Bioloids Teach Multiplications
(Technical Report)

HUANG Ivy S.
HOORN Johan F.

Contents

* Significant effects

Objectives

We conducted an experiment with different designs of social robots, rehearsing the multiplication tables with primary school children in Hong Kong. We analyzed the effects of School (2) / Partake (3) × Robot Design (3) × Gender (2) × Advancement (4) on
(a) learning gain, and
(b) experience of a robot tutor

Variables

Before analysis, we list all the variables and explain the items on the questionnaire. Where in this report Partake is mentioned, in the paper the name Sessions is used.

1. Summary of variables

Variable	Values	Label	Measure
School	1 = "GoodShepherd" 2 = "ChunLei"	S.K.H. Good Shepherd Primary School, Hong Kong SAR	Nominal
		Free Methodist Bradbury Chun Lei Primary School, Hong Kong SAR	

Robot	1 = “Humanoid” 2 = “Puppy” 3 = “Droid”	The appearance of the robot tutor	Nominal
Baseline (from pre-test)	[0,147]	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	Scale
Advancement	1 = “Challenged” 2 = “Below average” 3 = “Above average” 4 = “Advanced”	Advancement level of pupils, according to baseline.	Ordinal
BasePerf	1 = “High” 2 = “Medium” 3 = “Low”	High and Low are the five maximum outliers and the five minimum outliers detected in a data exploration process. Medium are those who are not outliers.	Nominal
MuSc1	[0, max]	Multiplication Score at t1 during interaction	Scale
MuSc2	[0, max]	Multiplication Score at t2 during interaction	Scale
MuSc3	[0, max]	Multiplication Score at t3 during interaction	Scale
Partake / Sessions	[1,2,3]	Number of times participant interacted with the robot	Scale
FinMSco (from post-test)	[0,147]	Final multiplication score based on multiplying 1-or-2 digit with 2-digit numbers	Scale
Fin_min_Base	[0, 147]	Difference between pre-test Baseline and post-test Final score. Also calculated as difference-score of FinMSco minus Baseline	Scale
Per_Fin_min_Base	[min, max]	The percentage of Fin_min_Base compared with Baseline	
Representation	Human_like = [1,6] Animal_like = [1,6] Machine_like = [1,6]	What does the robot look like to the participant?	Nominal
Social role	Friend = [1,6] Classmate = [1,6] Teacher = [1,6] Acquaintance = [1,6] Stranger = [1,6] Machine = [1,6] Other = [1,6]	What does the robot feel like to the participant?	Nominal
Bonding	Bon_1…5 = [1,6]	How is the social-affective relationship between participant and robot tutor? (Konijn & Hoorn, 2017)	Scale
Mbond	[1,6]	The mean value of bonding items: Bon_1…5	Scale
Anthropomorphism	Anth_1…4 = [1,6]	Does participant attribute human traits or emotions to robot tutor? (Konijn & Hoorn, 2017)	Scale
Perceived realism	Real_1…4 = [1,6]	Does robot tutor feel like a real creature or is it a fake? (Paauwe et al., 2015)	Scale
Perceived relevance	Rel_1…4 = [1,6]	Is robot tutor significant for doing the multiplication exercise? (Konijn & Hoorn, 2017)	Scale
Perceived affordances	Aff_1…4 = [1,6]	What can I do with the robot (in view of the multiplication exercise)? (Konijn & Hoorn, 2017)	Scale
Engagement	Eng_1…5 = [1,6]	Level of involvement with the robot	Scale
Use intentions	Use_Int_1...3= [1,6]	Want to use the robot again?	Scale
Novelty	[1,6]	To what extent is the robot tutor new to the participant?	Scale
Aesthetics	Aest_1 = [1,6]	To what extent is the robot attractive to the participant in terms of appearance?	Scale
Gender	0 = “Male” 1 = “Female”	The gender of the participant	Nominal
Age	[7, 10]	The age of the participant	Scale

Table 1. Variable details

2. Participants

A total of 95 pupils from two Hong Kong primary schools signed up for the experiment. Eventually 75 students were able to participate in at least one session with the robot and do the pre and post-test (N = 75; M_{Age} = 8.4, SD_{Age} = .82, range: 7-10, 44% female, Hongkongers). These 75 participants were randomly distributed over three differently Robot Designs (between-subjects): Humanoid (n = 21), Puppy (n = 27), and Droid (n = 27). A Chi-square test of independence checked for the distribution of Age over robots types but no significant relationship was found ($¡^2_{(6)}$ = 1.76, p = .94).

We planned for all pupils to participate in 3 robot tutoring sessions spread over more weeks (within-subjects). Due to the schools' tight time schedules, however, not every pupil could partake 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 once. Those who participated twice (13), and thrice (14) were all from Chun Lei. Also see next section.

Boys and girls were distributed over the Robot Designs as follows: Humanoid (15 males, 6 females), Puppy (15 males, 12 females), and Droid (12 males, 15 females). The schools' strict time scheduling caused inconsistencies in the ratios but these unequal distributions did not render a significant interaction effect ($¡^2_{(2)}$ = 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. 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). Also see next section. No significant effect of unequal distributions was found between Advancement Level and Robot Design ($¡^2_{(6)}$ = 1.73, p = .943).

3. Participant distribution

School & Robot / Advancement & Gender		GoodShepherd + ChunLei			ChunLei						Total	
		Humanoid	Puppy	Droid	Humanoid		Puppy		Droid			
					2 times	3 times	2 times	3 times	2 times	3 times		
Challenged	Female	0	1	0	1	0	0	1	0	1	4	11
					1		1		1			
	Male	1	2	1	0	1	0	0	1	1	7	
					1		0		2			
Below Average	Female	1	2	6	1	0	1	1	2	0	14	34
					1		2		2			
	Male	5	4	3	1	0	2	2	2	1	20	
					1		4		3			
Above Average	Female	1	2	2	1	0	0	1	0	1	8	19
					1		1		1			
	Male	4	2	1	0	2	0	2	0	0	11	
					2		2		0			
Advanced	Female	1	4	2	0	0	0	0	0	0	7	11
					0		0		0			
	Male	1	0	2	0	0	0	0	1	0	4	
					0		0		1			
Total		14	17	17	4	3	3	7	6	4	75	
					7		10		10			
		48			27							

Table 2. Participant distribution

4. Questionnaire overview

Index	Section	Description	Number of items	Abbreviation & Value
Design factors				
1	Representation	What does the robot look like to the participant?	3	Human_like = [1,6] Animal_like = [1,6] Machine_like = [1,6]
2	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]
Experiment factors				
3	Engagement	Level of involvement with the robot	5	Eng_1…5 = [1,6]
4	Bonding	How is the social-affective relationship between participant and robot tutor? (Konijn & Hoorn, 2017)	5	Bon_1…5 = [1,6]
5	Anthropomorphism	Do the participants attribute human traits or emotion to the robot tutor? (Konijn & Hoorn, 2017)	4	Anth_1…4 = [1,6]
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]
7	Perceived relevance	Does the robot tutor have importance for doing the multiplication exercise? (Konijn & Hoorn, 2017)	4	Rel_1…4 = [1,6]
8	Perceived affordances	What can I do with the robot (in view of multiplication exercise)? (Konijn & Hoorn, 2017)	4	Aff_1…4 = [1,6]
9	Use intentions	Want to use the robot again?	3	Use_Int_1..3= [1,6]
Control factors				
10	Novelty	To what extent is the robot tutor new to the participant?	1	Nov_1 = [1,6]
11	Aesthetics	To what extent is the robot attractive to the participant in terms of appearance?	1	Aesr_1 = [1,6]
12	Gender	/	1	Gender = [Male, Female]
15	Age	/	1	Age = [7, 10]

Table 3. Questionnaire details

5. Data Analysis in Brief

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, η_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.

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.

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 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.

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).[1] 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. The effects were caused by the Challenged pupils (n = 10), indicating that if weak students benefited, they benefited relatively

[1] Even with worse performers included, the effect was significant.

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.

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

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 Affordances, and Use Intentions), none but Bonding achieved convergent *and* divergent measurement reliability. 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.

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.

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 ($F_{(1,58)} = 194.76$, $p = .000$, $\eta_p^2 = .77$). However, none of the main effects or interaction was significant ($F < 1$). 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.

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

6. Data Analysis Extended

We did our data analysis in SPSS (version 23...0) and started with the effects on learning gains and then the effects on the experience during robot interaction. After exploring the learning gains, we inspected the theoretically less interesting variables, such as Age, School, and Gender in the hope that they did not sort significant effects. For the experiential variables, we started with reliability analysis of the scales and tested the effects of various design factors on experience. Lastly, we looked into the interaction between experience and learning gains.

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.

We found 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 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 (see next) except for Advancement Level, which was a trivial finding obviously. H1 and H2 were refuted for learning gain counted in absolute numbers. Next we run a number of checks and controls for possible confounds.

7. Effect of Age, School, Gender, and Novelty on Baseline and FinMSco

There are several variables of little theoretical interest (e.g., Age, School, Gender), so we wanted to check if they have a significant effect. If not, we would dismiss them from the main analyses.

7.1. Outlier Analysis on Baseline - find the extreme values

To find out the outliers who did very good or bad on their Baseline, we conducted an outlier analysis on Baseline to find the extremes (Table 4).

Extreme Values

			Case Number	Value
Baseline	Highest	1	73	78.00
		2	71	69.00
		3	72	69.00
		4	70	65.00
		5	75	64.00
	Lowest	1	25	10.00
		2	1	10.00
		3	3	17.00
		4	54	20.00
		5	22	20.00

Table 4: Extreme values of Baseline

Table 4 shows the extreme values found for Baseline. Five people had extremely high performance and five people extremely poor performance.

7.2. Two-tailed Correlation Between Baseline-Age and FinMSco-Age*

To verify whether Age can be omitted from analysis, we ran a two-tailed bivariate correlation analysis between Baseline and Age (Table 5) and between FinMSco and Age (Table 6).

Correlations

		Baseline	Age
Baseline	Pearson Correlation	1	.358**
	Sig. (2-tailed)		.002
	N	75	75
Age	Pearson Correlation	.358**	1
	Sig. (2-tailed)	.002	
	N	75	75

**. Correlation is significant at the 0.01 level (2-tailed).

Table 5. Correlation between Baseline and Age

Correlations

		Final Multiplication Score	Age
Final Multiplication Score	Pearson Correlation	1	.239*
	Sig. (2-tailed)		.039
	N	75	75
Age	Pearson Correlation	.239*	1
	Sig. (2-tailed)	.039	
	N	75	75

*. Correlation is significant at the 0.05 level (2-tailed).

Table 6. Correlation between FinMSco and Age

Table 5 shows that the correlation between Baseline and Age is significant (r = .36, p = .002). Similarly, Table 6 shows that the correlation between FinMSco and Age is significant (r = .24, p = .039), implying that Age should not be omitted from analysis. The older participants performed better on both Baseline and FinMSco. On the other hand, the correlation coefficient for the Age-Baseline test is .36 (medium correlation) and after robot intervention that for the Age-FinMSco is .24 (low correlation), indicating that the robot tutoring perhaps diminished the effect of Age on learning.

7.3. Two-tailed Correlation Between Novelty and Fin_min_Base

To check whether the difference scores (Fin_min_Base) were affected by the newness of the robot experience, we ran a two-tailed bivariate correlation analysis between Novelty and Fin_min_Base (Table 7).

