & Breazeal, 2017). In teaching vocabulary during a storytelling game, cuddly toy robots that tended to the child's oral language skills were more successful than robots that did not (Kory-Westlund & Breazeal, 2014).

In arithmetic and mathematics, such social aspects may play less of a role (e.g., Hindriks & Liebens, 2019; Konijn & Hoorn, 2020b). In STEM related topics, a robot's social behaviors such as greeting, following gaze, motivational feedback, and humanoid appearance do not seem to matter too much (e.g., Hindriks & Liebens, 2019; Saerbeck, Schut, Bartneck, & Janse, 2010) or may exert adverse effects (e.g., Kennedy, Baxter, & Belpaeme, 2015). Moreover, robots appear to be successful at maintenance rehearsal and repeated exercise (e.g., Wei, Hung, Lee, & Chen, 2011; Huang & Hoorn, 2018). In other words, if students are to practice multiplication tables as a kind of remedial teaching, perhaps more social behaviors of the robot tutor may be insignificant or even distracting (Kennedy, Baxter, & Belpaeme, 2015; Konijn, Smakman, & Van den Berghe, 2020).

Yet, in the on-screen community of virtual tutors and avatars, researchers do report positive effects of building rapport while learning STEM. For example, a virtual agent was most successful in supporting STEM learning when it showed rapport behavior (Krämer, Karacora, Lucas, Dehghani, Rüther, & Gratch, 2016). Although learners were not aware of the increased rapport, the agent that showed rapport fostered better performance (ibid.). Arroyo, Royer, and Park Woolf (2011) reported that during basic math operations, their adaptive Wayang Tutoring System embodied by an affective learning companion improved the students' working memory and math fluency (the speed to recover or compute answers).

The theory of affective bonding (Konijn & Hoorn, 2017; 2020a) also would expect that stronger bonding of the learner with the robot enhances learning performance. The affective bond would be fed by the relevance of the robot to the task (here, learning multiplications) and by the robot's 'affordances' or action possibilities (cf. Jamone et al., 2018) to execute that task. On the more affective side, emotional bonding would be nurtured by a realistic, human-like embodiment and human-like behaviors (cf. 'anthropomorphism').

In the design literature, the importance assigned to realistic anthropomorphic design can hardly be overstated (e.g., Syrdal, Dautenhahn, Woods, Walters, & Koay, 2007; Köse et al., 2015). For instance, Moshkina, Trickett, and Trafton (2014) reported that more humanlike features in a robot, such as a voice, a face, and gestures, invoked more engagement with its audience. Nonetheless, Li, Rau, and Li (2010) suggest that dependent on someone's cultural background, a robot's appearance may exert different levels of likeability, engagement, trust, and satisfaction. From their empirical work, Paauwe, Hoorn, Konijn, and Keyson (2015) concluded that the perceived realism of a robot's embodiment played a modest role in intentions to use the robot and feeling engaged with it. In robot design also, realism is not all (Van Vugt, Konijn, Hoorn, Eliëns, & Keur, 2007).

To study the effects of robot tutoring on learning a STEM related task such as rehearsing multiplications, we varied different forms of human-likeness in the design of the robot (cf. Syrdal, Dautenhahn, Woods, Walters, & Koay, 2007). Our hypothesis (H1) was that we expected positive effects of a more humanlike design on rehearsing the multiplications.

As our H2, we presumed that working with a robot tutor potentially would be more beneficial to lower-ability pupils than for advanced students. For below-average students, larger progress may be achieved whereas for the high performers, the added value may be minimal.

From Konijn and Hoorn (2017; 2020a), one could infer that robot tutoring improves learning the multiplications the more the child emotionally bonds with the robot tutor. Bonding is stimulated when the robot's design looks and behaves like a human and in the perception of the child is experienced as high on anthropomorphism, relevance, realism, and affordances. Therefore, H3 supposed that building rapport or establishing an emotional bond with the robot would lead to better task performance, perhaps in a mediating or moderating manner. As a control, we queried the social role the robot played for these children (cf. Chen, Park Hae, & Breazeal, 2020) and how appealing ('beautiful') and new they felt their robot tutors were.

2. Materials and Methods

2.1. Participants and Design

After obtaining approval from institutional IRB (lbn344; April 1, 2019; FSW Research Ethics Review Committee - RERC), parental consent letters were distributed through two Hong Kong primary schools. Due to strict time planning by the schools and because parents picked up their children early, eventually 75 students were able to participate in at least one session with a robot tutor and completed the pre- and post-test (N = 75; M_{Age} = 8.4, SD_{Age} = .82, range: 7-10, 44% female, Hongkongers). For more details on demographics, consult the technical report in Supplementary Materials 1.

We planned for all pupils to participate in 3 robot tutoring sessions spread over three weeks (within-subjects). Due to the schools' tight time schedules, however, not every pupil could participate in every session. Children from the S.K.H. Good Shepherd Primary School only took one session. This number plus those from the Free Methodist Bradbury Chun Lei Primary School that took but one session, resulted into 48 children participating only once. Those who participated twice (n = 13), and thrice (n = 14) were all from Chun Lei.[1] For a complete overview of the participatory division, consult the technical report in Supplementary Materials 1.

To test our hypotheses, we administered an experiment with the between-subjects factors of Robot Design (3) and Advancement Level (4) to measure their effects on the within-subjects scores at the multiplication test, before-and-after robot tutoring. We also examined the mediating or moderating effects of affective Bonding with the robot on learning the multiplications. We invited the children to participate in three sessions with the tutoring robot.

Participants (N = 75) were randomly distributed over three different Robot Designs (between-subjects): Humanoid (n = 21), Puppy (n = 27), and Droid (n = 27) (Figure 1). A Chi-square test of independence checked for the distribution of Age over robot types but no significant relationship was found ($\chi^2_{(6)}$ = 1.76, p = .94).

[1] Those participating twice or thrice were different children.

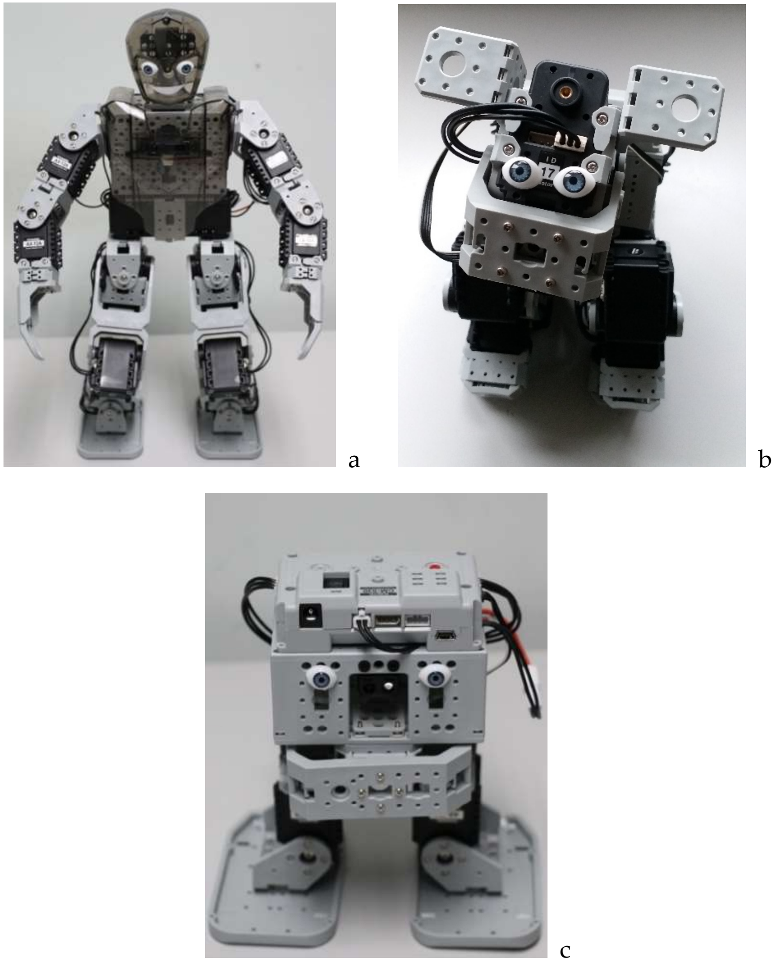

Figure 1. Robotis' Bioloids:[2] Humanoid (a), Puppy (b), and Droid (c).

Boys and girls were distributed over the Robot Design conditions as follows: Humanoid (15 males, 6 females), Puppy (15 males, 12 females), and Droid (12 males, 15 females). The schools' strict time scheduling caused unequal distributions of Gender over the three robots but this did not render a significant effect ($\chi^2_{(2)} = 3.49$, $p = .174$).

To determine the Advancement Level of the pupils, we took the average Baseline score ($N = 75$, $M = 37.16$, $SD = 12.88$) established in the pre-test and categorized the children into four groups for further exploration. Those who scored lower than one standard deviation below average (Baseline ≤ 22.28) were categorized as 'Challenged' students ($n = 11$). Those between one negative standard deviation and the average were categorized as 'Below average' (22.8 < Baseline ≤ 37.16) ($n = 34$). Those between average and one positive standard deviation were categorized as 'Above average' (37.16 < Baseline ≤ 52.04) ($n = 19$), and those beyond one positive standard deviation were categorized as 'Advanced' students (Baseline > 52.04) ($n = 11$). No significant effect of unequal distributions was found between Advancement Level and Robot Design ($\chi^2_{(6)} = 1.73$, $p = .943$). For more details, see the technical report in Supplementary Materials 1.

