Table B.1
Quality Assessment of Studies

Last Name of First Author	Selection Bias Randomization present: was method disclosed? (Yes = Y, no = N)	Allocation Concealment	Performance Bias Blinding of Participants and Personnel	Detection Bias Blinding of Outcome Assessment	Attrition Bias Incomplete Outcome Data	Reporting Bias Selective Reporting	Other Bias
1. Alemi et al. (2014)	N	-	-	-	+	+	No fidelity checks reported. Self –selection of school and year. Female students only. Small sample size. No ethics.
2. Alemi et al. (2015)	N	-	-	-	+	+	No fidelity checks reported. Self –selection of school. Female students only. Small sample size. No ethics.
3. Baxter et al. (2017)	N	-	+	-	+	+	Self-selected school and year. No control for children's

							academic ability. Gender imbalance. Different teachers used.
4. Chang and Chen (2010)	N	-	-	-	-	+	No fidelity checks reported. Self –selection of school and classes. Heterogeneity among English teachers. No control of English ability. No ethics.
5. Chang et al. (2010)	N	-	-	-	+	-	No fidelity checks reported. Self –selection of school and classes. Heterogeneity among English teachers. No control of English ability. No ethics.
6. Chen et al. (2011)	N	-	-	-	-	-	No fidelity checks reported. Self –selection of school and students. Self-report measures.

							No comparator. No baseline. Small sample size. No ethics.
7. Fernandez-Llama et al. (2018)	N	-	-	-	+	-	No fidelity checks reported. No report of sampling method. Wide age range. Cognitive development variation. No ethics.
8. Fridin (2014a)	N	-	-	-	?	+	No fidelity checks reported. No report of sampling method. Subjective interpretation of video footage. Small sample size. No ethics.
9. Fridin (2014b)	N	-	-	-	?	+	No fidelity checks reported. No report of sampling method. Subjective interpretation of video footage.

10. Hashimoto et al. (2011)	N	-	-	-	?	+	Small sample size. No fidelity checks reported. No report of sampling method. Validity of outcome measures unknown. Small sample size. No ethics.
11. Hong et al. (2016)	N	-	-	-	-	-	No fidelity checks reported. No baseline. Self –selection of school and classes. Small sample size. Lack of control over student language ability. No ethics.
12. Hsiao et al. (2015)	N	-	-	-	?	+	No fidelity checks reported. Selection of kindergartens specific to districts. Non-equivalent control group. Small

							sample size. No ethics.
13. Jones and Castellano (2018)	N	-	-	-	?	+	No fidelity checks reported. Self –selection of school and participants. Small sample size.
14. Kanda et al. (2004)	N	-	-	-	+	+	No fidelity checks reported. Self –selection of school and participants. Small sample size. No ethics.
15. Keren and Fridin (2014)	N	-	-	-	+	+	No fidelity checks reported. Self –selection of school and participants. Small sample size. No ethics.
16. Kory-Westlund, Dickens et al. (2017)	N	-	-	-	-	+	No fidelity checks reported. Self –selection of school and participants. No control over children's learning ability.

17. Kory-Westlund, Jeong et al. (2017)	N	-	-	-	-	-	Small sample size. No ethics. No fidelity checks reported. Self –selection of school and participants. Unequal language learning status & ages. No control of children's learning ability. Experimenter effect. Incomplete and underpowered data.
18. Mazzoni and Benvenuti (2015)	N	-	-	-	+	+	Small sample size. No ethics. No fidelity checks reported. Self –selection of school and participants. Potential experimenter effect. Small sample size. No ethics.

19. Serholt (2018)	N	-	-	-	?	+	No fidelity checks reported. Self –selection of school and participants. Teacher subjectivity. Small sample size.
20. Shiomi et al. (2015)	N	-	-	-	+	+	No fidelity checks reported. Self –selection of school and participants. Subjectivity of video and audio analysis. Reliability of perception measure unknown. No ethics.
21. Wei et al. (2011)	N	-	-	-	?	+	No fidelity checks reported. Self –selection of school and participants. Subjectivity of instructors. Small sample size. No ethics.