Correlations

		Nov_1	Fin_min_Base
Nov_1	Pearson Correlation	1	.187
	Sig. (2-tailed)		.121
	N	70	70
Fin_min_Base	Pearson Correlation	.187	1
	Sig. (2-tailed)	.121	
	N	70	75

Table 7. Correlation between Fin_min_Base and Novelty

Table 7 shows that the correlation between Novelty and Fin_min_Base (r = .187, p = .12) was not significant. That means novelty of the robot did not influence the learning.

7.4. Two-tailed Independent Samples T-test of School on Baseline

To check whether School could be omitted from the analyses, we ran a two-tailed independent samples t-test on Baseline with the two schools (ChunLei and GoodShepherd) (Table 8).

Group Statistics

	School	N	Mean	Std. Deviation	Std. Error Mean
Baseline	Good Shepherd	48	39.7083	15.84897	2.28760
	Chun Lei	27	32.6296	11.94265	2.29836

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Baseline	Equal variances assumed	4.355	.040	2.018	73	.047	7.07870	3.50697	.08933	14.06808
	Equal variances not assumed			2.183	66.777	.033	7.07870	3.24278	.60570	13.55171

Table 8. Independent Samples T-test on Baseline with groups School

The results in Table 8 show that the mean difference between the Baseline of GoodShepherd (n = 48, M = 39.71, SD = 15.85) and ChunLei (n = 27, M = 32.63, SD = 11.94) is significant ($t_{(73)}$ = 2.02, p = .047). Therefore, School should not be omitted in later analyses.

7.5. Two-tailed Independent Samples T-test of Gender on Baseline

To check whether Gender (Male and Female) can be omitted from the later analyses, we ran a two-tailed independent samples t-test on Baseline with male vs. female (Table 9).

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Baseline	Male	42	34.0714	13.80949	2.13085
	Female	33	41.0909	15.46238	2.69166

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Baseline	Equal variances assumed	2.531	.116	-2.073	73	.042	-7.01948	3.38630	-13.76837	-.27059
	Equal variances not assumed			-2.045	64.810	.045	-7.01948	3.43301	-13.87605	-.16291

Table 9. Independent Samples T-test on Baseline with Gender

Table 9 shows that the difference between the means of Baseline of Males (n = 42, M = 34.07, SD = 13.81) and Females (n = 33, M = 41.09, SD = 15.46) is significant ($t_{(73)}$ = -2.08, p = .042). Therefore, Gender should not be omitted from the analyses. Girls did more multiplications correct on the baseline test.

7.6. School (2) × Gender (2) ANOVA on Baseline with Age as covariate

To inspect the interaction effect of School and Gender on Baseline, we ran a School (2) × Gender (2) ANOVA on Baseline with Age as a covariate.

Descriptive Statistics

Dependent Variable: Baseline

School	Gender	Mean	Std. Deviation	N
Good Shepherd	Male	36.3077	15.59428	26
	Female	43.7273	15.53853	22
	Total	39.7083	15.84897	48
Chun Lei	Male	30.4375	9.65380	16
	Female	35.8182	14.56584	11
	Total	32.6296	11.94265	27
Total	Male	34.0714	13.80949	42
	Female	41.0909	15.46238	33
	Total	37.1600	14.87792	75

Table 10. Mean of School (2) × Gender (2) ANOVA on Baseline with Age as covariate

Tests of Between-Subjects Effects

Dependent Variable: Baseline

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2569.199[a]	4	642.300	3.255	.017	.157
Intercept	55.249	1	55.249	.280	.598	.004
Age	858.595	1	858.595	4.352	.041	.059
School	.282	1	.282	.001	.970	.000
Gender	414.992	1	414.992	2.103	.151	.029
School * Gender	1.268	1	1.268	.006	.936	.000
Error	13810.881	70	197.298			
Total	119945.000	75				
Corrected Total	16380.080	74				

a. R Squared = .157 (Adjusted R Squared = .109)

Table 11. School (2) × Gender (2) ANOVA on Baseline with Age as covariate

According to Table 11, the only significant difference is caused by Age ($F_{(1,70)} = 4.35$, $p = .041$). With age, pupils performed better, which is consistent with the result of (1.2). School, Gender, and their interaction had not influence on the Baseline. This is inconsistent with the findings in the t-test on School (1.4) and the t-test on Gender (1.5). This finding may indicate that the significant effects on the detailed level are spurious when more factors are added.

7.7. School (2) × Gender (2) ANOVA on FinMSco with Age as covariate

To inspect the interaction effect of School and Gender on FinMSco, we ran a School (2) × Gender (2) ANOVA on FinMSco with Age as a covariate.

Descriptive Statistics

Dependent Variable: Final Multiplication Score

School	Gender	Mean	Std. Deviation	N
Good Shepherd	Male	44.8462	17.75318	26
	Female	52.0000	19.62506	22
	Total	48.1250	18.78051	48
Chun Lei	Male	38.4375	13.52020	16
	Female	45.9091	13.95317	11
	Total	41.4815	13.94045	27
Total	Male	42.4048	16.40056	42
	Female	49.9697	17.94694	33
	Total	45.7333	17.39551	75

Table 12. Means of School (2) × Gender (2) on FinMSco with Age as covariate

Tests of Between-Subjects Effects

Dependent Variable: Final Multiplication Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2012.546[a]	4	503.136	1.728	.154	.090
Intercept	108.532	1	108.532	.373	.543	.005
Age	276.110	1	276.110	.948	.333	.013
School	42.657	1	42.657	.147	.703	.002
Gender	697.505	1	697.505	2.396	.126	.033
School * Gender	5.621	1	5.621	.019	.890	.000
Error	20380.121	70	291.145			
Total	179258.000	75				
Corrected Total	22392.667	74				

a. R Squared = .090 (Adjusted R Squared = .038)

Table 13. School (2) × Gender (2) ANOVA on FinMSco with Age as covariate

According to Table 13, none of the differences is significant. Therefore, the interaction of School and Gender with Age had no influence on the FinMSco. In comparing Table 11 and 13, we found that Age

had an effect on Baseline but no effect on FinMSco. Therefore, robot interaction probably diminished the effect of Age on learning. The result is consistent with that of the two-tailed bivariate correlation analysis on Age with Baseline and FinMSco (1.3), respectively.

In comparing Table 10 and 12, we found both girls and boys showed an increase in mean difference scores regardless of the schools they came from. For the pupils from Good Shepherd, boys and girls showed a similar mean increase (8.5 for boys and 8.3 for girls). For the pupils from Chun Lei, girls had greater increases in mean difference score (10.1 for girls vs. 8.0 for boys).

7.8. Conclusion

From the analyses of this part, we found:
1) Novelty had no effect on Fin_min_Base,
2) Robot Design diminished the effect of Age on learning,
3) We could not omit School and Gender because they exerted significant effects on Baseline and FinMSco, albeit inconsistently.

8. Effect of School, Robot Design, and Gender on Fin_min_Base

In the previous analyses, the between-subject variables (School, Gender) had inconsistent effects on the Baseline and FinMSco, which indicated that we cannot omit the above variables when exploring their interaction effect on Fin_min_Base (learning gain). In this part, we investigated factors (School, Gender, and Robot Design) that may contribute to Fin_min_Base (learning gains).

8.1. Fin_min_Base calculation

To study the learning gains, we calculated the difference score Fin_min_Base from FinMSco – Baseline. Whereas 64 pupils gained from robot tutoring, there were 11 who did not perform better but *worse* after robot exposure (Fin_min_Base = -1 to -35) (Table 14).

	School	Robot	Advancement	Baseline	BasePerf	MuSc1	MuSc2	MuSc3	ParTake	FinMSco	Fin_min_Base
1	1.00	1.00	2.0	27.00	2.00	10.00	.	.	1.00	24.00	-3.00
2	1.00	1.00	2.0	27.00	2.00	12.00	.	.	1.00	22.00	-5.00
3	1.00	1.00	2.0	28.00	2.00	13.00	.	.	1.00	27.00	-1.00
4	2.00	1.00	2.0	30.00	2.00	8.00	6	.	2.00	27.00	-3.00
5	2.00	2.00	2.0	24.00	2.00	.00	0	.00	.	13.00	-11.00
6	1.00	2.00	2.0	31.00	2.00	3.00	.	.	1.00	28.00	-3.00
7	2.00	2.00	2.0	36.00	2.00	.00	4	3.00	3.00	35.00	-1.00
8	2.00	3.00	1.0	22.00	2.00	.00	4	.	2.00	16.00	-6.00
9	1.00	3.00	2.0	35.00	2.00	9.00	.	.	1.00	33.00	-2.00
10	2.00	3.00	2.0	31.00	2.00	4.00	3	.	2.00	30.00	-1.00
11	1.00	3.00	4.0	65.00	1.00	11.00	.	.	1.00	30.00	-35.00

Table 14. Reverse influence of the robot on teaching the participants

8.2. School (2) × Robot (2) × Gender (2) ANOVA on Fin_min_Base with Age as covariate

To explore how the three between-subject factors (School, Gender, and Robot Design) affected the learning gains (Fin_min_Base), we ran a GLM univariate (ANOVA) of School (2) × Robot (2) × Gender (2) on Fin_min_Base with Age as a covariate (Table 15).

Tests of Between-Subjects Effects

Dependent Variable: Fin_min_Base

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1377.151[a]	12	114.763	1.091	.383	.174
Intercept	230.934	1	230.934	2.196	.143	.034
Age	106.528	1	106.528	1.013	.318	.016
School	26.302	1	26.302	.250	.619	.004
Robot	162.531	2	81.265	.773	.466	.024
Gender	12.784	1	12.784	.122	.729	.002
School * Robot	700.133	2	350.067	3.328	.042	.097
School * Gender	12.301	1	12.301	.117	.734	.002
Robot * Gender	54.255	2	27.127	.258	.773	.008
School * Robot * Gender	338.231	2	169.116	1.608	.209	.049
Error	6521.195	62	105.181			
Total	13411.000	75				
Corrected Total	7898.347	74				

a. R Squared = .174 (Adjusted R Squared = .015)

Table 15. School (2) × Robot (2) × Gender (2) ANOVA on Fin_min_Base

According to Table 15, the only significant difference is caused by the interaction of School × Robot ($F_{(2,62)} = 3.33$, $p = .042$). Therefore, we looked into the details of their interaction effect.

8.3. Two-tailed Independent Samples T-test of School and Robot Design on Fin_min_Base

To study the effect of School on Fin_min_Base, we ran a two-tailed independent samples t-test (Table 16).

Group Statistics

	School	N	Mean	Std. Deviation	Std. Error Mean
Fin_min_Base	Good Shepherd	48	8.4167	10.29115	1.48540
	Chun Lei	27	8.8519	10.59283	2.03859

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Fin_min_Base	Equal variances assumed	.333	.566	-.174	73	.862	-.43519	2.50176	-5.42118	4.55081
	Equal variances not assumed			-.173	52.716	.864	-.43519	2.52235	-5.49502	4.62465

Table 16. Independent Samples T-test of School on Fin_min_Base

In Table 16, the mean difference between the schools is not significant ($t_{(73)} = -.17$, $p = .86$), so School does not have a significant effect on Fin_min_Base. To study the effect of Robot on the Fin_min_Base, we ran three two-tailed independent t-tests of Robot Design (Humanoid-Puppy, Humanoid-Droid, and Puppy-Droid) on Fin_min_Base.