2.2. Procedure

At the Free Methodist Bradbury Chun Lei Primary school, the experiment took place during three weeks on every Tuesday. The S.K.H. Good Shepherd Primary school had time for but one

[2] http://www.robotis.us/robotis-premium/

session. In class, the topic and procedure was introduced and pupils took a 5 minute multiplication pre-test of 147 equations (Table 1, Figure 2). One week later, after class, the pupils from Chun Lei were asked to wait in the corridor before entering the experiment classroom (Figure 3).

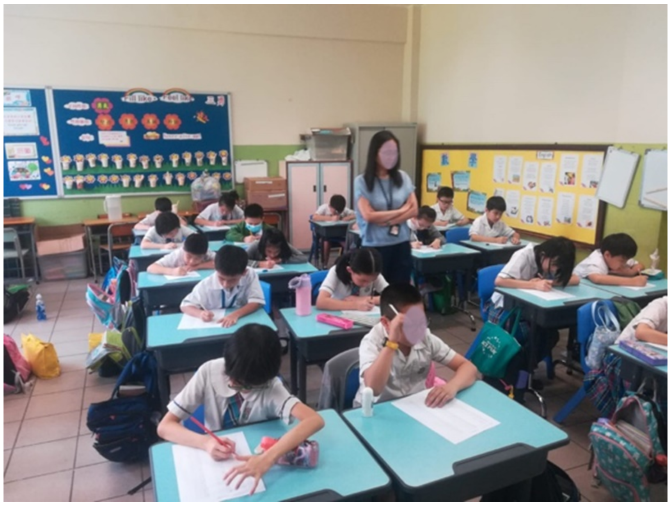

Figure 2. Set up for pre- and post-test.

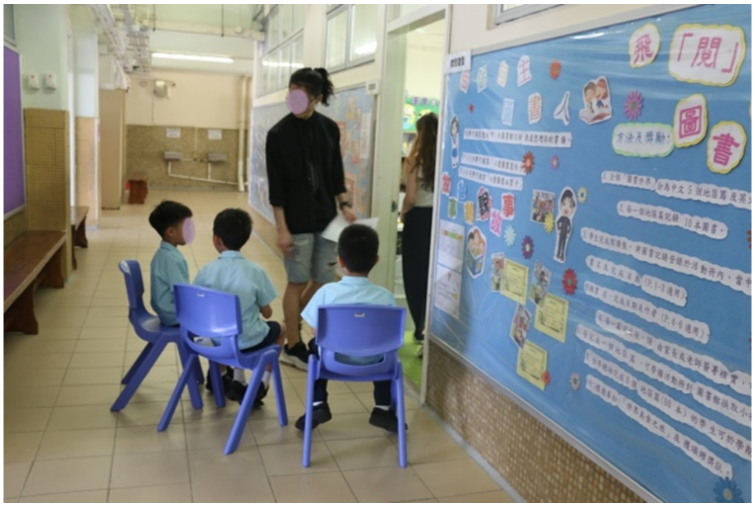

Figure 3. Waiting area at Chun Lei.

Those from Good Shepherd were taken out of class one at a time by one of the research assistants and entered the experiment room upon arrival. When one of the pupils of either school entered the room, they were brought by one of the assistants to the table where the robot stood (Figure 4). With three Bioloid robots available, three children were tutored simultaneously such that they did not disturb each other.

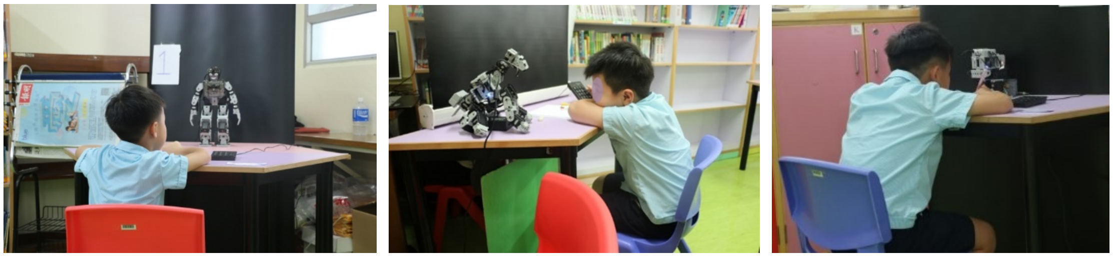

Figure 4. Experimental set-up. Humanoid left, Puppy middle, Droid right.

The assistant explained that the robot would ask a question and that the pupil could answer through the numpad and pressing Enter afterwards (Figure 4). All interactions, tests, and questionnaires were recorded in Cantonese. The robot started the session by asking if the pupil was ready. Upon confirmation, the multiplication program started, automatically drawing 147 equations randomly from various multiplication tables. Equations consisted of one-digit numbers times two-digit numbers (see Table 1). Questioning went on for 5 minutes, after which the program thanked the child, reported on the number of correct answers, and dismissed the pupil from the session. After one and after two weeks, the same procedure was repeated (at Chun Lei).

The three assistants that operated the robots were sitting behind a curtain. This way, the pupil had the illusion that the robot was fully autonomous while in fact someone was pressing buttons on a remote control. The answers that participants typed in on the numpad could be read by the assistant. When the answer was correct, the assistant pressed the button that triggered positive feedback such as clapping or nodding; when incorrect, the assistant pressed the button that triggered feedback about the mistake such as shaking the head or head scratching (對不起。那是不對的。“I am sorry. That is not right”).

Each time the pupils completed their sessions, they took another multiplication test as post-test (once, twice, thrice). The same procedure as in the pre-test was used. After the post-test, pupils filled out a questionnaire about their experiences with the robot. At Chun Lei, the questionnaire was a homework assignment; the pupils from the Good Shepherd did the questionnaire in class.

2.3. Apparatus and Materials

Humanoid, Puppy, and Droid (Figure 1) were built from three identical Bioloid Premium DIY kits and programmed on the same CM530 computer.[3] To tease out bonding tendencies, we put comparable eyes on the three machines (Figure 1) so each robot would ‘look’ at the participants. Attached to the Bioloids were Fresh ‘n Rebel Rockbox Cube Fabriq Army (59×59×59mm, Bluetooth 4.0, 1 channel mono 3W) front speakers that were connected to a self-written speech engine in Node.js (a Javascript framework), which ran independently from the robot software.

Trials consisted of prerecorded Cantonese male speech (23 years of age) of multiplication equations, for instance, “5 times 12?” and the child’s input was followed by various feedback such as “I’m sorry that is incorrect,” “Well done, that’s correct.” Trials were composed from separate audio files of the numbers 1 to 99, of the word “times,” and of “equals.” The program would then randomly select a number audio file, followed by the “times” audio file, followed by another random number audio file, followed by the “equals” audio file.

The speech program kept track of the pupil’s answers but motoric functions of the robot were controlled remotely because the speech program in Node.js was incompatible with the Robotis+ code language of the robot.[4] Therefore, a wireless Bluetooth receiver was attached to the robot’s computer, communicating with a wireless controller (Figure 5). Code can be found in Supplementary Materials 1.

[3] http://www.robotis.us/robotis-premium/

[4] https://nodejs.org/en/about/

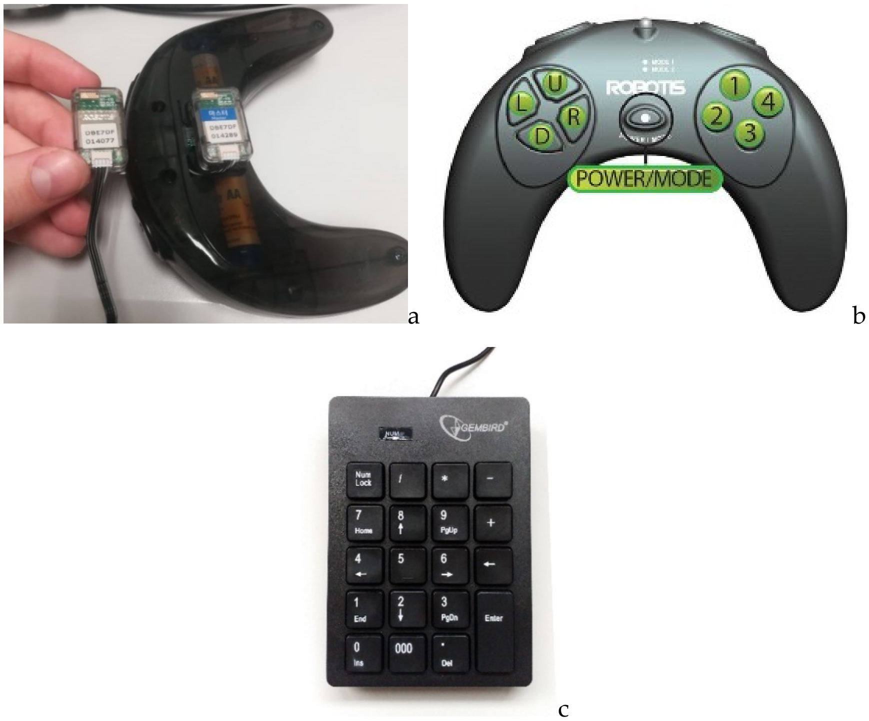

Figure 5. Wireless Bluetooth receiver (a), wireless controller (b), and numpad (c).