Note. + = low risk of bias; - = high risk of bias; ? = unsure risk of bias.

Table B.2
Data Extraction Table

Authors Year	Aim	Participants and Sampling	Theoretical/Conceptual Framework (TCF), Methodology and Data Collection	Intervention	Main Findings
1. Alemi et al. 2014 (Iran)	Study the effects of Robotics Assisted Language Learning (RALL) on vocabulary learning and retention.	46 private junior high children (mean age 12 years). 46 female. Non-probability convenience sampling.	TCF: social and interactive learning. Quantitative pre-post quasi-experimental design. Assigned to either control group or RALL group. 45 item vocabulary test.	NAO humanoid robot. The robot played various games and tests with students.	RALL system was more successful in learning and retention of vocabulary in the domain of English.
2. Alemi et al. 2015 (Iran)	Investigate the effect of robots in class on anxiety levels of students during the language learning process.	46 private junior high children (mean age 12.5 years). 46 female. Non-probability convenience sampling.	TCF: N/A. Mixed experimental design. Post measures only. Random allocation to one of three classes; 2 classes of experimental condition and 1 class of control. Foreign Language Classroom Anxiety Scale (FLCAS). Attitude Scale.	NAO humanoid robot. Programmed for 6 different modes; playing games, calling on students to answer, singing songs, teaching, cheering the correct answer, and making mistakes on purpose.	RALL group students demonstrated significantly lower levels of anxiety than those in the non-RALL group. Students displayed positive attitudes towards the use of this educational system.
3. Baxter et al. 2017 (UK)	Examine the effect of personalised robots on child learning.	59 primary school children (mean age 7.5 years). 24 males and 35 females.	TCF: Vygotskian (collaborative, personalised, active learning), Revised Bloom's Taxonomy. Quantitative pre-post experimental design. Group randomisation at class level. Comparison between	NAO humanoid robot. Touchscreen (Sandtray). 2 category sorting tasks played with the robot on the touchscreen.	The personalised robot appeared superior to the non-personalised robot for the novel learning task but not the mathematical based learning task.

		Non-probability convenience sampling.	personalised and non-personalised robot group. Pre and post knowledge tests, questionnaires assessing perception of social presence of the robot and perceived social support provided by the robot.		Children's perception of robots largely favoured the personalised robot although not statistically significant.
4.Chang et al. 2010 (Taiwan)	Evaluate the effects of robots on the learning of language.	100 elementary school children (mean age 10.5 years). 3 teachers. Non-probability convenience sampling.	TCF: interactive and personalised learning, dual coding theory, natural approach and the affective filter hypothesis, communicative approach and task-based language teaching, total physical response approach. Quasi-experimental design. No control group. Post study interviews. Post-hoc analysis of video footage.	Robosapien. The humanoid robot interacted with children during 5 different scenarios; storytelling mode, oral reading mode, cheerleader mode, action-command mode and question and answer mode.	The combination of the children's reaction to the robot and the teachers' opinions suggested that the robot could create an engaging and interactive learning experience.
5. Chang and Chen 2010 (Taiwan)	Evaluate the interaction and usability of a classroom using a robot.	134 elementary school children (mean age 7.5 years). 7 English teachers. Non-probability convenience sampling.	TCF: interactive learning. Mixed quasi-experimental design. Traditional teaching compared to teaching with a robot. Within-subject post measures only. The usability for the teacher – subjects were interviewed and completed a questionnaire. Interaction of students – Questionnaire about subjects' satisfaction and interaction in the class. Classroom dynamic was video recorded.	RoboSapien V2. Robot performed a dialogue, encouraged positive behaviour and gave quizzes. The robot used affective behavioural responses.	Usability was rated unsuccessfully due to budget constraints and fear of damage to the robot. Positive impact of the robot on interest, motivation and concentration in learning English.