Group Statistics

	Robot	N	Mean	Std. Deviation	Std. Error Mean
Fin_min_Base	Humanoid	21	9.4762	10.71736	2.33872
	Puppy	27	9.6296	9.50364	1.82898

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Fin_min_Base	Equal variances assumed	.456	.503	-.052	46	.958	-.15344	2.92394	-6.03902	5.73214
	Equal variances not assumed			-.052	40.338	.959	-.15344	2.96897	-6.15238	5.84550

Table 17. Independent samples t-test on Fin_min_Base between Humanoid and Puppy

Group Statistics

	Robot	N	Mean	Std. Deviation	Std. Error Mean
Fin_min_Base	Humanoid	21	9.4762	10.71736	2.33872
	Droid	27	6.8148	10.95809	2.10889

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Fin_min_Base	Equal variances assumed	.244	.624	.843	46	.404	2.66138	3.15807	-3.69550	9.01825
	Equal variances not assumed			.845	43.582	.403	2.66138	3.14913	-3.68699	9.00975

Table 18. Independent samples t-test on Fin_min_Base between Humanoid and Droid

Group Statistics

	Robot	N	Mean	Std. Deviation	Std. Error Mean
Fin_min_Base	Puppy	27	9.6296	9.50364	1.82898
	Droid	27	6.8148	10.95809	2.10889

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Fin_min_Base	Equal variances assumed	.005	.943	1.008	52	.318	2.81481	2.79151	-2.78677	8.41640
	Equal variances not assumed			1.008	50.980	.318	2.81481	2.79151	-2.78944	8.41907

Table 19. Independent samples t-test on Fin_min_Base between Puppy and Droid

In Table 17, 18, and 19, the three comparisons among 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) yielded no significant effects (Humanoid-Puppy: $t_{(46)}$ = -.52, p = .96; Humanoid-Droid: $t_{(46)}$ = .84, p = .40; Puppy-Droid: $t_{(52)}$ = 1.01, p = .32). Neither School nor Robot had a significant effect on Fin_min_Base (learning gains).

We reran the analyses on the group that performed *worse* after robot tutoring. However, Robot and School again did not exert significant effects on Fin_min_Base.

In all, the differences between schools, gender, and types of robot improved nor worsened the children's learning gains as measured by Fin_min_Base.

8.4. Paired Samples T-test on Baseline and FinMSco*

As said, 64 children showed learning gains after robot intervention. We ran a paired samples t-test on Baseline and FinMSco to see how much those children gained.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Baseline	37.9844	64	15.29342	1.91168
	Final Multiplication Score	49.1406	64	16.36379	2.04547

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Baseline - Final Multiplication Score	-11.15625	7.97261	.99658	-13.14775	-9.16475	-11.195	63	.000

Table 20. Paired Samples T-test on Baseline and FiMSco

In Table 20, the difference between Baseline (n = 64, M = 37.98, SD = 1.91) and FinMSco (n = 64, M = 49.14, SD = 2.05) is highly significant ($t_{(63)}$ = -11.20, p = .000). On average, those who learned from the robot did over one-third better compared to baseline.

8.5. Conclusion

We found:
1) Those who learned had an average of more than one-third gain after robot tutoring,
2) None of the factors (School, Gender, Robot Design) contributed to the learning gain.

9. Effects of Advancement and Partake on Fin_min_Base

We analyzed the level of Advancement of the various pupils and the number of times they participated in the tutoring sessions (Partake: 1, 2, or 3 times) for their effects on Fin_min_Base.

9.1. Advancement calculation

Before we explored the effect of level of advancement on Fin_min_Base, we categorized pupils according to their Baseline results. First, we calculated the average Baseline value (n = 75, M = 37.16, SD = 12.88) (Table 21).

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Baseline	75	10.00	78.00	37.1600	14.87792
Valid N (listwise)	75				

Table 21. Standardized value of Baseline

Then we categorized the students into four groups. Those who scored lower than one standard deviation below average (< 22.28) were categorized as ‘challenged’ students. Those between one negative standard deviation and the average were categorized as ‘below average.’ Those between average and one positive standard deviation were categorized as ‘above average,’ and those beyond one positive standard deviation were categorized as ‘advanced’ students (see Table 22, upper panel).

Categorization Name	Criteria	Number of students
Challenged	Baseline <= 22.8	11
Below Average	22.8 < Baseline <= 37.16	34
Above Average	37.16 < Baseline <= 52.04	19
Advanced	Baseline > 52.04	11

9.2. Inspection of the students who learned worse

We looked into the participants who suffered from learning loss after robot exposure and found that ten of them were categorized as Below Average and Challenged, the remaining one being Advanced (Table 22, lower panel). For both types of students, we know that (for different reasons) they can be easily distracted and have learning disabilities (e.g., Beckmann & Minnaert, 2018).

Index	School	Robot	Gender	Advancement	Partake	Fin_min_Base
6	1	1	0	2	1	-3
7	1	1	1	2	1	-5
8	1	1	0	2	1	-1
11	2	1	0	2	2	-3
27	2	2	0	2	3	-11

31	1	2	0	2	1	-3
41	2	2	1	2	3	-1
53	2	3	0	1	2	-6
60	1	3	1	2	1	-2
63	2	3	1	2	2	-1
70	1	3	1	4	1	-35

Table 22. Advancement distribution (upper panel) and details of worse performers after robot tutoring (lower panel)

9.3. One-way ANOVA of Partake on Fin_min_Base

To explore the effect of Partake (the number of tutoring sessions) on Fin_min_Base (learning gains), we ran a one-way ANOVA on Fin_min_Base with Partake as the independent variable.

Descriptives

Fin_min_Base

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	Between-Component Variance
						Lower Bound	Upper Bound			
1.00		48	8.4167	10.29115	1.48540	5.4284	11.4049	-35.00	27.00	
2.00		16	7.6875	9.82323	2.45581	2.4531	12.9219	-6.00	31.00	
3.00		10	12.7000	10.03383	3.17298	5.5222	19.8778	-1.00	34.00	
Total		74	8.8378	10.14285	1.17908	6.4879	11.1877	-35.00	34.00	
Model	Fixed Effects			10.16152	1.18125	6.4825	11.1932			
	Random Effects				1.18125[a]	3.7553[a]	13.9204[a]			-.72694

a. Warning: Between-component variance is negative. It was replaced by 0.0 in computing this random effects measure.

ANOVA

Fin_min_Base

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	178.850	2	89.425	.866	.425
Within Groups	7331.204	71	103.256		
Total	7510.054	73			

Table 23. One-way ANOVA on Fin_min_Base with Partake as independent variable

Table 23 shows that the differences among the levels of Partake are not significant ($F_{(2,71)}$ = .866, p = .425). Therefore, Partake had no influence on Fin_min_Base. Entering more robot-tutoring sessions did not improve learning performance.

9.4. One-way ANOVA of Advancement on Fin_min_Base

To explore the effects of the Advancement on Fin_min_Base, we ran the one-way ANOVA.

Descriptives

Fin_min_Base

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	Between-Component Variance
						Lower Bound	Upper Bound			
Challenged (Baseline < 22.28)		11	13.5455	12.25858	3.69610	5.3100	21.7809	-6.00	34.00	
Below Average (22.28 < Baseline < 37.16)		34	6.6765	9.34108	1.60198	3.4172	9.9357	-11.00	27.00	
Above Average (37.16 < Baseline < 52.04)		19	10.2632	6.49651	1.49040	7.1319	13.3944	.00	21.00	
Gifted (Baseline > 52.04)		11	6.5455	15.04236	4.53544	-3.5601	16.6511	-35.00	24.00	
Total		75	8.5733	10.33123	1.19295	6.1963	10.9503	-35.00	34.00	
Model	Fixed Effects			10.21224	1.17921	6.2221	10.9246			
	Random Effects				1.57733	3.5536	13.5931			3.50939

ANOVA

Fin_min_Base

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	493.767	3	164.589	1.578	.202
Within Groups	7404.580	71	104.290		
Total	7898.347	74			

Table 24. One-way ANOVA on Fin_min_Base with Advancement as independent variable

In Table 24, the differences among the levels of Advancement are not significant ($F_{(3,71)}$ = 1.58, p = .202). Therefore, Advancement had no influence on the Fin_min_Base. No matter how good children were initially, this did not affect their one-third learning gain on average.

9.5. Partake (3) × Advancement (4) ANOVA on Fin_min_Base with Age as covariate

From Section 3.3 and 3.4, we found neither Partake nor Advancement had effect on Fin_min_Base. To inspect whether the interaction of Partake and Advancement affected Fin_min_Base, we ran a GLM univariate (ANOVA) of Partake (3) × Advancement (4) on Fin_min_Base with Age as a covariate. Yet, Table 25 shows that the interaction of Partake and Advancement had no significant effect on Fin_min_Base.

Tests of Between-Subjects Effects

Dependent Variable: Fin_min_Base

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	1000.731[a]	11	90.976	.867	.577	.133
Intercept	365.277	1	365.277	3.479	.067	.053
Age	187.796	1	187.796	1.789	.186	.028
Advancement	144.962	3	48.321	.460	.711	.022
ParTake	93.751	2	46.876	.446	.642	.014
Advancement * ParTake	350.867	5	70.173	.668	.649	.051
Error	6509.323	62	104.989			
Total	13290.000	74				
Corrected Total	7510.054	73				

a. R Squared = .133 (Adjusted R Squared = -.021)

Table 25. Partake (3) × Advancement (4) ANOVA on Fin_min_Base with Age as a covariate

The conclusions of Section 3.3, 3.4, and 3.5 are consistent: Partake and Advancement had no effect on Fin_min_Base. Therefore, every student could benefit from robot tutoring, regardless of their academic performance and the number of times they worked with the robot.

9.6. Paired Samples T-test of Partake on Baseline and FinMSco*

In the above analyses, we did not find effects of Partake on Fin_min_Base. It made us assume that one robot intervention is enough and that subsequent sessions are superfluous. To verify this idea, we ran three paired samples t-tests of Partake on Baseline and FinMSco.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Baseline	39.7083	48	15.84897	2.28760
	Final Multiplication Score	48.1250	48	18.78051	2.71073

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Baseline - Final Multiplication Score	-8.41667	10.29115	1.48540	-11.40491	-5.42843	-5.666	47	.000

Table 26. Paired Samples T-test on Baseline and FinMSco as variables with Partake = 1 session

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Baseline	35.3750	16	11.80889	2.95222
	Final Multiplication Score	43.0625	16	14.94643	3.73661

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Baseline - Final Multiplication Score	-7.68750	9.82323	2.45581	-12.92193	-2.45307	-3.130	15	.007

Table 27. Paired samples t-test on Baseline and FinMSco as variables with Partake = 2 sessions

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Baseline	28.6364	11	11.49150	3.46482
	Final Multiplication Score	39.1818	11	12.66348	3.81818

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Baseline - Final Multiplication Score	-10.54545	11.90264	3.58878	-18.54175	-2.54916	-2.938	10	.015

Table 28. Paired samples t-test on Baseline and FinMSco as variables with Partake = 3 sessions

We summarised the results of the t-test in Table 29:

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

(1) Fin_min_Base = FinMSco – Baseline
(2) Per_Fin_Min_Base = Fin_min_Base / Baseline

Table 29. Improvement after robot tutoring

Counted as the absolute number of correctly answered multiplications, we could see that those who worked with the robot once improved by 8.42 correct. Those who did two sessions had a 7.68 improvement. Those who interacted thrice had a 10.54 improvement. According to Oneway ANOVA (Section 3.3), however, the differences between the number of sessions followed were not significant for Fin_min_Base.