Pupils could input their answers on a Gembird numerical keyboard or numpad (OS independent, plug-and-play, 124×81×21mm, USB 2.0 powered with type A-plug) (Figure 5). Apart from audio feedback, a correct answer was rewarded with Humanoid clapping its hands, Puppy nodding its head, and Droid moving up and down. For negative feedback, Humanoid scratched its head, Puppy shook its head, and Droid wiggled from left to right.

The program terminated after 5 minutes, counted the number of correct answers, and based on the results played "Well done" or "I'm sorry." It then thanked the child for its participation and asked to leave the room.

2.4. Measures

Table 1 offers a synopsis of the variables investigated in this study. The full record of variables can be found in Supplementary Materials 1. Table 1 has two types of dependent measures that are theoretically relevant: learning and experience. Additionally, several control variables are tabulated as well.

Table 1. Overview of dependent variables.

Variable name	Description	Number of items	Abbreviation & Value
	Learning variables		
Baseline (from pre-test)	The scores in the pre-test, which established baseline. Pupils multiplied 1-or-2 digit numbers with 2-digit numbers from the range 1-99, most difficult equation being 23 × 67	147	Base [0,147]

doi:10.20944/preprints202012.0061.v1

Final score (from post-test)	Final multiplication score based on multiplying 1-or-2 digit with 2-digit numbers	147	FinMSco [0,147]
Learning gain	Difference between pre-test Baseline and post-test Final score. Also calculated as difference-score of FinMSco minus Baseline	147	Fin_min_Base [0,147]
Gain percentage	The percentage of Fin_min_Base compared with Baseline		Per_Fin_min_Base [min, max]
	Experiential variables		

Representation	What does the robot look like to the participant?	3	Human_like = [1,6]
			Animal_like = [1,6]
			Machine_like = [1,6]

Social role	What does the robot feel like to the participant?	7	Friend = [1,6]
			Classmate = [1,6]
			Teacher = [1,6]
			Acquaintance = [1,6]
			Stranger = [1,6]
			Machine = [1,6]
			Other = [1,6]

Bonding	How is the social-affective relationship between participant and robot tutor? (Konijn & Hoorn, 2017; 2020a)	5	Bon_1…5 = [1,6]
Anthropomorphism	Does participant attribute human traits or emotions to robot tutor? (Konijn & Hoorn, 2017; 2020a)	4	Anth_1…4 = [1,6]
Perceived realism	Does robot tutor feel like a real creature or is it a fake? (Paauwe et al., 2015)	4	Real_1…4 = [1,6]
Perceived relevance	Is robot tutor significant for doing the multiplication exercise? (Konijn & Hoorn, 2017; 2020a)	4	Rel_1…4 = [1,6]
Perceived affordances	What can I do with the robot (in view of the multiplication exercise)? (Konijn & Hoorn, 2017; 2020a)	4	Aff_1…4 = [1,6]
Engagement	Level of involvement with the robot	5	Eng_1…5 = [1,6]
Use intentions	Want to use the robot again?	3	Use_Int_1...3= [1,6]
	Controls		
Novelty	To what extent is the robot tutor new to the participant?	1	Nov_1 = [1,6]
Aesthetics	To what extent is the robot attractive to the participant in terms of	1	Aest_1 = [1,6]

appearance?		
Gender	1	Gender = [Male, Female]
Age	1	Age = [7,10]

Learning variables were derived from pre- and post-test in which pupils solved 147 equations drawn from the range [1-99] with the second number always having two digits (e.g., 3 × 12 or 15 × 31). In the analysis, our focus will be on Learning gain (the absolute difference between pre- and post-test) and Gain percentage (learning gain relative to a child's baseline knowledge).

We created the measure of Gain percentage because, for example, 5 more correct answers after robot tutoring may be a relatively big gain for those who performed poorly before but a small gain for those who already performed at a high level (cf. ceiling effect). Percentage_Fin_min_Base, then, was calculated as Fin_min_Base divided by Baseline (Table 1).

The **experiential** variables were measured by a 43-item paper-and-pencil structured questionnaire that was filled out after pupils completed their tutoring session(s) (Appendix A). Indicative and counter-indicative Likert-type items were scored on a 6-point rating scale (1 = totally disagree, 6 = totally agree). The counter-indicative items on the questionnaire were recoded into new variables, after which we calculated Cronbach's α for all scales followed by Principal Component Analysis (PCA). From the remaining items, we calculated Cronbach's α again.

Representation. To check the manipulation with the three different Robot Designs, participants rated to what degree they felt the design of their robot represented a human being, an animal, and a machine. All three dimensions were rated for each robot. In addition, they evaluated the *Social role* of the robot (e.g., a friend, a teacher, etc.).

Bonding was measured with 5 items (i.e. bond, interested, connected, friends, understand). Two examples of indicative items are "I felt a bond with the robot" and "The robot understands me." Cronbach's $\alpha = .88$.

Anthropomorphism contained 4 items (machine, talk like human, humanlike reaction, humanlike interaction). Two examples are: "It felt just like a human was talking to me" and "I reacted to the robot just as I react to a human." Only these two items were left after psychometric analysis: Spearman-Brown Correlation ($r = .68$, $p = .000$).

Perceived realism was based on Paauwe, Hoorn, Konijn, and Keyson (2015) and Van Vugt, Hoorn, Konijn, and De Bie Dimitriadou (2006). The scale had 4 items (real creature, like real, feels fabricated, real conversation), two examples of which are: "The robot resembled a real-life creature" and "It was just like real to me." Psychometric analysis indicated 3 items for sufficient reliability: Cronbach's $\alpha = .75$.

Perceived relevance was based on Van Vugt, Hoorn, Konijn, and De Bie Dimitriadou (2006) and consisted of four items (important, help, useless, need). Two examples are: "The robot was important to do my exercises" and "The robot is what I need to practice the multiplication tables." With four items, Cronbach's $\alpha = .73$.

Perceived affordances also was based on Van Vugt, Hoorn, Konijn, and De Bie Dimitriadou (2006) (immediately clear, took a while, puzzled). Two examples are: "I understood the task with the robot immediately" and "The robot was clear in its instructions." Only these two items achieved just sufficient reliability ($r = .61$, $p = .000$).

Engagement was included in addition to Bonding and was measured based on two scales by Paauwe, Hoorn, Konijn, and Keyson (2015) and Van Vugt, Hoorn, Konijn, and De Bie Dimitriadou (2006). Engagement was constructed from 5 items (e.g., like, dislike, feeling uncomfortable, fun). Examples are "I like the robot" and "I felt uncomfortable with the robot." Cronbach's $\alpha = .79$.

Use intentions also was based on Van Vugt, Hoorn, Konijn, and De Bie Dimitriadou (2006). It consisted of 3 items (use again, another time, help again), an example being: "I would use the robot again." Cronbach's $\alpha = .63$, which is just sufficient for group comparisons.

Control variables were single items, pertaining to Novelty (“Have played with robots before”), Aesthetics (“The robot looked beautiful”), Age, and Gender.

2.4.1. Principal component analysis

In a 7- and a 5-factor solution, divergent validity of the questionnaire items was weak and the only scale having good measurement quality overall, clearly distinguishable from other components, was Bonding (5 items, Cronbach’s α = .88), which will be the experiential measure we use for further analysis. For in-depth PCA analysis, consult Supplementary Materials 1.

3. Results

3.1. Preliminary analyses

To check the Robot Design manipulation, participants rated the extent to which they believed their robot resembled a human, an animal, and a machine (i.e. Human-like, Animal-like, and Machine-like). We ran a General Linear Model Multivariate Analysis (MANOVA) of Robot Design (3) on the Representation ratings of Human-like, Animal-like, and Machine-like. Pupils judged their robots as significantly different in what they represented: The effects of Robot Design on the rating of Representation was significant (Wilks’ λ = .57, $F_{(6,134)}$ = 7.17, p < .000, $\eta_p{}^2$ = .24). Significant effects were found for Human-like ($F_{(2,69)}$ = 8.32, p = .001) and Animal-like ($F_{(2,69)}$ = 12.41, p = .000). Thus, the robots did not differ in their machine-likeness but they did differentiate according to their representation of a human being or an animal.

Six two-tailed independent t-tests of Robot Design (Humanoid-Puppy, Humanoid-Droid, and Puppy-Droid) on ratings of Human-like and Animal-likeness showed that Human-likeness of the Humanoid robot (n = 19, M = 3.89, SD = 1.91) was significantly higher than that of Puppy (n = 26, M = 1.88, SD = 1.42) ($t_{(43)}$ = 4.05, p = .000). Human-likeness of Humanoid (n = 19, M = 3.89, SD = 1.91) also was significantly higher than that of Droid (n = 27, M = 2.26, SD = 1.79) ($t_{(44)}$ = 2.97, p = .005). Human-likeness of Puppy (n = 26, M = 1.88, SD = 1.42) and that of Droid (n = 27, M = 2.26, SD = 1.79) did not significantly differ ($t_{(51)}$ = -.84, p = .40). Animal-likeness of Humanoid (n = 19, M = 1.95, SD = 1.58) was significantly lower than that of Puppy (n = 26, M = 4.23, SD = 1.82) ($t_{(43)}$ = -4.39, p = .000). The Animal-likeness of Humanoid (n = 19, M = 1.95, SD = 1.58) and that of Droid (n = 27, M = 2.22, SD = 1.78) did not significantly differ ($t_{(44)}$ = -.54, p = .59) but the Animal-likeness of Puppy (n = 26, M = 4.23, SD = 1.82) was significantly higher than that of Droid (n = 27, M = 2.22, SD = 1.78) ($t_{(51)}$ = 4.06, p = .000). Therefore, Humanoid was rated as more human-like and Puppy was more animal-like, whereas for Droid, no differences were significant. Thus, all robots were machine-like with Droid as the starting point, while Puppy added an animalistic and Humanoid a more humanlike impression.