6. Chen et al. 2011 (Taiwan)	To assess whether the integration of robot, computer and book creates a joyful and novel English learning environment.	5 elementary school students (mean age 10.5 years). 3 male and 2 female. Non-probability convenience sampling.	TCF: Vygotskian interactive learning. Quasi-experimental design. No comparison group. Interviews conducted with students and teachers. Data collected from video recording for system evaluation (usability and feasibility).	Humanoid robot (unspecified). Integrated with computer and books. The robot interacted with children through conversation, dancing and other motions.	The proposed system promoted a positive learning experience and motivates students to learn English. Concerns expressed regarding the stability, interface design and cost of the system.
7.Fernandez-Llamas et al. 2018 (Spain)	Analyse students' attitudes towards robots between the two conditions of human and robot teacher.	190 primary and secondary school children (aged 6 to 16 years). Non-probability convenience sampling.	TCF: problem-based, project-based, challenge-based learning approaches. Quantitative quasi-experimental design. Randomisation at group level. Post measures only. Comparison of both age groups and condition (robot teacher vs. human teacher). The Negative Attitudes Toward Robots Scale (NARS). The Robot Anxiety Scale (RAS).	Baxter. The robot gave a brief lecture and explained the exercises.	Human vs. Robot condition: Students in the human conditions showed greater concern over the robots capabilities than those in the robot condition. Age Differences Older children appeared more apprehensive or cautious around robots and technology compared to younger children.
8. Fridin 2014a (Israel)	Design a procedure to effectively introduce Kindegarten SAR (KindSAR) to	11 kindergarten children (mean age 3.3 years). 5 male and 6 female. Israeli-born.	TCF: social interactionism. Quantitative cross sectional design. No comparison group. Post measures only. Post-hoc analysis of video footage.	NAO humanoid robot. The robot engaged children in brief conversation. Explained the rules and played a game with the children	Interaction level was positive throughout all but one of the procedures.

	children in a natural setting.	Non-probability convenience sampling.		whilst providing positive reinforcement.	
9. Fridin 2014b (Israel)	Aim to see if KindSAR can engage preschool children in constructive learning.	10 kindergarten children (mean age 3 years). 5 male and 5 female. Non-probability convenience sampling.	TFC: social constructivist theory, stealth education. Quantitative cross-sectional design. No control group. Post-hoc analysis of video footage.	NAO humanoid robot. The robot interacted with the children and guided them by singing, playing games, movements and conversation.	Preschool children enjoyed interacting with the KindSAR and accepted the authority of the robot as a teaching assistant.
10. Hashimoto et al. 2011 (Japan)	Investigate the effectiveness of the proposed educational system (where robot is used as a teacher) by conducting two field experiments	Field Experiment 1: 38 elementary school children (aged 6-12 years). 30 university students (aged in their 20s). Field Experiment 2: 22 elementary school children (mean age 10.5 years). Non-probability convenience sampling.	TFC: N/A. Quantitative quasi-experimental design. Field Experiment 1: Comparison between elementary and university students. Questionnaire on age-dependent differences in students' interest, motivation, and concentration. Field Experiment 2: No comparison group or baseline measurements. Questionnaire on students' interest and motivation.	SAYA - android robot Field Experiment 1: SAYAs greeted and talked to the students giving advice, cautions and asking questions through the use of the operator. Field Experiment 2: SAYA conducted the science class through the use of the operator and interacted with students similar to field experiment 1.	Elementary school students were more accepting of the robot education system than the university students. They also displayed a greater a level of participation and concentration. The use of SAYA enhanced the elementary school students' motivation towards science class.
11. Hong et al. 2016 (Taiwan)	Assess the impact of robot-assisted English learning	52 elementary school children	TCF: N/A (motivation). Mixed quasi-experimental design. Control group. Post measures	Bioloid The humanoid robot assisted the learning	Students in the experimental group (with robot)