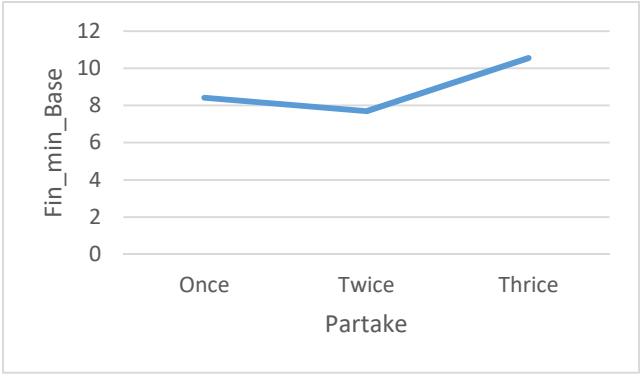

Figure 1. Improvement in absolute number of correct answers: Partake and Fin_min_Base

Yet, when we calculated the improvement as a percentage of the Baseline, Partake did exact positive effects on Per_Fin_min_Base (see Section 3.9). Those who had one session with the robot improved 21.20%, those who did two sessions improved with 21.73%, and those who had three sessions gained 36.83%.

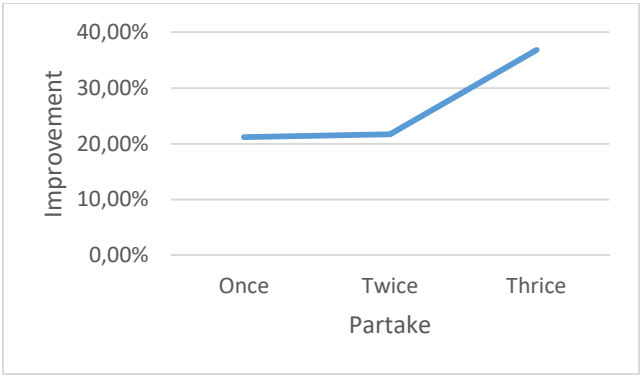

Figure 2. Improvement as a percentage of the Baseline: Partake and Per_Fin_min_Base

9.7. Per_Fin_min_Base calculation

In view of Section 3.6, we calculated the percentage of learning gains Per_Fin_min_Base from Fin_min_Base / Baseline (Table 29).

School	Robot	Advancement	Baseline	BasePerf	MuSc1	MuSc2	MuSc3	ParTake	FinMSco	Fin_min_Base	Per_Fin_min_Base	Machine_like
1.00	1.00	1.0	10.00	3.00	12.00	.	.	1.00	21.00	11.00	1.10	6.00
2.00	1.00	1.0	22.00	2.00	7.00	7	.	2.00	53.00	31.00	1.41	3.00
2.00	1.00	1.0	17.00	3.00	.00	1	1.00	3.00	51.00	34.00	2.00	3.00
1.00	1.00	2.0	25.00	2.00	10.00	.	.	1.00	28.00	3.00	.12	6.00
1.00	1.00	2.0	26.00	2.00	15.00	.	.	1.00	28.00	2.00	.08	6.00
1.00	1.00	2.0	27.00	2.00	10.00	.	.	1.00	24.00	-3.00	-.11	6.00
1.00	1.00	2.0	27.00	2.00	12.00	.	.	1.00	22.00	-5.00	-.19	5.00
1.00	1.00	2.0	28.00	2.00	13.00	.	.	1.00	27.00	-1.00	-.04	1.00
1.00	1.00	2.0	28.00	2.00	11.00	.	.	1.00	30.00	2.00	.07	6.00
2.00	1.00	2.0	26.00	2.00	7.00	10	.	2.00	43.00	17.00	.65	.
2.00	1.00	2.0	30.00	2.00	8.00	6	.	2.00	27.00	-3.00	-.10	1.00
1.00	1.00	3.0	39.00	2.00	13.00	.	.	1.00	60.00	21.00	.54	4.00
1.00	1.00	3.0	42.00	2.00	12.00	.	.	1.00	47.00	5.00	.12	6.00
1.00	1.00	3.0	43.00	2.00	9.00	.	.	1.00	45.00	2.00	.05	4.00
1.00	1.00	3.0	44.00	2.00	17.00	.	.	1.00	53.00	9.00	.20	5.00
2.00	1.00	3.0	40.00	2.00	.00	6	2.00	3.00	47.00	7.00	.18	6.00
2.00	1.00	3.0	42.00	2.00	12.00	2	4.00	3.00	58.00	16.00	.38	1.00
1.00	1.00	3.0	52.00	2.00	18.00	.	.	1.00	69.00	17.00	.33	6.00
1.00	1.00	4.0	54.00	2.00	15.00	.	.	1.00	70.00	16.00	.30	6.00
1.00	1.00	4.0	58.00	2.00	16.00	.	.	1.00	71.00	13.00	.22	6.00

Table 30. Per_Fin_min_Base calculation

9.8. Correlations of Baseline / Partake / Baseline & Partake on Per_Fin_min_Base*

In Section 3.7, we found significant improvement in the percentage of learning gains. However, we had insufficient information to account for the different improvement caused by Partake since the pupils had a different Baseline (once: 39.85, twice: 35.38, thrice: 28.6). We assumed that the difference may come from the different levels of Baseline with those who had low Baseline learning relatively more. To verify the assumption, we ran correlation analysis on the Baseline and Per_Fin_min_Base (Table 30) and found a significant correlation (r = -.530, p = .000), indicating that those with poorer Baseline performance learned relatively more as measured by Per_Fin_min_Base.

Descriptive Statistics

	Mean	Std. Deviation	N
Baseline	38.0968	15.43331	62
Per_Fin_min_Base	.3835	.38453	62

Correlations

		Baseline	Per_Fin_min_Base
Baseline	Pearson Correlation	1	-.530**
	Sig. (2-tailed)		.000
	N	62	62
Per_Fin_min_Base	Pearson Correlation	-.530**	1
	Sig. (2-tailed)	.000	
	N	62	62

**. Correlation is significant at the 0.01 level (2-tailed).

Table 31. Correlation of Baseline and Per_Fin_min_Base

Thus, we observed that both Baseline and Partake contributed to the relative learning gain expressed in percentages (Per_Fin_min_Base). To verify this observation, we ran a Linear Regression Analysis of Baseline and Partake on Per_Fin_min_Base (Table 32).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.666	.156		4.262	.000
	Baseline	-.012	.003	-.473	-4.277	.000
	Number of times Ss participated	.114	.057	.219	1.978	.053

a. Dependent Variable: Per_Fin_min_Base

Table 32. Linear Regression of Baseline and Partake on Per_Fin_min_Base

However, the correlation between Partake and Per_Fin_min_Base was not significant. Therefore, we eliminated the Partake factor and concluded that those with worse Baseline benefited relatively more from the robot tutoring, not necessarily from doing it more often.

9.9. One-way ANOVA of Advancement on Per_Fin_min_Base

To strengthen the conclusion that we made in Section 3.8, we ran a One-way ANOVA of Advancement on Per_Fin_min_Base, resulting into Table 33.

Descriptives

Per_Fin_min_Base

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	Between-Component Variance
						Lower Bound	Upper Bound			
Challenged (Baseline < 22.28)		11	.7929	.67497	.20351	.3395	1.2464	-.27	2.00	
Below Average (22.28 < Baseline < 37.16)		34	.2266	.32832	.05631	.1120	.3411	-.46	.93	
Above Average (37.16 < Baseline < 52.04)		19	.2233	.14370	.03297	.1540	.2925	.00	.54	
Gifted (Baseline > 52.04)		11	.1089	.23744	.07159	-.0507	.2684	-.54	.38	
Total		75	.2915	.40956	.04729	.1973	.3858	-.54	2.00	
Model	Fixed Effects			.35699	.04122	.2093	.3737			
	Random Effects				.14068	-.1562	.7392			.05785

ANOVA

Per_Fin_min_Base

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.364	3	1.121	8.800	.000
Within Groups	9.048	71	.127		
Total	12.413	74			

Table 33. One-way ANOVA on Per_Fin_min_Base with Advancement as independent variable

Table 33 shows that the means of Per_Fin_min_Base decrease with the increase of Advancement level (Challenged: n = 11, M = .80, SD = .67; Below Average: n = 34, M = .23, SD = .33; Above Average: n = 19, M = .22, SD = .14; Advanced; n = 11, M = .11, SD = .24). The differences are significant ($F_{(3,71)}$ = 8.80, p = .000), which is consistent with the results of Section 3.8 that poor students benefited relatively more from robot tutoring than others.

9.10. Two-tailed Independent T-tests of Advancement on Per_Fin_min_Base

We ran six two-tailed independent t-tests of Advancement (Challenged - Below Average, Challenged - Above Average, Challenged - Advanced, Below Average – Above Average, Below Average – Advanced, Above Average - Advanced) on Per_Fin_min_Base. Table 34 –Table 39 show the results.

Group Statistics

	Advancement	N	Mean	Std. Deviation	Std. Error Mean
Per_Fin_min_Base	Challenged (Baseline < 22.28)	11	.7929	.67497	.20351
	Below Average (22.28 < Baseline < 37.16)	34	.2266	.32832	.05631

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Per_Fin_min_Base	Equal variances assumed	9.231	.004	3.759	43	.001	.56636	.15067	.26251	.87022
	Equal variances not assumed			2.682	11.569	.021	.56636	.21116	.10438	1.02834

Table 34. Independent Sample t-test of Challenged - Below Average on Per_Fin_min_Base

Group Statistics

	Advancement	N	Mean	Std. Deviation	Std. Error Mean
Per_Fin_min_Base	Challenged (Baseline < 22.28)	11	.7929	.67497	.20351
	Above Average (37.16 < Baseline < 52.04)	19	.2233	.14370	.03297

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Per_Fin_min_Base	Equal variances assumed	20.548	.000	3.584	28	.001	.56966	.15894	.24410	.89523
	Equal variances not assumed			2.763	10.528	.019	.56966	.20617	.11340	1.02593

Table 35. Independent Sample t-test of Challenged - Above Average on Per_Fin_min_Base

Group Statistics

	Advancement	N	Mean	Std. Deviation	Std. Error Mean
Per_Fin_min_Base	Challenged (Baseline < 22.28)	11	.7929	.67497	.20351
	Gifted (Baseline > 52.04)	11	.1089	.23744	.07159

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Per_Fin_min_Base	Equal variances assumed	8.873	.007	3.171	20	.005	.68405	.21574	.23403	1.13407
	Equal variances not assumed			3.171	12.438	.008	.68405	.21574	.21583	1.15228

Table 36. Independent Sample t-test of Challenged - Advanced on Per_Fin_min_Base

Group Statistics

	Advancement	N	Mean	Std. Deviation	Std. Error Mean
Per_Fin_min_Base	Below Average (22.28 < Baseline < 37.16)	34	.2266	.32832	.05631
	Above Average (37.16 < Baseline < 52.04)	19	.2233	.14370	.03297

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Per_Fin_min_Base	Equal variances assumed	11.545	.001	.042	51	.967	.00330	.07950	-.15630	.16290
	Equal variances not assumed			.051	48.956	.960	.00330	.06525	-.12782	.13442