As an extra control on the manipulation, we asked the pupils if they experienced the robot as a classmate, a teacher, a tutor, and other Social Roles. We ran three GLM Multivariate Analyses (MANOVA) of Social Role (Friend, Classmate, Teacher, etc.) on Human-like, Animal-like, and Machine-like as separate dependents so that effects would become significant easily. However, the different Social Roles were not significant for Human-likeness ($F_{(30,246)}$ = .94, p = .563) and had no significant effect on Animal-likeness ($F_{(30,246)}$ = 1.18, p = .246). The different Social Roles *were* significant for Machine-likeness ($F_{(30,246)}$ = 1.75, p = .012): Between-subject effects indicated that the effect of Teacher ($F_{(5,66)}$ = 2.75, p = .026) and the effect of Machine ($F_{(5,66)}$ = 5.53, p = .000) on Machine-likeness were significant. However, there were six dependent variables in the analysis so that the rejection area α should be corrected, according to Bonferroni (.05 / 6 = .0083). Hence, only the categorization as Machine ($F_{(5,66)}$ = 5.53, p = .000) exerted significant effects on Machine-likeness, indicating that students perceived a machine-like robot indeed as a machine.

To check on possible confounding effects of non-theoretical variables, we ran a School (2) × Gender (2) ANCOVA on the Baseline score from the pre-test with Age as a covariate (N = 75). The only significant difference was caused by Age ($F_{(1,70)}$ = 4.35, p = .041) (r = .36, p = .002). With age, pupils performed better. School, Gender, and their interaction had no significant effect on Baseline performance. Only as isolated effects, while disregarding omnibus variance, did a two-tailed independent samples *t*-test show that the mean Baselines of Good Shepherd (n = 48, M = 39.71, SD = 15.85) and Chun Lei (n = 27, M = 32.63, SD = 11.94) significantly differed ($t_{(73)}$ = 2.02, p = .047) in favor of Good Shepherd. Likewise, while ignoring overall variance, the Baseline means of Boys (n = 42, M = 34.07, SD = 13.81) versus Girls (n = 33, M = 41.09, SD = 15.46) significantly differed ($t_{(73)}$ = -2.08, p = .042): Girls did more multiplications correct during the pre-test (not on the post-test after robot intervention as we shall see later). It seems that effects of School and Gender while significant on the detailed level (*t*-test) were spurious when more factors were added (*F*-test).

In a School (2) × Gender (2) ANCOVA on FinMSco with Age as a covariate (N = 75), none of the differences were significant. Although in an isolated correlation analysis, Age significantly affected the FinMSco (r = .24, p = .039), this relationship dissolved in the ANCOVA. Probably, the interaction with the robot countered the effect of Age on learning.

In addition, the correlation between Novelty and Fin_min_Base was not significant (r = .187, p = .12). Thus, novelty of the robot did not affect learning.

To explore the effects of the number of tutoring sessions on learning, we ran a number of tests with the factor Sessions (partaking once, twice, thrice). To see whether advancement level and number of sessions had an effect, we ran a GLM Univariate (ANCOVA) of Sessions (3) × Advancement Level (4) on Fin_min_Base with Age as a covariate. Yet, the interaction was not significant (F = .668).

We also conducted a One-way ANOVA of Sessions (participating once, twice, thrice) on Fin_min_Base without other variables involved but still no significant effects were established ($F_{(2,71)}$ = .866, p = .425). More robot-tutoring sessions did not improve learning performance any further.

Notwithstanding that there was not much difference among the groups that took one, two, or three tutorial sessions, yet, within each group, we wanted to know how big the learning gain was. We conducted three paired samples *t*-tests of Sessions on Baseline score versus FinMSco, representing the gain in absolute numbers and in percentages (Table 2).

Table 2. Mean improvement after robot tutoring once (N = 75), twice (n = 13), or thrice (n = 14).

Number of	$M_{Baseline}$	$M_{FinMSco}$	t	**Sig. (2-tailed)**	$M_{Fin\_min\_Base}$[a]	$M_{Per\_Fin\_min\_Base}$[b]
Sessions = 1	39.71	48.13	$t_{(48)}$ = -5.66	.000	8.42	21.20%
Sessions = 2	35.38	43.06	$t_{(16)}$ = -3.13	.007	7.68	21.70%
Sessions = 3	28.64	39.18	$t_{(11)}$ = -2.94	.015	10.54	36.80%

[a] Fin_min_Base = FinMSco – Baseline

[b] Per_Fin_Min_Base = Fin_min_Base / Baseline

Those who worked once with the robot improved by 8.42 more answers correct (21.20%). Those who did two sessions had a 7.68 improvement (21.73%) compared to Baseline. Those who interacted thrice had a 10.54 improvement (36.83%) compared to Baseline. Although at face value, three times tutoring seems to be better, later in the paper we see that Oneway ANOVA pointed out that statistically, the differences among the number of sessions were not significant.

3.2. Learning effects

H1 expected positive effects of Robot Design on learning with a significant advantage for Humanoid. H2 assumed differences in learning as a function of Advancement Level of the students, the Challenged students gaining significantly more from robot tutoring.

To test H1 and H2, we ran a GLM Repeated Measures of Robot Design (3) × Advancement Level (4) (between-subjects) on the (within-subjects) number of equations correctly solved before (Baseline) and after (Final Score) robot tutoring (N = 75). Note that this was the score in absolute numbers, not the percentage of gain relative to Baseline.

Our key finding was a significant and moderately strong main before-after effect on the absolute number of multiplications solved correctly (V = .50, $F_{(1,63)}$ = 62.43, p = .000, η_p^2 = .50). The mean score M_{Final} = 45.73 (SD = 17.40) was significantly larger than $M_{Baseline}$ = 37.16 (SD = 14.88) ($t_{(74)}$ = 7.19, p = .000), the mean difference being 8.57 equations more solved correctly after one session of robot tutoring, irrespective of Robot Design or Advancement Level.

Multivariate tests also showed a significant second-order interaction among Robot Design, Advancement Level, and before-after score (V = .22, $F_{(6,63)}$ = 2.99, p = .012, η_p^2 = .22). Inspection of the mean scores showed that the largest difference was established for Challenged pupils working with Humanoid ($M_{Baseline}$ = 16.33, SD = 6.03; M_{Final} = 41.67, SD = 17.93) and a small reverse effect was found for Advanced pupils, working with Droid ($M_{Baseline}$ = 69.33, SD = 5.52; M_{Final} = 68.00, SD = 18.61). Paired-samples t-test, however, showed that the effect for Challenged pupils working with Humanoid (n = 3) was not significant (not even preceding Bonferroni correction): $t_{(2)}$ = 3.51, p = .072; probably due to the large SDs and lack of power. No other main or interaction effects were significant (Supplementary Materials 1) except for the main effect of Advancement Level, which was a trivial finding obviously. H1 and H2 were refuted for learning gain in absolute numbers of correctly answered multiplications.

3.2.1. Learning gain (difference scores)

GLM Repeated Measures accounts for multiple sources of variance and is therefore the strictest test on our hypotheses. To assess if nothing was gained at all from Robot Design or Advancement Level, we included fewer sources of variance in our analysis from the reasoning that if lenient tests do not render significant effects either, we can dismiss Robot Design and Advancement Level from our theorizing altogether.

Therefore, we calculated the difference score from the Final Mean Score (FinMSco) – Baseline Score = Final_minus_Baseline (Fin_min_Base). Whereas 64 pupils gained from robot tutoring, there were 11 (about 15%) who did not perform better but *worse* after robot interaction (Fin_min_Base = -1 to -35). Ten of the worse performers came from the categories Below Average and Challenged, the remaining one coming from Advanced.

For H1 on Robot Design, we ran a GLM univariate (ANOVA) of Robot Design (2) × School (2) × Gender (2) on Fin_min_Base with Age as a covariate (N = 75). The only significant effect was the interaction of Robot Design × School (2) ($F_{(2,62)}$ = 3.33, p = .042). Yet, a two-tailed independent samples t-test indicated that the main effect of School on Fin_min_Base was not significant ($t_{(73)}$ = -.17, p = .86). The factor Robot Design had three levels: Humanoid (n = 21, M = 9.47, SD = 1.72), Puppy (n = 27, M = 9.50, SD = 1.83), and Droid (n = 27, M = 6.81, SD = 1.96). Therefore, we ran three two-tailed independent t-tests on Fin_min_Base but no significant effects occurred (Humanoid-Puppy: $t_{(46)}$ = -.52, p = .96; Humanoid-Droid: $t_{(46)}$ = .84, p = .40; Puppy-Droid: $t_{(52)}$ = 1.01, p = .32). Neither Robot Design nor School had a significant effect on learning gains as measured by Fin_min_Base.