	on the learning performance and motivation of students.	(mean age 10.5 years). Non-probability convenience sampling.	only. School examination to assess four types of language ability; listening, speaking, reading and writing. Motivation measured using IMMS (29 items assessing attention, relevance, confidence and satisfaction). Interview with teacher.	during four types of classroom activities; storytelling; reading aloud; listening and acting and questioning and answering.	demonstrated greater increase in confidence, attention, motivation, satisfaction with teaching materials and learning process and language ability.
12. Hsiao et al. 2015 (Taiwan)	Examine the effect of a robot learning companion (RLC) or tablet-PC on children's reading motivation, literacy and behaviour (Mandarin reading literacy).	57 kindergarten children (mean age 2.5 years). Non-probability convenience sampling.	TCF: N/A. Quantitative quasi-experimental design. Comparison group used. Pre and post tests to measure reading literacy (reading comprehension, storytelling ability, word recognition and retelling of stories). Behaviour indicators of on-task behaviour observed and attitude toward learning with RLC measured with questionnaire.	iRobiQ. An intelligent robot which used three main functions; broadcasting sound; the ability to express human-like emotion and the touch screen.	The robot improved children's reading ability and enhance their interest in Mandarin reading.
13. Jones and Castellano 2018 (UK)	Examine if using long-term adaptive self-regulated learning (SRL) scaffolding improves learners' SRL skills and learning gain.	24 primary school children (mean age 11 years). 14 male and 10 female. Non-probability convenience sampling.	TCF: Self-regulating learning and scaffolding. Quantitative experimental design. Comparison group used. Participants randomly split into SRL or control group. SRL questionnaires (pre and post activity domain), Self-assessment accuracy (pre and post ranking of skills. SRL indicators in task performance data.	NAO humanoid robot torso. Robot introduced the geography based learning task, provided domain tutoring and performed motions throughout the session. In the SRL condition the robot	A more personalised and adaptive scaffolding of SRL processes using open learner model (OLM) improved SRL processes more than domain tutoring alone. Learning gain in both conditions, although no significant differences.

				offered SRL scaffolding	
14. Kanda et al. 2004 (Japan)	Examine whether interacting with robots could improve children's ability to speak English.	119 students (mean age 6.5 years) and 109 students (mean age 11.5 years). 112 male and 116 female. Non-probability convenience sampling.	TCF: Clark's theory for communication and Hall's theory on proximity during communication. Quantitative pre and post quasi-experimental design. No comparison group. The time spent interacting with robot was measured by logging ID tags and calculating interaction time per day. Social interaction was measured by comparing log data to the names of each child's friends. Pre-mid-post English skill tests were administered.	Robovie. Interactive humanoid robot used a wireless ID tag system to recognize different children. Programmed for human-like expression. It made interactive behaviours such as shaking hands and idle behaviours such as scratching its head.	The robot encouraged some children to improve their English but this was more successful for children who had higher baseline of English knowledge.
15. Keren and Fridin 2014 (Israel)	Examined the hypothesis that children can learn to think geometrically through their interaction with a KindSAR.	17 kindergarten children (mean age 4.1 years). Israeli-born 10 male and 7 female. Non-probability convenience sampling.	TCF: Van Hiele Theory of geometric thinking levels. Quantitative cross-sectional design. No control group. Post-hoc analysis of video footage.	NAO humanoid robot. The robot interacts with the children and guides them through the different procedures by instructing them to complete tasks such as asking children to look at the screen and note the basic shape of its picture shown there.	A KindSAR assisted the teacher in promoting geometric thinking learning via an educational game-based activity.