Table 37. Independent Sample t-test of Below Average – Above Average on Per_Fin_min_Base

Group Statistics

	Advancement	N	Mean	Std. Deviation	Std. Error Mean
Per_Fin_min_Base	Below Average (22.28 < Baseline < 37.16)	34	.2266	.32832	.05631
	Gifted (Baseline > 52.04)	11	.1089	.23744	.07159

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Per_Fin_min_Base	Equal variances assumed	3.772	.059	1.096	43	.279	.11769	.10738	-.09887	.33425
	Equal variances not assumed			1.292	23.476	.209	.11769	.09108	-.07051	.30589

Table 38. Independent Sample t-test of Below Average – Advanced on Per_Fin_min_Base

Group Statistics

	Advancement	N	Mean	Std. Deviation	Std. Error Mean
Per_Fin_min_Base	Above Average (37.16 < Baseline < 52.04)	19	.2233	.14370	.03297
	Gifted (Baseline > 52.04)	11	.1089	.23744	.07159

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Per_Fin_min_Base	Equal variances assumed	.328	.572	1.652	28	.110	.11439	.06925	-.02746	.25624
	Equal variances not assumed			1.451	14.333	.168	.11439	.07882	-.05429	.28307

Table 39. Independent Sample t-test of Above Average – Advanced on Per_Fin_min_Base

Table 34 -Table 36 show that the percentage of learning gain (Per_Fin_min_Base) of pupils that are Challenged (n = 11, M = .79, SD = 0.67) was significantly higher than those of level of Below Average (n = 34, M = .23, SD = 0.33), Above Average (n = 19, M = .22, SD = 0.14) or Advanced (n = 11, M = .11, SD = 0.24) (Challenged – Below Average: $t_{(43)} = 3.76$, $p = .001$; Challenged – Above Average: $t_{(28)} = 3.59$, $p = .001$; Challenged – Advanced: $t_{(20)} = 3.17$, $p = .005$). However, Table 37 - Table 39 also show that the differences between Below Average, Above Average and Advanced were not significant (Below Average – Above Average: $t_{(51)} = .042$, $p = .967$; Below Average – Advanced: $t_{(43)} = 1.10$, $p = .28$; Above Average – Advanced: $t_{(28)} = 1.65$, $p = .11$). It indicated that the significant difference of the One-way ANOVA (Section 3.9) came from being Challenged as compared to the other three levels. Thus, extremely poor performers benefited most from robot tutoring.

9.11. Conclusion

From the analyses in Section 3, we learned:

1) Those with extremely low Baseline benefited relatively more from robot tutoring expressed as a percentage (Per_Fin_min_Base), regardless of the number of sessions they took (Partake).

10. Effects of Robot Design, Representation, and Social Role

We analyzed the effect of Robot Design (3) × Representation (3) × Social Role (6) on children's experience of the robot tutor. We took out those who did not fill out the questionnaire and kept 72 valid cases. To see whether the participants experienced the different robots as the entities they were supposed to resemble (manipulation check), each 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).

10.1. MANOVA of Robot Design (3) on Human-like, Animal-like, and Machine-like*

We ran a General Linear Model Multivariate Analysis (MANOVA) of Robot (3) on Human-like, Animal-like, and Machine-like, resulting into Table 40.

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.936	328.109[b]	3.000	67.000	.000	.936
	Wilks' Lambda	.064	328.109[b]	3.000	67.000	.000	.936
	Hotelling's Trace	14.691	328.109[b]	3.000	67.000	.000	.936
	Roy's Largest Root	14.691	328.109[b]	3.000	67.000	.000	.936
Robot	Pillai's Trace	.460	6.779	6.000	136.000	.000	.230
	Wilks' Lambda	.573	7.167[b]	6.000	134.000	.000	.243
	Hotelling's Trace	.686	7.548	6.000	132.000	.000	.255
	Roy's Largest Root	.586	13.288[c]	3.000	68.000	.000	.370

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	Machine_like	4.726[a]	2	2.363	.577	.564	.016
	Human_like	48.149[b]	2	24.075	8.321	.001	.194
	Animal_like	75.646[c]	2	37.823	12.414	.000	.265
Intercept	Machine_like	1256.264	1	1256.264	306.783	.000	.816
	Human_like	504.325	1	504.325	174.316	.000	.716
	Animal_like	550.737	1	550.737	180.759	.000	.724
Robot	Machine_like	4.726	2	2.363	.577	.564	.016
	Human_like	48.149	2	24.075	8.321	.001	.194
	Animal_like	75.646	2	37.823	12.414	.000	.265
Error	Machine_like	282.552	69	4.095			
	Human_like	199.629	69	2.893			
	Animal_like	210.229	69	3.047			
Total	Machine_like	1554.000	72				
	Human_like	718.000	72				
	Animal_like	881.000	72				
Corrected Total	Machine_like	287.278	71				
	Human_like	247.778	71				
	Animal_like	285.875	71				

Table 40. General Linear Model Multivariate Analysis (MANOVA) of Robot (3) on Human-like, Animal-like, and Machine-like

Table 40 shows that pupils judged their robots as significantly different in what they represented: The effects of Robot type on the rating of Representation was significant (Wilks' Lambda = .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.

10.2. Two-tailed Independent T-tests of Robot on Human-like and Animal-likeness*

We ran six two-tailed independent t-tests of Robot Design (Humanoid-Puppy, Humanoid-Droid, and Puppy-Droid) on ratings of Human-like and Animal-likeness. Table 41 –Table 46 show the results.

Group Statistics

	Robot	N	Mean	Std. Deviation	Std. Error Mean
Human_like	Humanoid	19	3.89474	1.911798	.438596
	Puppy	26	1.88462	1.423430	.279158

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Human_like	Equal variances assumed	4.387	.042	4.047	43	.000	2.010121	.496668	1.008496	3.011747
	Equal variances not assumed			3.866	31.782	.001	2.010121	.519900	.950836	3.069407

Table 41. Independent Sample t-test of Humanoid - Puppy on Human-likeness

Group Statistics

	Robot	N	Mean	Std. Deviation	Std. Error Mean
Human_like	Humanoid	19	3.89474	1.911798	.438596
	Droid	27	2.25926	1.788695	.344235

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Human_like	Equal variances assumed	.168	.684	2.968	44	.005	1.635478	.550998	.525014	2.745941
	Equal variances not assumed			2.933	37.227	.006	1.635478	.557552	.506002	2.764953

Table 42. Independent Sample t-test of Humanoid - Droid on Human-likeness

Group Statistics

	Robot	N	Mean	Std. Deviation	Std. Error Mean
Human_like	Puppy	26	1.88462	1.423430	.279158
	Droid	27	2.25926	1.788695	.344235

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Human_like	Equal variances assumed	3.598	.064	-.842	51	.404	-.374644	.445119	-1.268257	.518969
	Equal variances not assumed			-.845	49.278	.402	-.374644	.443200	-1.265161	.515873

Table 43. Independent Sample t-test of Puppy - Droid on Human-likeness

Group Statistics

	Robot	N	Mean	Std. Deviation	Std. Error Mean
Animal_like	Humanoid	19	1.9474	1.58021	.36253
	Puppy	26	4.2308	1.81786	.35651

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Animal_like	Equal variances assumed	2.546	.118	-4.392	43	.000	-2.28340	.51984	-3.33176	-1.23504
	Equal variances not assumed			-4.491	41.622	.000	-2.28340	.50845	-3.30978	-1.25702

Table 44. Independent Sample t-test of Humanoid - Puppy on Animal-likeness

Group Statistics

	Robot	N	Mean	Std. Deviation	Std. Error Mean
Animal_like	Humanoid	19	1.9474	1.58021	.36253
	Droid	27	2.2222	1.78311	.34316

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Animal_like	Equal variances assumed	.967	.331	-.539	44	.593	-.27485	.50997	-1.30263	.75292
	Equal variances not assumed			-.551	41.591	.585	-.27485	.49918	-1.28254	.73283

Table 45. Independent Sample t-test of Humanoid - Droid on Animal-likeness

Group Statistics

	Robot	N	Mean	Std. Deviation	Std. Error Mean
Animal_like	Puppy	26	4.2308	1.81786	.35651
	Droid	27	2.2222	1.78311	.34316

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Animal_like	Equal variances assumed	.317	.576	4.061	51	.000	2.00855	.49465	1.01550	3.00160
	Equal variances not assumed			4.059	50.830	.000	2.00855	.49483	1.01505	3.00205

Table 46. Independent Sample t-test of Puppy - Droid on Animal-likeness

Table 41 shows that the 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). Table 42 shows that the 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). Table 43 shows that the 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). Table 44 shows that the 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). Table 45 shows that 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). Table 46 shows that 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 human impression.

10.3. Regression of Human-like, Animal-like, and Machine-like on Fin_min_Base / Per_Fin_min_Base

To check whether the students' perceptions of the Representation (Human-like, Animal-like, and Machine-like) had an effect on learning (i.e. Fin_min_Base), we did regression analysis of Human-like, Animal-like, and Machine-like on Fin_min_Base and Per_Fin_min_Base, respectively (Table

47-48). However, no significant relationship was established with Fin_min_Base (Human-like: t = -.47, p = .64; Animal-like: t = -.52, p = .61; Machine-like: t = -.50, p = .62), nor with Per_Fin_min_Base (Human-like: t = -.26, p = .80; Animal-like: t = -1.16, p = .25; Machine-like: t = -.71, p = .48).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	11.882	4.808		2.471	.016
	Machine_like	-.339	.674	-.065	-.503	.617
	Human_like	-.327	.696	-.059	-.469	.640
	Animal_like	-.343	.663	-.066	-.517	.607

a. Dependent Variable: Fin_min_Base

Table 47. Regression of Representation on Fin_min_Base

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	.478	.189		2.524	.014
	Machine_like	-.019	.027	-.092	-.712	.479
	Human_like	-.007	.027	-.032	-.256	.799
	Animal_like	-.030	.026	-.147	-1.159	.250

Table 48. Regression of Representation on Per_Fin_min_Base

Combined with the results from Sections 4.2 and 4.3, students perceived the robot as we expected but their perception had no effect on learning gains; not in absolute numbers of correct answers and not as a percentage of improvement from the Baseline. Thus, the design of the embodiment did not matter for learning multiplication tables.

10.4. MANOVA of Social Role on Human-like, Animal-like, and Machine-like, separately*

We ran three GLM Multivariate Analyses (MANOVA) of Social Role (Friend, Classmate, Teacher, Acquaintance, Stranger) on Human-like, Animal-like, and Machine-like as separate dependents for effects to become significant easily. Results are in Table 49 – 51.

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.771	34.158[b]	6.000	61.000	.000
	Wilks' Lambda	.229	34.158[b]	6.000	61.000	.000
	Hotelling's Trace	3.360	34.158[b]	6.000	61.000	.000
	Roy's Largest Root	3.360	34.158[b]	6.000	61.000	.000
Human_like	Pillai's Trace	.402	.948	30.000	325.000	.548
	Wilks' Lambda	.648	.938	30.000	246.000	.563
	Hotelling's Trace	.468	.927	30.000	297.000	.580
	Roy's Largest Root	.205	2.218[c]	6.000	65.000	.052

a. Design: Intercept + Human_like

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Table 49. GLM Multivariate Analysis (MANOVA) of Social Role on ratings of Human-like

Table 49 shows that the different Social Roles are not significant for Human-likeness ($F_{(30,246)}$ = .94, p = .563).