We conjectured that perhaps certain Robot Designs exercised negative effects on learning. Therefore, we reran the analyses on the group that performed *worse* after robot tutoring. However, Robot Design and School again did not exert significant effects on Fin_min_Base. In all, the effects of schools, gender, and robot designs improved nor worsened the children's learning as measured through the difference scores.

For the 64 children (about 85%) that did show learning gains after robot intervention, we ran a paired samples t-test on Baseline versus FinMSco to see *how much* those children gained. The difference between Baseline (n = 64, M = 37.98, SD = 1.91) and FinMSco (n = 64, M = 49.14, SD = 2.05) was highly significant ($t_{(63)}$ = -11.20, p = .000). On average, those who learned from the robot did over one-third better compared to Baseline. Although most children learned significantly from robot tutoring, the various robot designs did not significantly differentiate the learning effects, therefore countering H1.

Although Robot Design did not exact significant effects on learning, perhaps the experience of the design as Human-like, Animal-like, or Machine-like would, allowing yet another chance for H1 to come to expression; albeit in a more perceptual way. To check the effects of the childrens' perceptions of their robot on learning, we did regression analysis of Human-like, Animal-like, and Machine-like on Fin_min_Base. However, no significant relationship was established (Human-like: t = -.47, p = .640; Animal-like: t = -.52, p = .610; Machine-like: t = -.50, p = .620). Also with Gain percentage as dependent (Table 1: Per_Fin_min_Base) significant effects remained absent (Human-like: t = -.26, p = .800; Animal-like: t = -1.16, p = .250; Machine-like: t = -.71, p = .480).

Combined with the results from the section on Learning effects, students perceived the robot as we expected but their perception had no effect on learning; not in absolute numbers of correct answers and not as a percentage of improvement from the Baseline. Although overall learning gains were achieved, the design of the robot embodiment or what it represented to the children did not matter, rejecting H1.

For H2 on Advancement Level, we ran a One-way ANOVA of Advancement Level on the difference score Fin_min_Base but none of the effects were significant ($F_{(3,71)}$ = 1.58, p = .202). No matter how well or poor children performed initially, it did not affect their learning gain on average.

As stated under Measures, we devised another measure from the notion that children may not have gained differently in absolute numbers but that 8.57 more multiplications correct is a relatively stronger gain for a poor performer than for an excellent student. Learning gain, then, was calculated from the percentage of gain (Fin_min_Base) in relation to the Baseline (Per_Fin_min_Base = Fin_min_Base / Baseline). With this measure, we ran a One-way ANOVA of Advancement Level on Per_Fin_min_Base for N = 64, excluding those with a learning loss. This time, we *did* find significant effects ($F_{(3,60)}$ = 12.66, p = .000).[5] On average, the gain percentage (Per_Fin_min_Base) increased with the decrease of Advancement Level (r = -.53, p = .000) (Advanced: n = 10, M = .17 (17%), SD = .11; Above Average: n = 19, M = .22 (22%), SD = .14; Below Average: n = 25, M = .35 (35%), SD = .28; Challenged: n = 10, M = .90 (90%), SD = .61).

To scrutinize the individual contrasts, we did 6 two-tailed independent t-tests of Advancement Level with Bonferroni correction (Challenged – Below Average, Challenged – Above Average, Challenged – Advanced, Below Average – Above Average, Below Average – Advanced, Above Average – Advanced) on Per_Fin_min_Base. The percentage of learning gain (Per_Fin_min_Base) of pupils that were Challenged (n = 10, M = .90, SD = .61) was significantly higher than those Below Average (n = 25, M = .35, SD = .28), Above Average (n = 19, M = .22, SD = .14), or Advanced (n = 10, M = .17, SD = .11) (Challenged – Below Average: $t_{(33)}$ = 3.68, p = .001; Challenged – Above Average: $t_{(27)}$ = 4.69, p = .000; Challenged – Advanced: $t_{(18)}$ = 3.73, p = .002). Yet, the differences among Below Average, Above Average, and Advanced were not significant (Supplementary Materials 1). The effects were caused by the Challenged pupils (n = 10), indicating that if weak students benefited, they benefited relatively more (90% improvement on Baseline) from robot tutoring than others. Calculated as the relative improvement to their individual baselines, H2 was confirmed for Challenged students but not for other.

3.3. Summary of findings for learning

1. Prior to robot intervention, pupils performed better with age and girls did better on baseline performance than boys. After 5 minutes of robot interaction, these differences disappeared
2. Most children (~85%) learned from the robot, a small group (~15%) performed worse
3. Those who learned from the robot had an average of more than one-third gain after tutoring
4. The weakest students that gained from robot tutoring did so in percentage of gain (90%), not in absolute numbers, compared to their earlier achievements
5. School, gender, design of the robot, the number of times these children were tutored, nor the experience of novelty of the robot were influential for learning through robot tutoring

3.4. Experience

Although we had a range of psychometric scales on our questionnaire to measure dimensions of affect (i.e. Engagement, Bonding, Anthropomorphism, Perceived Realism, Relevance, Perceived

[5] Even with worse performers included, the effect was significant (Supplementary Materials 1).

Affordances, and Use Intentions), none but Bonding achieved convergent *and* divergent measurement reliability (Supplementary Materials 1). Therefore, we decided to work with the only clear-cut case we had, Bonding, and not make *ad-hoc* decisions.

H3 expected that emotional bonding with the robot would positively affect the learning outcomes in a mediating or moderating way. To examine H3, we ran the previous GLM Repeated Measures again of Robot Design (3) × Advancement Level (4) (between-subjects) on the (within-subjects) number of equations correctly solved before and after robot tutoring but now with mean Bonding as the covariate. However, mean Bonding exerted no significant main or interaction effects on the multiplication scores and the earlier pattern of results was not altered (Supplementary Materials 1).

To let the presumed relation between bonding and learning happen more easily, we ran a two-tailed bivariate correlation analysis between M_{Bond} and Fin_min_Base (r = .007, p = .951) and between M_{Bond} and Per_Fin_min_Base (r = -.076, p = .531). Yet, neither were significant.

Therefore, H3 was rejected. Bonding tendencies were independent from the design of the robot or the advancement level of the children. The level of bonding with a robot tutor seemed not to have any substantial correlation with learning, not in absolute numbers nor in relative gain.

To check if any of the non-theoretical variables would affect the level of learning and bonding, we conducted GLM Multivariate Analysis (MANOVA) of Robot Design (3) × Advancement Level (4) × School (2) × Gender (2) on Fin_min_Base and M_{Bond} and on Per_Fin_min_Base and M_{Bond} with Age, Novelty, and Aesthetics as covariates. The following results were obtained:

1. The interaction of Robot Design × School × Gender on Fin_min_Base ($F_{(1,30)}$ = 6.44, p = .017) was significant. However, earlier we showed that none of the contrasts in the factors Robot Design, School, and Gender were significant so that 1. can be considered a false positive
2. The interaction of Robot Design × School × Gender on Per_Fin_min_Base ($F_{(1,30)}$ = 9.56, p = .004) was significant. To scrutinize the contrasts of the factor Robot Design, we ran three independent samples t-tests on Per_Fin_min_Base. Yet, none of the differences were significant (Humanoid – Puppy: $t_{(43)}$ = .14, p = .89; Humanoid – Droid: $t_{(44)}$ = 1.03, p = .31; Puppy – Droid: $t_{(51)}$ = 1.18, p = .24). Additionally, neither the difference between School ($t_{(70)}$ = -1.23, p = .22) nor that between Gender ($t_{(70)}$ = .13, p = .90) was significant. We therefore conclude that the significant F-value for 2. came from the accumulation of noise in the contrasts
3. The interaction of Robot Design × Advancement Level on Per_Fin_min_Base ($F_{(6,30)}$ = 4.15, p = .004) was the product of 4. and 5.
4. The main effect of Robot Design on Per_Fin_min_Base ($F_{(2,30)}$ = 6.06, p = .006) was significant but as said in 2., the contrasts of the factor Robot Design were not so that the inconsistency between ANOVA and t-test indicates the propagation of noise from a set of non-significant contrasts, resulting in a false-positive for the F-value
5. The main effect of Advancement Level on Per_Fin_min_Base ($F_{(3,30)}$ = 4.12, p = .015). As shown earlier, we saw that Per_Fin_min_Base decreased with the increase of Advancement, which was due to the group we regarded as Challenged
6. The only significant effect that included Bonding was that Aesthetics covaried with M_{Bond} ($F_{(1,71)}$ = 13.21, p = .001): A robot experienced as 'prettier' raised stronger bonding tendencies

3.4.1. Effects on Bonding

We ran a Univariate Analysis of Variance (ANOVA) of Robot Design and Advancement Level directly on mean Bonding. Not all children who took the multiplication test also filled out the questionnaire, therefore N = 70. The intercept was significantly different from zero so that Bonding tendencies did occur ($F_{(1,58)}$ = 194.76, p = .000, η_p^2 = .77). However, none of the main effects or interaction was significant (F < 1) (Supplementary Materials 1). Robot Design nor Advancement Level exerted significant effects on Bonding.

As an extra exploration, we conducted an ANOVA of Robot Design (3) × Advancement Level (4) × School (2) × Gender (2) on the grand averages of M_{Bond}, showing that only the difference in

School was significant ($F_{(1,34)} = 4.57$, $p = .04$). We ran an independent samples t-test of School on M_{Bond}, showing that Bonding at Good Shepherd was significantly higher than at Chun Lei ($t_{(68)} = 2.99$, $p = .004$). Theoretically, this is an irrelevant finding.