16. Kory-Westlund, Dickens et al. 2017 (USA)	Examined whether having human or robot partner during word learning task effects learning and attention to social cues.	34 preschool children (mean age 3.69 years). 13 male and 21 female. Non-probability convenience sampling.	TCF: N/A. Quantitative quasi-experimental design. Pre and post-test measured child's perception of robot's similarity to an iPad or human. Score on animal recall test. Video and audio recordings of each session to assess gaze of child.	DragonBot named "Blue". Robot's animated face displayed on a screen. The tele-operator triggered speech, movements and facial expressions as a means of interaction with the children during the task.	No difference found in the recall test between human and robot conditions. Performance was inhibited by the spatial distinctiveness of nonverbal orientation cues available to determine which animal was being referred to during naming.
17. Kory-Westlund, Jeong et al. 2017 (USA)	Investigate whether children learn from a robot and whether the expressiveness of the robot's voice would impact students' learning.	45 preschool children (mean age 5.2 years). 23 male and 22 female. 18 children native English speakers, 17 English language learners, 8 bilingual and 3 unknown. Non-probability convenience sampling.	TCF: N/A. Mixed experimental design. Post measures only. Participants randomly assigned to expressive or flat tone group. Children's responses to the Peabody Picture Vocabulary Test. Recorded children's facial expressions using Affdex, emotion measurement software.	Tega robot. A teddy like interactive robot interacted with children through greetings, storytelling, question and answering and encouragement.	Children in the expressive robot condition had higher levels of engagement and concentration than the flat condition. Of the children who responded to the robots questions, those in the expressive condition were more likely to identify the target words correctly.
18. Mazzoni and	Compared the effectiveness of a humanoid robot to	10 kindergarten children (mean age 5 years old).	TCF: Vygotskian social constructivism. Quantitative quasi-experimental design.	MecWilly. A humanoid ecological robot. The	Humanoid robot was effective in improving

Benvenuti 2015 (Italy)	a human counterpart in helping Italian children learn English words.	4 male and 6 female. Non-probability convenience sampling.	Control group used (human condition including 6 participants). Experimental condition (robot) consisted of 4 children. Pre and post test scores on word-picture association task.	robot worked collaboratively with children and offered suggestions and comments for answers to the task.	children's knowledge of English words.
19. Serholt 2018 (Sweden)	Explore the causes of breakdowns in children's interactions with the robotic tutor.	46 children (mean age 10.8 years). A sample of 6 students was selected for in-depth analysis. Each partnered with another student from their class. 11 male and 1 female. Non-probability convenience sampling.	TCF: N/A. Quasi-experimental design. Video recordings of interaction sessions.	NAO T14 humanoid robot. Robot interacted with the children by greeting them, explaining the task, demonstrating enthusiasm, instructing and engaging in the task.	Potential causes for the breakdown of child-robot interaction included: (a) inability of robot to evoke initial engagement and identify misunderstandings; (b) confusing scaffolding; (c) lack of consistency and fairness; and (d) controller problems.
20. Shiomi et al. 2015 (Japan)	Investigate whether interaction with a social robot raises students' interest in science.	114 elementary school students (mean age 10.5 years). Non-probability convenience sampling.	TCF: no explicit theory, but inferences to inquiry-based learning, motivation and peer influence. Mixed pre-post quasi-experimental design. No comparison group. Science curiosity measure consisting of 5 items. Perception of robots post-test. Audio and video data recorded to identify	Robovie. Social robot was programmed with relational behaviour such as calling children by their names and behaviours that encourage science questions such as	Children who asked the robot more science questions displayed a greater increase in science curiosity. The cause and effect relationship is unclear but it is suggested that a curiosity loop encourages questions

			the number of times the robot was asked science questions and robot question.	chatting about science topics.	which further increase curiosity.
21. Wei et al. 2011 (Taiwan)	Examine whether Joyful Classroom Learning System (JCLS) can lead child learners to have better learning experiences.	47 elementary school children (mean age 7.5 years). 19 male and 28 female. Non-probability convenience sampling.	TCF: Experiential Learning and Constructivist Learning Theory. Mixed pre-post quasi-experimental design. Control group used. Questionnaire examining the three constructs of experiential learning, constructivist learning and joyful learning. Observational data. One-to-one interview of 6 participants.	LEGO Mindstorms NXT. The robot reacted to participants whilst they engaged in the activity by offering sounds and actions based on their inputs.	JCLS improved the learning experience of students and children perceived JCLS could increase their learning motivations and concentration. Results also found that children perceived the JCLS to be useful and easy to use.