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.872	68.957[b]	6.000	61.000	.000
	Wilks' Lambda	.128	68.957[b]	6.000	61.000	.000
	Hotelling's Trace	6.783	68.957[b]	6.000	61.000	.000
	Roy's Largest Root	6.783	68.957[b]	6.000	61.000	.000
Animal_like	Pillai's Trace	.491	1.180	30.000	325.000	.242
	Wilks' Lambda	.584	1.180	30.000	246.000	.246
	Hotelling's Trace	.591	1.170	30.000	297.000	.253
	Roy's Largest Root	.284	3.080[c]	6.000	65.000	.010

a. Design: Intercept + Animal_like

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Table 50. GLM Multivariate Analysis (MANOVA) of Social Role on ratings of Animal-like

Table 50 shows that Social Roles had no significant effect on Animal-likeness ($F_{(30,246)}$ = 1.18, p = .246).

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.883	76.811[b]	6.000	61.000	.000
	Wilks' Lambda	.117	76.811[b]	6.000	61.000	.000
	Hotelling's Trace	7.555	76.811[b]	6.000	61.000	.000
	Roy's Largest Root	7.555	76.811[b]	6.000	61.000	.000
Machine_like	Pillai's Trace	.639	1.587	30.000	325.000	.029
	Wilks' Lambda	.462	1.746	30.000	246.000	.012
	Hotelling's Trace	.959	1.899	30.000	297.000	.004
	Roy's Largest Root	.710	7.693[c]	6.000	65.000	.000

a. Design: Intercept + Machine_like

b. Exact statistic

c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Friend	25.961[a]	5	5.192	1.267	.289
	Classmate	14.997[b]	5	2.999	1.062	.389
	Teacher	45.313[c]	5	9.063	2.752	.026
	Acquaintance	3.138[d]	5	.628	.302	.910
	Stranger	7.509[e]	5	1.502	.417	.835
	Machine	93.856[f]	5	18.771	5.527	.000
Intercept	Friend	357.622	1	357.622	87.263	.000
	Classmate	246.676	1	246.676	87.378	.000
	Teacher	372.274	1	372.274	113.049	.000
	Acquaintance	147.410	1	147.410	70.857	.000
	Stranger	263.867	1	263.867	73.244	.000
	Machine	360.315	1	360.315	106.096	.000
Machine_like	Friend	25.961	5	5.192	1.267	.289
	Classmate	14.997	5	2.999	1.062	.389
	Teacher	45.313	5	9.063	2.752	.026
	Acquaintance	3.138	5	.628	.302	.910
	Stranger	7.509	5	1.502	.417	.835
	Machine	93.856	5	18.771	5.527	.000

Table 51. GLM Multivariate Analysis (MANOVA) of Social Role on ratings of Machine-like

Table 51 shows that 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 are 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.

10.5. Conclusion

From the analyses in Section 4, we found:
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 is a trivial finding.

11. Reliability Analysis of Questionnaire Items (# = 43)

In this section, we scrutinize the convergent and divergent validity of measuring the experiential factors.

11.1. Recoding Counter-indicative Items

The counter-indicative items on the questionnaire were recoded into new variables (1→6, 2→5, 3→4, 4→3, 5→2, 6→1). Items Eng_3, Eng_5, Anth_1, Anth_4, Real_3, Rel_3, Aff_3, Aff_4, Use_Int_2 were recoded into Eng_3CR, Eng_5CR, Anth_1CR, Anth_4CR, Real_3CR, Rel_3CR, Aff_3CR, Aff_4CR, and Use_Int_2CR.

11.2. Convergent Validity

For the test on convergent validity (do items on a scale measure the same construct?) we calculated Cronbach's Alpha. For divergent validity (do items on different scales not measure the same construct?), we did Principal Component Analysis (PCA).

11.2.1. Engagement scale

We ran reliability analysis on Engagement (5 items) and found that Cronbach's Alpha = .79. The result was high enough to confirm that the scale of Engagement is reliable.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.785	.791	5

Table 52. Convergent Validity of Engagement

11.2.2. Bonding scale

We ran reliability analysis on Bonding (5 items) and found that Cronbach's Alpha = .88. The result was high enough that we could say the reliability of the scale of Bonding was reliable.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.876	.877	5

Table 53. Convergent Validity of Bonding

11.2.3. Anthropomorphism scale

We ran reliability analysis on Anthropomorphism and found Cronbach's Alpha to be very low (.34). We enhanced the reliability by taking out Anth_4CR (Table 54).

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.335	.334	4

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Anth_1CR	10.4493	14.810	.177	.056	.274
Anth_2	11.5507	10.839	.433	.478	-.077[a]
Anth_3	11.5942	12.568	.286	.478	.135
Anth_4CR	11.1014	18.622	-.112	.064	.574

a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

Table 54. Convergent Validity of Anthropomorphism

After taking out the Anth_4CR, we ran the analysis again and got Table 55.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.574	.562	3

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Anth_1CR	6.6522	13.613	.121	.040	.807
Anth_2	7.7536	7.747	.604	.478	.088
Anth_3	7.7971	8.752	.490	.463	.297

Table 55. Convergent Validity of Anthropomorphism after dismissing Anth_4CR

The reliability of Anthropomorphism still was not satisfactory (Cronbach's Alpha = .57). We could further improve the reliability by taking out Anth_1CR (Table 55). However, because there were only two items left on the scale, we calculated Spearman-Brown Correlation (r = .68, p = .000)and got the results of Table 56.

Correlations

			Anth_2	Anth_3
Spearman's rho	Anth_2	Correlation Coefficient	1.000	.684**
		Sig. (2-tailed)	.	.000
		N	69	69
	Anth_3	Correlation Coefficient	.684**	1.000
		Sig. (2-tailed)	.000	.
		N	69	70

**. Correlation is significant at the 0.01 level (2-tailed).

Table 56. Spearman-Brown Correlation with Anth_2 and Anth_3

The items Anth_2 and Anth_3 were significantly correlated (r = .68, p = .000).

11.2.4. Realism scale

We ran reliability analysis on Realism and found Cronbach's Alpha to be low (.37). We enhanced the reliability by taking out Real_3CR (Table 57).

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.367	.287	4

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Real_1	10.6377	11.646	.420	.295	.016
Real_2	10.5507	10.780	.438	.372	-.035[a]
Real_3CR	9.1159	25.516	-.393	.177	.752
Real_4	10.5652	11.043	.510	.519	-.100[a]

a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

Table 57. Convergent Validity of Realism

After taking out Real_3CR, we ran the analysis again and found that the reliability of Realism (3 items) improved (Cronbach's Alpha = .75).

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.752	.754	3

Table 58. Convergent Validity of Realism without Real_3CR

11.2.5. Relevance scale

We ran reliability analysis on Relevance and found that Cronbach's Alpha = .73. The result shows that the reliability of the items was positive.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.734	.728	4

Table 59. Convergent Validity of Relevance

11.2.6. Affordance scale

We ran reliability analysis on Affordance and found that Cronbach's Alpha was very low (.13). We enhanced the reliability by taking out Aff_3CR (Table 60).

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.128	.066	4

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Aff_1	13.6000	7.287	.229	.414	-.259[a]
Aff_2	13.9714	8.289	.181	.445	-.132[a]
Aff_3CR	12.8286	13.130	-.086	.275	.271
Aff_4CR	12.7000	13.662	-.098	.307	.265

a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

Table 60. Convergent Validity of Affordance

After taking out Aff_3CR, we ran the analysis again and obtained the results of Table 61.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.271	.078	3

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
Aff_1	8.7286	4.201	.485	.404	-.911[a]
Aff_2	9.1000	5.280	.392	.445	-.494[a]
Aff_4CR	7.8286	14.405	-.304	.113	.776

a. The value is negative due to a negative average covariance among items. This violates reliability model assumptions. You may want to check item codings.

Table 61. Convergent Validity of Affordance after taking Aff_3CR out

However, Affordances remained unreliable (Cronbach's Alpha = .27). We could further improve the reliability by taking out Aff_4CR (Table 61). However, because there were only two items left on the scale, we calculated Spearman-Brown Correlation and got the result of Table 62.

Correlations

			Aff_1	Aff_2
Spearman's rho	Aff_1	Correlation Coefficient	1.000	.610**
		Sig. (2-tailed)	.	.000
		N	71	70
	Aff_2	Correlation Coefficient	.610**	1.000
		Sig. (2-tailed)	.000	.
		N	70	70

**. Correlation is significant at the 0.01 level (2-tailed).

Table 62. Spearman-Brown Correlation between Aff_1 and Aff_2

Aff_1 and Aff_2 were significantly correlated (r = .61, p = .000).

11.2.7. Use Intentions scale

We ran reliability analysis on Use Intentions (3 items) and found Cronbach's Alpha = .63. Although reliability was not too high, we still kept the scale intact to enter the analysis of divergence.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.629	.615	3

Table 63. Convergent Validity of Use Intentions

In the analyses of 5.2.1 up to 5.2.7, we removed the variables of Eng_1-5, Bon_1-5, Rel_1-4, Anth_2, Anth_3, Real_1, Real_2, Real_4, Aff_1, Aff_2, Use_Int_1-3. The remaining items were tested for divergence.

11.3. Divergent Validity: PCA

As the indicators of each factor were chosen beforehand, we executed a *Principal Component Analysis* or PCA, forcing the items into a 7-factor solution. Because we expected that all factors would correlate, we used the *Direct Oblimin* method. *Maximum Iterations for convergence* was set to 25, coefficients under .30 were disregarded, and the number of factors were based on the *Kaisers Criterium* (>1). Results are in Table 64.

Component Matrix[a]

	Component						
	1	2	3	4	5	6	7
Eng_1	.677	.327				-.306	
Eng_2	.739	.300					
Eng_3CR	.329	.665	.366			.354	
Eng_4	.773						
Eng_5CR	.337	.722			.355		
Bon_1	.679		-.332		.410		
Bon_2	.757						
Bon_3	.772				.318		
Bon_4	.803						
Bon_5	.658			.437			
Anth_2	.701	-.305					
Anth_3	.768						
Real_1	.699					.377	-.320
Real_2	.693						
Real_4	.720						.313
Rel_1	.732				-.308		
Rel_2	.715		.337				
Rel_3CR	.308		.687				

Rel_4	.501	-.393	.389				
Aff_1	.691				-.309		
Aff_2	.703			-.302			
Use_Int_1	.694				.305		
Use_int_2CR			.689				
Use_Int_3	.765			-.352		-.316	

Extraction Method: Principal Component Analysis.

a. 7 components extracted.

Total Variance Explained

Component	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	10.560	44.001	44.001	10.389	43.288	43.288
2	1.970	8.210	52.211	2.024	8.434	51.723
3	1.916	7.982	60.193	1.864	7.767	59.489
4	1.276	5.317	65.510	1.338	5.574	65.063
5	1.009	4.203	69.713	1.116	4.650	69.713

Extraction Method: Principal Component Analysis.