We then ran three t-tests with Sessions as the grouping variable (once – twice, once – thrice, twice – thrice). The effects on M_{Bond} of Once and Thrice and that of Twice and Thrice were not significant (Once – Thrice: $t_{(54)} = 1.31$, $p = .20$; Twice – Thrice: $t_{(20)} = .97$, $p = .34$). However, the difference between Once and Twice was significant for M_{Bond} (Once – Twice: $t_{(60)} = 3.01$, $p = .004$), even if α was corrected to .017 with respect to Bonferroni. Apparently, mean Bonding became less upon second encounter ($M_{Bond1} = 3.60$, $SD = 1.64$; $M_{Bond2} = 2.19$; $SD = 1.70$), which was due to Chun Lei pupils alone. The insignificant difference with those encountering the robot thrice might indicate a ceiling effect.

We wondered if the high bonding upon first encounter was due to a novelty effect, wearing off after multiple encounters. Therefore, we correlated M_{Bond} with Novelty and found that the correlation was significant but not very strong ($r = .31$, $p = .01$). Children from Chun Lei saw the robot more often so that less novelty may have led to lower rates of bonding. M_{Bond} also correlated with Aesthetics ($r = .56$, $p = .000$), indicating that the experience of 'prettier' led to stronger bonding tendencies as supported by the covariance analysis earlier on.

3.5. Summary of findings for experience

With respect to the experience of the robot tutor as a social entity, we found that:

1. The pupils perceived the robot as intended (manipulation successful)
2. The social role they attributed to the robots had no significant effect on their perceptions of human, animal, or machine-likeness, except that the role of 'machine' indeed raised significant machine-likeness, which a trivial finding
3. From a design perspective, the Bioloids to these children were basically all machines like Droid, while Puppy added animal-like features to that basic frame and Humanoid added human-like features to it. However, type of robot (humanoid, animal, or machine) did not affect the bonding tendencies
4. Only the Bonding scale was psychometrically reliable and all other measures for these children seemed to be related to that experience or were confusing
5. Bonding had no significant relation with learning gains. In 5 minutes of robot training, children improved their skills irrespective of the quality of the established relationship
6. The Good Shepherd children experienced more bonding with their robot tutor than Chun Lei pupils, maybe owing to a novelty effect
7. Stronger perceptions of the robot's attractiveness ('beautiful') were associated with stronger bonding tendencies

4. Discussion and Conclusions

We found that 5 minutes of robot tutoring improved learning the multiplications irrespective of the design of the robot or the advancement level of the pupils. This result counters our hypothesis H1 that a more anthropomorphic design would enhance performance. It also counters H2 on different effects for advancement level when dealt with as the absolute number of equations solved correctly. H2 is confirmed when seen as the relative gain pupils get from robot tutoring as compared to their earlier achievements; then, the more challenged children ($n = 10$) gain relatively more than the others. H3 was disconfirmed that the child learns more while developing a stronger emotional bond with the robot tutor. While rehearsing multiplication equations in this study, learning and bonding seemed to be two different strands of processing, both happening, but not affecting each other significantly.

Thus, our conclusion is very straightforward: Apparently, children improved on the multiplication tables with 5 minutes of exercise with a robot; more sessions were unnecessary. Initial differences between gender, age, or school disadvantages were compensated for and the novelty of the method had no significant effect on learning. The type of robot or its social role (teacher, peer, friend) also did not matter (cf. Onyeulo & Gandhi, 2020): A more human-like machine did not improve performance, a teacher role was no better than a peer, and the level of emotional bonding of the child with the tutoring machine (because it is new and beautiful) made no difference for their learning outcomes.

This is good news for teaching practice (cf. UN, 2020) because cheap and simple robots of whatever kind may help the larger part of pupils gain more than 33% better scores with little time investment. The weakest pupils should be treated with caution because many may have a 90% progress but some challenged and under-average children may be set back by robot tutoring. For different reasons, challenged as well as certain advanced students can be easily distracted and may experience learning difficulties (e.g., Beckmann & Minnaert, 2018).

The theory of affective bonding (Konijn & Hoorn, 2017; 2020a) was not supported. For the children, all the different conceptualizations of affordances, relevance, realism, and anthropomorphism seemed to be diffuse except for the notion of bonding ("I felt connected to the robot") and such bonding may be present but was not influential for rational performance.

Robots are not human beings (cf. Onyeulo & Gandhi, 2020). It may be that a warm relationship with a human teacher makes a child want to work harder and may improve its social-emotional development (e.g., Frymier & Houser, 2000; Hattie, 2015; Skinner & Belmont, 1993; Hamre & Pianta, 2001). Yet, for a simple drill like quickly practicing multiplications with a little robot, warm relationships did not seem to be necessary in our case, perhaps because the interaction was so short. According to Serholt and Barendregt (2016), it may be that children do not develop bonds with robots in the human sense but engage in a different sort of relationship and what that is, needs to be found out.

Our work does coincide with the results of Hindriks and Liebens (2019) that social behavior during a maths task is not conducive to learning. Moreover, for certain challenged pupils, the effects we found were even counter-productive. It seems that matching the robot's appearance with its task is insignificant despite some individual preferences for specific robot appearances in some tasks (Li, Rau, & Li, 2010; Imai, Ono, & Ishiguro, 2003; Mutlu, Forlizzi, & Hodgins, 2006; Konijn, Smakman, & Van den Berghe, 2020). Our robots were successful at maintenance rehearsal and repeated exercise (e.g., Wei, Hung, Lee, & Chen, 2011; Huang & Hoorn, 2018) and during the remedial teaching of a strongly rational task, the bonding aspects of the robot appeared to be unimportant.

Strong point in our study is the comparability of the three robot designs. It is quite hard to compare existing factory robots of a different make, telling which design elements are responsible for the differences in user responses. Our basic design, materials, and general appearance of the robots was similar but differentiated in representation: It is a rather unique finding that the children recognized the basic design of all three robots as a machine with human features added for the humanoid and animal characteristics for the puppy. Unexpectedly, these representational variations were not conducive to learning, which brings us to the limitations of this study.

Field studies add to ecological validity and plausibility yet at the cost of methodological soundness. The time schedules of schools and parents left us with 75 children that could participate in but one session so the insignificant progress after the second and third session may have been due to a lack of power. Also effects of the advancement level (weaker-stronger pupils) may have been disturbed by the small numbers in a cell. Working with children in itself already yields nosier data than with adults, which may have drowned some effects of taking multiple sessions, the mix-up of psychometric constructs (e.g., anthropomorphism, realism), or the effects on bonding. It may be argued that 5 minutes of interaction is too short to become attached to a machine.

4.1. Future outlook

Due to severe budget cuts and fewer teachers, education faces a lack of human resources to serve an increasingly larger number of pupils with a wider variety of individual needs. Owing to changes in care systems (in Europe), children with special needs are integrated in regular rather than special schools (e.g., Mader, 2017; for the situation in Hong Kong, see Lee, Yeung, Tracey, & Barker, 2015). Migration causes new mixes of children with diverse backgrounds, cultural and educational differences. The current pandemic asks for novel teaching solutions to make up for learning loss (UN, 2020). These transitions demand ways of teaching that differ from class-wise instructions (ibid.). As is, the teaching level converges to the middle whereas children learn most if instruction matches their level of proficiency (Leyzberg, Spaulding, & Scassellati, 2014).

Social robots may provide support, which probably has far-reaching implications for classroom instruction and organization. For example, repetitive tasks may be performed by the robot while the teacher focuses on special cases or develops and teaches advanced topics. This actually asks from the teachers to recalibrate their profession. In the near future, teachers may have to consider working in teams that also consist of synthetic colleagues. However, before the role of this new robot colleague can be outlined, we have to understand how a robot's (limited) capabilities can match the teaching needs of pupils but also of teachers. In this respect, moral deliberations on robots in education should be proliferated (e.g., Smakman & Konijn, 2020).

Our results suggest that the robot does not have to be fancy in looks or behavior to help the child increase its performance quickly in arithmetic rehearsal tasks. In this study, weak pupils benefited strongly from robot instruction with the exception of a few challenged children. Robot teachers in motion pictures and comic books do not have to remain mere science fiction. Educators and parents may apply a simple and cheap machine equipped with the proper software to make up for knowledge deficits and gaps in the learning process without having to fear the lack of face-to-face interaction. That makes robot tutoring feasible in times of a COVID-19 pandemic.

Supplementary Materials: Technical Report S1: Bioloids, multiplication TechRep, Software S1: Bioloids, code, Audio S1: Bioloids, audio files.

Author Contributions: Conceptualization, E.A.K. and L.v.B; methodology, E.A.K., L.v.B. and J.F.H.; software, L.v.B.; validation, L.v.B., J.F.H. and I.S.H.; formal analysis, I.S.H. and J.F.H.; investigation, L.v.B. and I.S.H.; resources, J.F.H. and I.S.H.; data curation, J.F.H.; writing—original draft preparation, J.F.H., I.S.H., L.v.B.; writing—review and editing, J.F.H. and E.A.K.; visualization, J.F.H. and L.v.B.; supervision, J.F.H. and E.A.K.; project administration, J.F.H.; funding acquisition, N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgments: We would like to thank the pupils, parents, and staff of the Free Methodist Bradbury Chun Lei Primary school (i.e. principal Liu) and of the S.K.H. Good Shepherd Primary school (i.e. principal Chan) in the Hong Kong SAR for their kind cooperation in the experiments. Tom Ootes and Prashant Sridhar are kindly acknowledged for their help in programming the software. Ruowei Li (Tristan) is kindly acknowledged for assembling two of the Bioloid robots. Without the hands-on support of Stefanie de Boer, Marijn Hagenaar, Eddie Hong, Kelvin Lai, Gigi Lee, Edison Liang, Dominique Roos, and Dickson Soon, we would not have been able to conduct the experiments.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Appendix A

Structured questionnaire on the experience of a tutoring robot (English translated from the Cantonese). Variable names (between brackets) were left out from the original questionnaire.

What did the robot look like to you? The more circles you fill in, the more you agree with the statement. Only one circle filled in means you don't agree at all, all circles filled in means you totally agree.

機器人對你來說像什麼呢？你填滿越多的圈圈代表你越認同對應的陳述。只填滿一個圈圈代表你完全不同意，如果所有圓圈都被你填滿了，代表你十分認同這個陳述。

[Representation]

The robot looked like a…

機器人看起來像…

1. Machine
   機器

○ ○ ○ ○ ○ ○

2. Human
   人類

○ ○ ○ ○ ○ ○

3. Animal
   動物

○ ○ ○ ○ ○ ○

[Social Role]

What did the robot feel like to you? To me the robot felt like a…

(choose one answer that suits you best)

你怎麼看待機器人呢？對我來說，機器人像一個…

(選擇一個最接近你想法的)

4. Friend
   朋友

○ ○ ○ ○ ○ ○

5. Classmate
   同學

○ ○ ○ ○ ○ ○

6. Teacher
   老師

○ ○ ○ ○ ○ ○

7. Acquaintance
   熟人

○ ○ ○ ○ ○ ○

8. Stranger
   陌生人

○ ○ ○ ○ ○ ○

9. Machine

機器

○ ○ ○ ○ ○ ○

10. Other…
其它

○ ○ ○ ○ ○ ○

How did you feel about your connection with the robot? The more circles you fill in the more you agree with the statement.
你覺得你跟機器人的關係怎麼樣呢？越多的圈圈代表你越認同對應的陳述。

[Engagement]

The robot…
這個機器人

11. I like the robot
我喜欢機器人

○ ○ ○ ○ ○ ○

12. The robot gave me a good feeling
它讓我感覺很好

○ ○ ○ ○ ○ ○

13. I felt uncomfortable with the robot
機器人令我感觉不舒服

○ ○ ○ ○ ○ ○

14. It was fun with the robot
機器人好好玩

○ ○ ○ ○ ○ ○

15. I dislike the robot
我不喜欢機器人

○ ○ ○ ○ ○ ○

[Bonding]

16. I felt a bond with the robot
我觉得和機器人有联结

○ ○ ○ ○ ○ ○

17. I felt like the robot was interested in me
我觉得機器人对我有兴趣

○ ○ ○ ○ ○ ○

18. I felt connected to the robot
    我对機器人有联结的感觉

○ ○ ○ ○ ○ ○

19. I want to be friends with the robot
    我想和機器人做朋友

○ ○ ○ ○ ○ ○

20. The robot understands me
    機器人明白我

○ ○ ○ ○ ○ ○

What did you think about your interaction with the robot? The more circles you fill in the more you agree with the statement.
你覺得你跟機器人的互動怎麼樣？越多的圓圈代表你越同意。

[Anthropomorphism]

21. To me the robot was a machine
    我覺得機器人只是一个物件

○ ○ ○ ○ ○ ○

22. It felt just like a human was talking to me
    我觉得好像一个人和我说话

○ ○ ○ ○ ○ ○

23. I reacted to the robot just as I react to a human
    我跟機器人对话犹如和人类对话一样

○ ○ ○ ○ ○ ○

24. It differed from a human-like interaction
    和機器人交流和人类不一样

○ ○ ○ ○ ○ ○

[Perceived Realism]

25. The robot resembled a real-life creature
    機器人犹如活物一样

○ ○ ○ ○ ○ ○

26. It was just like real to me

機器人好真实

○ ○ ○ ○ ○ ○

27. The robot was fabricated
    機器人好做作

○ ○ ○ ○ ○ ○

28. It felt just like a real conversation
    和機器人对话好真实

○ ○ ○ ○ ○ ○

[Relevance]

29. The robot was important to do my exercises
    機器人对我学习很重要

○ ○ ○ ○ ○ ○

30. The robot helped me to practice the multiplication tables
    機器人帮到我练习乘法表

○ ○ ○ ○ ○ ○

31. The robot was useless for rehearsing the multiplication tables
    機器人帮不到我练习乘法表

○ ○ ○ ○ ○ ○

32. The robot is what I need to practice the multiplication tables
    我需要機器人才能练习乘法表

○ ○ ○ ○ ○ ○

[Perceived Affordances]

33. I understood the task with the robot immediately
    我明白機器人的指示

○ ○ ○ ○ ○ ○

34. The robot was clear in its instructions
    機器人的指示好清晰

○ ○ ○ ○ ○ ○

35. It took me a while before I understood what to do with the robot
    我需要一点时间明白機器人的操作

○ ○ ○ ○ ○ ○

36. I puzzled to understand how to work with the robot

我对于機器人的用法有点疑问

○ ○ ○ ○ ○ ○

[Use Intentions]

For the next time practicing multiplications, I would….

下次练习乘法表的时候，我会。。

37. use the robot again

再次用機器人

○ ○ ○ ○ ○ ○

38. use another tool, like a tablet

使用其他学习工具

○ ○ ○ ○ ○ ○

39. want this robot to help me again

想要機器人再次帮我

○ ○ ○ ○ ○ ○

Then, some final questions

The more circles you fill in the more you agree with the statement.

最后几个问题，圈得越多代表你越同意。

[Novelty]

40. I played with robots before

我有玩过機器人

○ ○ ○ ○ ○ ○

[Aesthetics]

The robot looked…

機器人的外表。。

41. Beautiful

很漂亮

○ ○ ○ ○ ○ ○

[Demographics]

42. I am a…

我是一個

○ Boy 男孩

- Girl 女孩

43. How old are you 請問你幾歲？

_____

Thank you for all the help. See you next time!!

謝謝你的幫助。期待我們下次再見。

References

Alves-Oliveira, P., Tullio, E. D., Ribeiro, T., Paiva, A. (2014). Meet me halfway: eye behaviour as an expression of robot's language. In: *AAAI Fall Symposium Series*, pp. 13-15.

Arroyo, I., Royer, J. M., & Park Woolf, B. (2011). Using an intelligent tutor and math fluency training to improve math performance. *International Journal of Artificial Intelligence in Education, 21*(1-2), 135-152. doi: 10.3233/JAI-2011-020

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.

Beckmann, E., & Minnaert, A. (2018). Non-cognitive characteristics of gifted students with learning disabilities: an in-depth systematic review. *Frontiers in Psychology, 9*, 504. doi: 10.3389/fpsyg.2018.00504

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science Robotics, 3*(21), eaat5954. doi: 10.1126/scirobotics.aat5954

Beran, T. N., & Ramirez-Serrano, A. (2011). Can children have a relationship with a robot? *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering Human-Robot Personal Relationships*, 49-56. doi: 10.1007/978-3-642-19385-9_7

Boucher, J. D., Pattacini, U., Lelong, A., Bailly, G., Elisei, F., Fagel, S., Dominey, P. F., Ventre-Dominey, J. (2012). I reach faster when I see you look: gaze effects in human-human and human-robot face-to-face cooperation. *Frontiers in Neurorobotics 6*(3), 1-11. doi: 10.3389/fnbot.2012.00003

Brown, L., Kerwin, R., & Howard, A. M. (2013, October). Applying behavioral strategies for student engagement using a robotic educational agent. In *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 4360-4365). doi: 10.1109/SMC.2013.744

Chang, C. W., Lee, J. H., Chao, P. Y., Wang, C. Y., & Chen, G. D. (2010). Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school. *Journal of Educational Technology & Society*, *13*(2), 13-24.

Chen, H., Park Hae, W., & Breazeal C. L. (2020). Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement. *Computers & Education, 150,* 103836. doi: 10.1016/j.compedu.2020.103836.

Esposito, J. (2011). Negotiating the gaze and learning the hidden curriculum: a critical race analysis of the embodiment of female students of color at a predominantly white institution. *Journal for Critical Education Policy Studies*, *9* (2), 143-164.