Component Matrix[a]

	Component				
	1	2	3	4	5
Eng_1	.677				
Eng_2	.739				
Eng_3CR		.665			
Eng_4	.773				
Eng_5CR		.722			
Bon_1	.679				.410
Bon_2	.757				
Bon_3	.772				
Bon_4	.803				
Bon_5	.658			.437	
Anth_2	.701				
Anth_3	.768				
Real_1	.699				
Real_2	.693				
Real_4	.720				
Rel_1	.732				
Rel_2	.715				
Rel_3CR			.687		
Rel_4	.501				
Aff_1	.691				
Aff_2	.703				
Use_Int_1	.694				
Use_int_2CR			.689		
Use_Int_3	.765				

Table 64. Divergent validity of the experiential items in a 7-factor and 5-factor solution

The 7-factor solution showed that all items loaded on factor 1 with the two counter-indicative items of Engagement loading on factor 2 but with a lot of 'smear' to other components. Two other items formed a third component but they came from Relevance and Use Intentions. Because theoretically (Konijn & Hoorn, 2017), Bonding is the result of Anthropomorphism, Relevance, Realism, and Affordances and because Engagement and Use intentions were mere 'back-ups,' we reasoned that the bulk of the items sided with Bonding as the central component.

To give it another try, we reasoned that in a forced 5-factor solution, the two support scales Engagement and Use intentions should fall in line with Bonding, whereas the other theoretical variables should form their own component. The Total Variance Explained shows the actual five factors that were extracted while the Rotated Component Matrix shows the factor loadings of each variable. Almost all experiential items loaded on factor 1, the only scale remaining intact being Bonding (5 items). Again, factor 2 consisted of two Engagement items. Although the Spearman-Brown Correlation between Eng_3CR and Eng_5CR was significant, it was not very high (.51) (Table 65). And again, factor 3 was a combination of two items from different scales (Rel_3CR and Use_int_2CR).

Correlations

			Eng_3CR	Eng_5CR
Spearman's rho	Eng_3CR	Correlation Coefficient	1.000	.513**
		Sig. (2-tailed)	.	.000
		N	70	69
	Eng_5CR	Correlation Coefficient	.513**	1.000
		Sig. (2-tailed)	.000	.
		N	69	70

**. Correlation is significant at the 0.01 level (2-tailed).

Table 65. Spearman-Brown Correlation of Eng_3CR and Eng_5CR

All in all, divergent validity of the questionnaire items was weak and the only scale having good measurement quality overall was Bonding (5 items, Cronbach's α = .88), which will be the experiential factor we use for further analysis.

11.4. GLM Repeated Measures of Robot Design × Advancement with Mbond as covar

H3 expected that emotional bonding with the robot would positively affect the learning outcomes in a mediating or moderating way. We computed Mbond by calculating the average over Bon_1 to Bon_5. 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 with Mbond as the covariate. However, Mbond exerted no significant main or interaction effects on the multiplication scores and the earlier pattern of results was not altered.

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
BeforeAfter * Mbond	Pillai's Trace	,000	,012[b]	1,000	57,000	,912	,000
	Wilks' Lambda	1,000	,012[b]	1,000	57,000	,912	,000
	Hotelling's Trace	,000	,012[b]	1,000	57,000	,912	,000
	Roy's Largest Root	,000	,012[b]	1,000	57,000	,912	,000

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	BeforeAfter	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
BeforeAfter	Linear	607,079	1	607,079	14,698	,000	,205
BeforeAfter * Mbond	Linear	,504	1	,504	,012	,912	,000
BeforeAfter * Robot	Linear	124,672	2	62,336	1,509	,230	,050
BeforeAfter * Advancement	Linear	289,499	3	96,500	2,336	,083	,109
BeforeAfter * Robot * Advancement	Linear	990,164	6	165,027	3,995	,002	,296
Error(BeforeAfter)	Linear	2354,353	57	41,304			

Tests of Between-Subjects Effects

Measure: MEASURE_1

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	49974,367	1	49974,367	642,835	,000	,919
Mbond	15,595	1	15,595	,201	,656	,004
Robot	342,804	2	171,402	2,205	,120	,072
Advancement	24738,026	3	8246,009	106,071	,000	,848
Robot * Advancement	882,142	6	147,024	1,891	,098	,166
Error	4431,214	57	77,741			

We then ran a Univariate Analysis of Variance (ANOVA) of Robot Design and Advancement Level directly on Mbond. 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 ($F_{(1,58)} = 194.76$, $p = .000$, $\eta_p^2 = .77$). However, none of the main effects or interaction was significant ($F < 1$). Robot Design nor Advancement Level exerted significant effects on Mbond.

Tests of Between-Subjects Effects

Dependent Variable: Mbond

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	13,425[a]	11	1,220	,407	,947	,072
Intercept	583,657	1	583,657	194,755	,000	,771
Robot	1,045	2	,522	,174	,840	,006
Advancement	7,952	3	2,651	,884	,455	,044
Robot * Advancement	4,052	6	,675	,225	,967	,023
Error	173,819	58	2,997			
Total	914,320	70				
Corrected Total	187,243	69				

a. R Squared = .072 (Adjusted R Squared = -.104)

11.5. ANOVA School (2) × Robot Design (3) × Gender (2) × Advancement (4) on Mbond*

We conducted an ANOVA of School (2) × Robot (3) × Gender (2) × Advancement level (4) on Mbond, showing that only the difference in School was significant ($F_{(1,34)} = 4.57$, $p = .04$) (Table 66).

Tests of Between-Subjects Effects

Dependent Variable: Mbond

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	88.792[a]	35	2.537	.876	.651
Intercept	353.792	1	353.792	122.181	.000
Robot	1.231	2	.616	.213	.810
School	13.243	1	13.243	4.574	.040
Gender	.000	1	.000	.000	.995
Advancement	3.177	3	1.059	.366	.778
Robot * School	.075	2	.037	.013	.987
Robot * Gender	.134	2	.067	.023	.977
Robot * Advancement	8.293	6	1.382	.477	.820
School * Gender	2.183	1	2.183	.754	.391
School * Advancement	21.699	3	7.233	2.498	.076
Gender * Advancement	7.086	3	2.362	.816	.494
Robot * School * Gender	.984	1	.984	.340	.564
Robot * School * Advancement	3.653	4	.913	.315	.866
Robot * Gender * Advancement	8.025	3	2.675	.924	.440
School * Gender * Advancement	.055	1	.055	.019	.892
Robot * School * Gender * Advancement	.000	0	.	.	.
Error	98.451	34	2.896		
Total	914.320	70			
Corrected Total	187.243	69			

Table 66. School (2) × Robot (3) × Gender (2) × Advancement (4) ANOVA on Mbond

11.6. Independent Samples T-test of School on Mbond *

We ran an independent samples t-test of School on Mbond and got the results of Table 67.

Group Statistics

	School	N	Mean	Std. Deviation	Std. Error Mean
Mbond	Good Shepherd	48	3.6000	1.63941	.23663
	Chun Lei	22	2.4000	1.36626	.29129

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Mbond	Equal variances assumed	1.847	.179	2.987	68	.004	1.20000	.40169	.39845	2.00155
	Equal variances not assumed			3.198	48.437	.002	1.20000	.37529	.44561	1.95439

Table 67. Independent Samples t-test on Mbond with School as variable

The difference between Schools was significant ($t_{(68)} = 2.99$, $p = .004$), Good Shepherd showing higher mean Bonding than Chun Lei. The main difference between the schools was that Good Shepherd partook but once in the tutoring sessions and Chun Lei more than once (factor Partake).

11.7. Independent Samples T-test of Partake on Mbond *

We ran three t-tests with Partake as the grouping variable (once – twice, once – thrice, twice - thrice) and obtained the results tabulated in Table 68 up to Table 70.

Group Statistics

	Number of times Ss participated	N	Mean	Std. Deviation	Std. Error Mean
Mbond	1.00	48	3.6000	1.63941	.23663
	2.00	14	2.1857	1.14882	.30703

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Mbond	Equal variances assumed	3.045	.086	3.011	60	.004	1.41429	.46971	.47473	2.35384
	Equal variances not assumed			3.648	30.093	.001	1.41429	.38764	.62273	2.20584

Table 68. Independent samples t-test with Once and Twice on Mbond

Group Statistics

	Number of times Ss participated	N	Mean	Std. Deviation	Std. Error Mean
Mbond	1.00	48	3.6000	1.63941	.23663
	3.00	8	2.7750	1.70189	.60171

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Mbond	Equal variances assumed	.002	.965	1.311	54	.195	.82500	.62920	-.43648	2.08648
	Equal variances not assumed			1.276	9.299	.233	.82500	.64657	-.63048	2.28048

Table 69. Independent samples t-test with Once and Thrice on Mbond

Group Statistics

	Number of times Ss participated	N	Mean	Std. Deviation	Std. Error Mean
Mbond	2.00	14	2.1857	1.14882	.30703
	3.00	8	2.7750	1.70189	.60171

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Mbond	Equal variances assumed	1.793	.196	-.972	20	.343	-.58929	.60633	-1.85407	.67550
	Equal variances not assumed			-.872	10.728	.402	-.58929	.67552	-2.08070	.90213

Table 70. Independent samples t-test with Twice and Thrice on Mbond

The effects on Mbond of Once and Thrice and that of Twice and Thrice are not significant (Once – Thrice: $t_{(54)} = 1.31$, $p = .20$; Twice – Thrice: $t_{(20)} = .97$, $p = .34$). However, the difference between Once and Twice is significant for Mbond (Once – Twice: $t_{(60)} = 3.01$, $p = .004$), even if α is corrected to .017, using Bonferroni. Mean Bonding became less after first encounter ($M_1 = 3.60$, $SD = 1.64$; $M_2 = 2.19$; $SD = 1.70$), which is due to Chun Lei pupils alone. The insignificant different with those encountering the robot thrice might come from a lack of statistical power ($n = 11$).

We wondered if the high bonding upon first encounter was due to a novelty effect, wearing off after multiple encounters. Therefore, we did correlation analysis on Novelty and Mbond (Table 71).

Correlations

		Mbond	Nov_1
Mbond	Pearson Correlation	1	.309**
	Sig. (2-tailed)		.010
	N	70	69
Nov_1	Pearson Correlation	.309**	1
	Sig. (2-tailed)	.010	
	N	69	70

**. Correlation is significant at the 0.01 level (2-tailed).

Table 71. Correlation between Novelty and Mbond

We found that the correlation was significant but not very high (r = .31, p = .01). Children from Chun Lei saw the robot more often so that less novelty may have led to lower bonding.

11.8. Correlation between Mbond and Fin_min_Base and with Per_Fin_min_Base

We ran a two-tailed bivariate correlation analysis between Mbond and Fin_min_Base and between Mbond and Per_Fin_min_Base.

Correlations

		Fin_min_Base	Mbond
Fin_min_Base	Pearson Correlation	1	.007
	Sig. (2-tailed)		.951
	N	72	70
Mbond	Pearson Correlation	.007	1
	Sig. (2-tailed)	.951	
	N	70	70

Table 72. Correlation between Mbond and Fin_min_Base

Correlations

		Per_Fin_min_Base	Mbond
Per_Fin_min_Base	Pearson Correlation	1	-.076
	Sig. (2-tailed)		.531
	N	72	70
Mbond	Pearson Correlation	-.076	1
	Sig. (2-tailed)	.531	
	N	70	70

Table 73. Correlation between Mbond and Per_Fin_min_Base

Table 72 and Table 73 show that neither the correlation between Mbond and Fin_min_Base (r = .007, p = .951) nor that between Mbond and Fin_min_Base (r = -.076, p = .531) were significant, implying Mbond had no relation to learning gain in whatever form.