Frymier, A. B., & Houser, M. L. (2000). The teacher-student relationship as an interpersonal relationship. *Communication Education*, *49*(3), 207-219. doi:10.1080/03634520009379209

Gomoll, A., Šabanović, S., Tolar, E., Hmelo-Silver, C. E., Francisco, M., & Lawlor, O. (2017). Between the social and the technical: Negotiation of human-centered robotics design in a middle school classroom. *International Journal of Social Robotics*, *10*(3), 309-324. doi: 10.1007/s12369-017-0454-3

Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development*, *72*(2), 625-638. doi:10.1111/1467-8624.00301

Hattie, J. (2015). The applicability of Visible Learning to higher education. *Scholarship of Teaching and Learning in Psychology*, *1*(1), 79-91. doi:10.1037/stl0000021

Hindriks, K. V., & Liebens, S. (2019). A robot math tutor that gives feedback. In M. Salichs et al. (Eds.), *Social Robotics. ICSR 2019. Lecture Notes in Computer Science, vol. 11876* (pp. 601-610). Cham, CH: Springer. doi: 10.1007/978-3-030-35888-4_56

Huang, I. S., & Hoorn, J. F. (2018). Having an Einstein in class. Teaching maths with robots is different for boys and girls. In X. Wang, Z. Wang, J. Wu, & L. Wang (Eds.), *Proceedings of the 13th World Congress on Intelligent Control and Automation (WCICA 2018) July 4-8, 2018. Changsha, China* (pp. 424-427). New York: IEEE. doi: 10.1109/WCICA.2018.8630584 Available from https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8630584

Imai, M., Ono, T., & Ishiguro, H. (2003). Physical relation and expression: joint attention for human-robot interaction. *Proceedings of the 10th IEEE International Workshop on Robot and Human Interactive Communication. ROMAN 2001 (Cat. No.01TH8591)*, *50*(4), 636-643. doi: 10.1109/roman.2001.981955

Jamone, L., Ugur , E., Cangelosi, A., Fadiga, L., Bernardino, A., Piater, J., & Santos-Victor, J. (2018). Affordances in psychology, neuroscience, and robotics: a survey. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(1), 4-25. doi: 10.1109/TCDS.2016.2594134

Kennedy, J., Baxter, P., & Belpaeme, T. (2015). The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 67-74). ACM. doi: 10.1145/2696454.2696457

Kennedy, J., Baxter, P., Senft, E., & Belpaeme, T. (2016). Social robot tutoring for child second language learning. *Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 231-238. doi: 10.1109/hri.2016.7451757

Konijn, E. A., & Hoorn, J. F. (2017). Parasocial interaction and beyond: media personae and affective bonding. In P. Roessler, C. Hoffner, & L. van Zoonen (Eds.), *The international encyclopedia of media effects* (pp. 1-15). NY: Wiley-Blackwell. doi: 10.1002/9781118783764.wbieme0071

Konijn, E. A., & Hoorn, J. F. (2020a). Media psychological perspectives on the use of communication robots in health care. In J. van den Bulck (Ed.), *The international encyclopedia of media psychology* (pp. xx-xx). New York: Wiley. doi: 10.1002/9781119011071.iemp0317

Konijn, E. A., & Hoorn, J. F. (2020b). Robot tutor and pupils' educational ability: Teaching the times tables. *Computers & Education, 157*, 103970. doi: 10.1016/j.compedu.2020.103970

Konijn, E. A., Smakman, M., & Van den Berghe, R. (2020). Use of robots in education. In J. Van den Bulck, E. Sharrer, D. Ewoldsen, & M.-L. Mares (Eds.), *The international encyclopedia of media psychology*. New York: Wiley.

Köse, H., Uluer, P., Akalın, N., Yorgancı, R., Özkul, A., & Ince, G. (2015). The effect of embodiment in sign language tutoring with assistive humanoid robots. *International Journal of Social Robotics*, *7*(4), 537-548. doi:10.1007/s12369-015-0311-1

Kory-Westlund, J. M., & Breazeal, C. L. (2014). Storytelling with robots: Learning companions for preschool children's language development. In *23rd IEEE International Symposium on Robot and Human Interactive Communication, 2014 (RO-MAN), Edinburgh* (pp. 643- 648). IEEE. doi: 10.1109/roman.2014.6926325

Kory-Westlund, J. M., Jeong, S., Park Hae, W., Ronfard, S., Adhikari, A., Harris, P. L., DeSteno, D., & Breazeal, C. L. (2017). Flat versus expressive storytelling: Learning and retention of a robot's narrative. *Frontiers in Human Neuroscience, 11*, 643-648. doi: 10.3389/fnhum.2017.00295.

Kory-Westlund, J. M., & Breazeal, C. L. (2019). A long-term study of young children's rapport, social emulation, and language learning with a peer-like robot playmate in preschool. *Frontiers in Robotics and AI, 6,* 81. doi: 10.3389/frobt.2019.00081

Krämer, N. C., Karacora, B., Lucas, G., Dehghani, M., Rüther, G., & Gratch, J. (2016). Closing the gender gap in STEM with friendly male instructors? On the effects of rapport behavior and gender of a virtual agent in an instructional interaction. *Computers & Education, 99,* 1-13. doi: 10.1016/j.compedu.2016.04.002

Lee, F., Yeung, A., Tracey, D., & Barker, K. (2015). Inclusion of children with special needs in early childhood education. *Topics in Early Childhood Special Education, 35*(2), 10.1177/0271121414566014.

Leyzberg, D., Spaulding, S., Toneva, M., & Scassellati, B. (2012). The physical presence of a robot tutor increases cognitive learning gains. *Proceedings of the Annual Meeting of the Cognitive Science Society, 34,* 1882-1887. Retrieved from https://escholarship.org/uc/item/7ck0p200

Li, D., Rau, P. L. P., & Li, Y. (2010). A cross-cultural study: effect of robot appearance and task. *International Journal of Social Robotics*, *2*(2), 175-186. doi: 10.1007/s12369-010-0056-9

Mader, J. (2017, March 3). How teacher training hinders special-needs students. *The Atlantic Daily.* Retrieved from www.theatlantic.com/education/archive/2017/03/how-teacher-training-hinders-special-needs-students/518286/

Mann, J. A., MacDonald, B. A., Kuo, I., Li, X., & Broadbent, E. (2015). People respond better to robots than computer tablets delivering healthcare instructions. *Computers in Human Behavior, 43*, 112-117. doi: 10.1016/j.chb.2014.10.029

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 (HRI '14) March 3-6, 2014. Bielefeld, Germany* (pp. 382-389). New York: ACM. doi: 10.1145/2559636.2559678

Mutlu, B., Forlizzi, J., & Hodgins, J. (2006). A storytelling robot: modeling and evaluation of human-like gaze behavior. *2006 6th IEEE-RAS International Conference on Humanoid* Robots (pp. 518-523). doi: 10.1109/ichr.2006.321322

Nuse, I. P. (2017, September 30). Humanoid robot takes over as teacher. Retrieved from http://sciencenordic.com/humanoid-robot-takes-over-teacher

Onyeulo, E. B., & Gandhi, V. (2020). What makes a social robot good at interacting with humans? *Information*, *11*(1), 43. doi: 10.3390/info11010043

Paauwe, R. A., Hoorn, J. F., Konijn, E. A., & Keyson, D. V. (2015). Designing robot embodiments for social interaction: Affordances topple realism and aesthetics. *International Journal of Social Robotics*, *7*(5), 697-708. doi: 10.1007/s12369-015-0301-3

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 28th International Conference on Human Factors in Computing Systems (CHI '10)* (pp. 1613-1622). Atlanta: ACM. doi: 10.1145/1753326.1753567

Serholt, S., & Barendregt, W. (2016). Robots tutoring children: Longitudinal evaluation of social engagement in child-robot interaction. *Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI '16)* (pp. 1-10). doi: 10.1145/2971485.2971536

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*(4), 571-581. doi: 10.1037/0022-0663.85.4.571

Smakman, M., & Konijn, E. A. (2020). Robot tutors: welcome or ethically questionable? In M. Merdan, W. Lepuschitz, G. Koppensteiner, R. Balogh, & D. Obdržálek (Eds.), *Robotics in Education, Advances in Intelligent Systems and Computing (AISC), 1023,* (pp. 376-386). Cham, CH: Springer. doi: 10.1007/978-3-030-26945-6_34.

STEMex. (2019). STEMex Learning Centre. Retrieved from https://stemex.org/about/

Syrdal, D. S., Dautenhahn, K., Woods, S. N., Walters, M. L., & Koay, K. L. (2007). Looking good? Appearance preferences and robot personality inferences at zero acquaintance. *AAAI Spring Symposium: Multidisciplinary Collaboration for Socially Assistive Robotics* (pp. 26-28).

Tiberius, R. G., & Billson, J. M. (1991). The social context of teaching and learning. *New Directions for Teaching and Learning*, *1991*(45), 67-86. doi: 10.1002/tl.37219914509

UN (2020, August). *Policy brief: education during COVID-19 and beyond.* United Nations. Available from https://www.un.org/development/desa/dspd/wp-content/uploads/sites/22/2020/08/sg_policy_brief_covid-19_and_education_august_2020.pdf

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46* (4), 197-221. doi: 10.1080/00461520.2011.611369

Van Vugt, H. C., Hoorn, J. F., Konijn, E. A., & De Bie Dimitriadou, A. (2006). Affective affordances: Improving interface character engagement through interaction. *International Journal of Human-Computer Studies*, *64*(9), 874-888. doi:10.1016/j.ijhcs.2006.04.008

Van Vugt, H. C., Konijn, E. A., Hoorn, J. F., Eliëns, A., & Keur, I. (2007). Realism is not all! User engagement with task-related interface characters. *Interacting with Computers*, *19*(2), 267-280.

Vogt, P., Haas, M. D., Jong, C. D., Baxter, P., & Krahmer, E. (2017). Child-Robot Interactions for Second Language Tutoring to Preschool Children. *Frontiers in Human Neuroscience*, *11*, 73. doi: 10.3389/fnhum.2017.00073

Wei, C. W., Hung, I. C., Lee, L., & Chen, N. S. (2011). A joyful classroom learning system with robot learning companion for children to learn mathematics multiplication. *The Turkish Online Journal of Educational Technology*, *10*(2), 11-23.