11.9. Conclusion

From the analyses of Section 5, we found:
1) Only the Bonding scale was psychometrically reliable,
2) Bonding had no significant relation with learning gain.
3) The Good Shepherd children experienced more bonding probably in view of a novelty effect

12. Overview of the factors on Fin_min_Base and Mbond

12.1. MANOVA of School (2) × Robot (3) × Gender (2) × Advancement (4) on Fin_min_Base, Per_Fin_min_Base, and Mbond with Age, Novelty, and Aesthetics as covariates

We conducted a GLM Multivariate Analysis (MANOVA) of School (2) × Robot Design (3) × Gender (2) × Advancement level (4) on Fin_min_Base and Mbond and on Per_Fin_min_Base and Mbond with Age, Nov_1, Aest_1 as covariates.

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Fin_min_Base	4758.161[a]	38	125.215	1.510	.124
	Per_Fin_min_Base	9.538[b]	38	.251	3.793	.000
	Mbond	126.024[c]	38	3.316	1.651	.080
Intercept	Fin_min_Base	98.695	1	98.695	1.190	.284
	Per_Fin_min_Base	.077	1	.077	1.160	.290
	Mbond	1.238	1	1.238	.617	.438
Nov_1	Fin_min_Base	133.125	1	133.125	1.605	.215
	Per_Fin_min_Base	.055	1	.055	.825	.371
	Mbond	2.363	1	2.363	1.177	.287
Aest_1	Fin_min_Base	57.003	1	57.003	.687	.414
	Per_Fin_min_Base	.026	1	.026	.391	.536
	Mbond	26.520	1	26.520	13.205	.001
Age	Fin_min_Base	38.572	1	38.572	.465	.500
	Per_Fin_min_Base	.017	1	.017	.250	.621
	Mbond	.227	1	.227	.113	.739
School	Fin_min_Base	27.615	1	27.615	.333	.568
	Per_Fin_min_Base	.017	1	.017	.263	.612
	Mbond	4.896	1	4.896	2.438	.129
Robot	Fin_min_Base	304.470	2	152.235	1.836	.177
	Per_Fin_min_Base	.802	2	.401	6.056	.006
	Mbond	.774	2	.387	.193	.826

Advancement	Fin_min_Base	69.835	3	23.278	.281	.839
	Per_Fin_min_Base	.817	3	.272	4.115	.015
	Mbond	1.903	3	.634	.316	.814
Gender	Fin_min_Base	15.064	1	15.064	.182	.673
	Per_Fin_min_Base	.004	1	.004	.063	.804
	Mbond	.837	1	.837	.417	.524
School * Robot	Fin_min_Base	188.014	2	94.007	1.133	.335
	Per_Fin_min_Base	.009	2	.005	.071	.932
	Mbond	1.529	2	.765	.381	.687
School * Advancement	Fin_min_Base	295.762	3	98.587	1.189	.331
	Per_Fin_min_Base	.269	3	.090	1.354	.276
	Mbond	1.668	3	.556	.277	.842
School * Gender	Fin_min_Base	3.727	1	3.727	.045	.834
	Per_Fin_min_Base	.007	1	.007	.112	.740
	Mbond	.269	1	.269	.134	.717
Robot * Advancement	Fin_min_Base	1161.560	6	193.593	2.334	.057
	Per_Fin_min_Base	1.646	6	.274	4.145	.004
	Mbond	14.118	6	2.353	1.172	.347
Robot * Gender	Fin_min_Base	.446	2	.223	.003	.997
	Per_Fin_min_Base	.000	2	.000	.004	.996
	Mbond	1.533	2	.766	.382	.686
Advancement * Gender	Fin_min_Base	320.295	3	106.765	1.287	.297
	Per_Fin_min_Base	.345	3	.115	1.738	.180
	Mbond	5.721	3	1.907	.950	.429
School * Robot * Advancement	Fin_min_Base	375.706	4	93.926	1.132	.360
	Per_Fin_min_Base	.210	4	.052	.792	.539
	Mbond	1.204	4	.301	.150	.962
School * Robot * Gender	Fin_min_Base	534.201	1	534.201	6.441	.017
	Per_Fin_min_Base	.633	1	.633	9.558	.004
	Mbond	.273	1	.273	.136	.715
School * Advancement * Gender	Fin_min_Base	199.238	1	199.238	2.402	.132
	Per_Fin_min_Base	.242	1	.242	3.660	.065
	Mbond	.211	1	.211	.105	.748
Robot * Advancement * Gender	Fin_min_Base	335.869	3	111.956	1.350	.277
	Per_Fin_min_Base	.285	3	.095	1.436	.252
	Mbond	9.985	3	3.328	1.657	.197
School * Robot * Advancement * Gender	Fin_min_Base	.000	0	.	.	.
	Per_Fin_min_Base	.000	0	.	.	.
	Mbond	.000	0	.	.	.
Error	Fin_min_Base	2488.129	30	82.938		
	Per_Fin_min_Base	1.985	30	.066		
	Mbond	60.251	30	2.008		
Total	Fin_min_Base	12710.000	69			
	Per_Fin_min_Base	17.982	69			
	Mbond	896.680	69			
Corrected Total	Fin_min_Base	7246.290	68			
	Per_Fin_min_Base	11.523	68			
	Mbond	186.275	68			

a. R Squared = .657 (Adjusted R Squared = .222)

b. R Squared = .828 (Adjusted R Squared = .609)

c. R Squared = .677 (Adjusted R Squared = .267)

Table 74. School (2) × Robot Design (3) × Gender (2) × Advancement (4) MANOVA on Fin_min_Base and Per_Fin_min_Base together with Mbond with Age, Nov_1, Aest_1 as covariates

Table 74 shows that the following effects were significant:

(1) the interaction of School × Robot Design × Gender on Fin_min_Base ($F_{(1,30)} = 6.44$, $p = .017$)

> However, Section 2.2 and 2.3 showed that none of the contrasts in the factors School, Robot, and Gender were significant so that (1) can be considered a false positive.

(2) the interaction of School × Robot Design × Gender on Per_Fin_min_Base ($F_{(1,30)} = 9.56$, $p = .004$)

> To scrutinize the contrasts of the factor Robot, we ran three independent samples t-tests of Robot on Per_Fin_min_Base. Table 76 – 78 show that none of the differences were significant (Humanoid – Puppy: $t_{(43)} = .14$, $p = .89$; Humanoid – Droid: $t_{(44)} = 1.03$, $p = .31$; Puppy – Droid: $t_{(51)} = 1.18$, $p = .24$). Table 79 and 80 show that neither the difference between School ($t_{(70)} = -1.23$, $p = .22$) nor that between Gender ($t_{(70)} = .13$, $p = .90$) was significant. We conclude that the significant F-value for the interaction came from the accumulation of noise in the contrasts.

(3) the interaction of Advancement × Robot Design on Per_Fin_min_Base ($F_{(6,30)} = 4.15$, $p = .004$) as produced by
(4) the main effect of Robot Design on Per_Fin_min_Base ($F_{(2,30)} = 6.06$, $p = .006$)

> As said in (2), the contrasts of the factor Robot were not significant. 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) and the main effect of Advancement on Per_Fin_min_Base ($F_{(3,30)} = 4.12$, $p = .015$)

> As shown in Section 3.8, a significant positive correlation occurred between Advancement and Per_Fin_min_Base and in Section 3.9, we saw a significant effect of Advancement on Per_Fin_min_Base, indicating that Per_Fin_min_Base decreased with the increase of Advancement. The t-test in Section 10, however, revealed that the significance effect is due to the level of being Challenged compared to the other three levels. With higher Advancement, pupils beyond being Challenged statistically did not obtain more learning gains.

(6) Aest_1 covaried with Mbond ($F_{(1,71)} = 13.21$, $p = .001$), indicating that the experience of 'prettier' led to stronger bonding tendencies as supported by a two-tailed bivariate correlation analysis ($r = .56$, $p = .000$) (Table 75).

Correlations

		Mbond	Aest_1
Mbond	Pearson Correlation	1	.561**
	Sig. (2-tailed)		.000
	N	70	69
Aest_1	Pearson Correlation	.561**	1
	Sig. (2-tailed)	.000	
	N	69	71

**. Correlation is significant at the 0.01 level (2-tailed).

Table 75. Correlation between Aest_1 and Mbond

Group Statistics

	Robot	N	Mean	Std. Deviation	Std. Error Mean
Per_Fin_min_Base	Humanoid	19	.3504	.56347	.12927
	Puppy	26	.3306	.39127	.07673

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Per_Fin_min_Base	Equal variances assumed	.782	.381	.139	43	.890	.01982	.14218	-.26691	.30655
	Equal variances not assumed			.132	30.219	.896	.01982	.15033	-.28710	.32674

Table 76. Independent samples t-test of Humanoid and Puppy on Per_Fin_min_Base

Group Statistics

	Robot	N	Mean	Std. Deviation	Std. Error Mean
Per_Fin_min_Base	Humanoid	19	.3504	.56347	.12927
	Droid	27	.2179	.30065	.05786

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Per_Fin_min_Base	Equal variances assumed	3.330	.075	1.034	44	.307	.13253	.12820	-.12585	.39091
	Equal variances not assumed			.936	25.234	.358	.13253	.14163	-.15902	.42408

Table 77. Independent samples t-test of Humanoid and Droid on Per_Fin_min_Base

Group Statistics

	Robot	N	Mean	Std. Deviation	Std. Error Mean
Per_Fin_min_Base	Puppy	26	.3306	.39127	.07673
	Droid	27	.2179	.30065	.05786

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Per_Fin_min_Base	Equal variances assumed	1.650	.205	1.179	51	.244	.11271	.09563	-.07927	.30470
	Equal variances not assumed			1.173	46.925	.247	.11271	.09610	-.08063	.30606

Table 78. Independent samples t-test of Puppy and Droid on Per_Fin_min_Base

Group Statistics

	School	N	Mean	Std. Deviation	Std. Error Mean
Per_Fin_min_Base	Good Shepherd	48	.2512	.31030	.04479
	Chun Lei	24	.3781	.56574	.11548

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Per_Fin_min_Base	Equal variances assumed	8.772	.004	-1.231	70	.222	-.12686	.10302	-.33233	.07861
	Equal variances not assumed			-1.024	30.106	.314	-.12686	.12386	-.37978	.12606

Table 79. Independent samples t-test of School on Per_Fin_min_Base

Group Statistics

	Gender	N	Mean	Std. Deviation	Std. Error Mean
Per_Fin_min_Base	Male	42	.2988	.39024	.06022
	Female	30	.2862	.45096	.08233

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Per_Fin_min_Base	Equal variances assumed	.022	.883	.126	70	.900	.01256	.09956	-.18600	.21112
	Equal variances not assumed			.123	56.822	.902	.01256	.10200	-.19172	.21683

Table 80. Independent samples t-test of Gender on Per_Fin_min_Base

13. Questionnaire

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

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

Konijn, E. A., & Hoorn, J. F. (2017). Parasocial Interaction and Beyond: Media Personae and Affective Bonding. *The International Encyclopedia of Media Effects*, 1-25.

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.
Available from http://link.springer.com/article/10.1007/s12369-015-0301-3

Pisapia, J., Schlesinger, J., & Parks, A. (1993). Learning Technologies in the Classroom: Review of Literature.