International Journal of Artificial Intelligence in Education (2020) 30:38–73
https://doi.org/10.1007/s40593-020-00193-4

ARTICLE

Preschoolers' Understanding of a Teachable Agent-Based Game in Early Mathematics as Reflected in their Gaze Behaviors – an Experimental Study

Agneta Gulz[1,2] · Ludvig Londos[1] · Magnus Haake[1]

Published online: 7 February 2020

Abstract

This study investigated how preschool children processed and understood critical information in Magical Garden, a teachable agent-based play-&-learn game targeting early math. We analyzed 36 children's (ages 4–6 years) real-time behavior during game-use to explore whether children: (i) processed the information meant to support number sense development; (ii) showed an understanding of the teachable agent as an entity with agency. An important methodological goal was to go beyond observable behavior and shed some light on how cognitive processing and understanding in children of such young age can be studied. First, the children played Magical Garden for three weeks to get acquainted with the game. Second, in an experimental part of the study, the children's gaze behaviors were measured during 5 rounds of interaction with an experimental version of one of the sub-games. The analyses suggest that two of the gaze behaviors were positively correlated with the game performance measure, as hypothesized. Another result was that children looked at the teachable agent significantly more often when the teachable agent had been in charge of gameplay than when it had not. This can be interpreted as an indication that the children had an understanding of their teachable agent as an entity that, like themselves and unlike other dynamic visual elements in the game, made decisions based on own 'knowledge'. In a broader context, the findings are important in showing the potential gains of combining log data with eye-tracking data for developing and refining AI algorithms for adaptive individual feedback and scaffolding.

Keywords Teachable agent · Preschoolers · Eye-tracking · Early math · Number sense

✉ Agneta Gulz
agneta.gulz@lucs.lu.se

Extended author information available on the last page of the article

Introduction

In the present study, four- to six-year-old preschoolers used a digital play-&-learn game targeting early math. Our focus was to investigate how children of this young age processed and understood some of the crucial information in the game, particularly for developing their *number sense* (Griffin 2004; Griffin and Case 1997). A well-developed number sense, i.e., the conceptual grounding of numbers, includes understanding of the relevance of words such as more-less, higher-lower, larger-smaller and shorter-taller in the context of sets, as well as the linking number words to magnitude and a variety of visual number *representations*. Such representations range from the iconic representations (e.g., a hand with fingers held up) to the semi-symbolic (e.g., slashes as found in Roman numerals) and the fully symbolic representations (Arabic numbers). Without a well-developed number sense there is high risk for mechanical rote learning and memorizing without understanding what the numbers really stand for.

We used the research-based play-&-learn game Magical Garden, which was developed to support children in developing their number sense and has been employed in several studies (Gerholm et al. 2018; Gulz and Haake 2019; Haake et al. 2015a). Magical Garden uses a teachable agent (Blair et al. 2006). Implementing a teachable agent (hence TA), means that the child gets the role to instruct a digital character how to accomplish tasks while the character's actions will reflect the child's teaching. This entails two potential pedagogical gains. First, tasks are practiced many times without being all too repetitive, since the child first tries herself to accomplish a kind of tasks, then instructs the character how to accomplish such tasks, and thereafter supervises the character's trying to accomplish the same kind of tasks. Second, the child gets to see and potentially evaluate someone else – namely the character – performing the tasks. Whereas it is very hard for young children to simultaneously act themselves and think about their own actions due to cognitive load, reflecting upon someone else's actions is more feasible for them (Gelman and Meck 1983).

However, full benefit from a TA-based game requires a recognition that not all people know the exact same things, that one person may know things someone else does not know, and that it is possible for the first person to influence the second person's knowledge. Translated into the kinds of understanding required to meaningfully instruct a TA, one must ascribe some knowledge[1] to the TA that can differ from one's own knowledge, and that one can influence. In addition, for meaningful instruction to take place the 'teacher' must be able to evaluate whether the TA changes its knowledge in response to instruction. This means that one must understand that the TA's knowledge is reflected in its behavior and perceive the TA's behavior as goal-directed.

In other words, to fully profit from the game one needs to perceive that the TA has *agency*, meaning that the TA is able to act independently in a goal-directed manner based on its own knowledge. If one perceives the TA as an entity with agency, it becomes meaningful to (try to) understand the TA's behavior in the game and try to

[1] Note that knowledge as used here is not an epistemological-philosophical notion but can be equated with understanding and interpretation and can be explicit or implicit.

influence it by instruction (such as affirming and correcting choices, demonstrating adequate answers, etc.).[2]

One goal of our study was to investigate aspects of these young children's cognitive processes as they played Magical Garden. The other goal was to explore if analyses of children's gaze behaviors during gameplay could be used in this investigation. Specifically, we asked whether analyses of children's gaze behaviors could provide insights on whether children:

- processed the information in the game meant to support number sense development by connecting the different number representations to one another, rather than behaving on a trial-&-error-basis to solve the tasks in the game,
- perceived the TA as an entity with agency in the sense detailed above.

The study presented in this article is, to our knowledge, the first eye-tracking study with preschoolers and visual stimuli consisting of dynamical, visually rich scenes and animations partly unfolding in response to the preschoolers' own actions. As researchers, we use the term 'experimental study', whereas the participating preschoolers probably saw the activities only as game-playing.

All researchers who have worked with preschoolers know that it is very hard to get insights into their understanding by methods such as think-aloud or different form of interviews compared to how these methods can be used with adults and school children. Preschoolers are less verbal. Their vocabulary is limited and, in addition, some are not inclined to talk much at all with an adult even if they know the person. Also, they are less inclined than adults and school children to comply with an interviewer, for example, some talk extensively but not on the topics introduced by the researcher. This was a reason for our choice to explore whether eye-tracking measurements – which do not rely on subjects' verbalization – could shed some light on the cognitive phenomena and processes we were interested in, i.e., children's understanding of number sense as well as their understanding of agency.

To learn more about processes of understanding is central for cognitive science and learning science. It is also relevant for those interested in educational software, not the least for the implementation of Artificial Intelligence (AI). Behavioral data (at least for the time being) is a main input for AI-algorithms – but to really take advantage of the potential of AI in educational software, we need to better understand the cognitive processes behind the behavioral log data. As a *first* step strict behavioral comparisons between control-groups and experimental groups using pre- and post-tests are indeed useful for the evaluation of an educational game, but with more complex AI-based educational software it is essential to analyze what occurs in terms of different learners' processing and understanding.

Importantly, learning outcomes can be identical and yet the road taken there be different for different groups of learners. Misunderstandings and failures can be of different kinds and occur at different stages with the consequence that different groups of learners will be helped by different kinds of support and feedback, at different stages

[2] There are clouds that move on the sky in the game. Like the TA these are dynamical, visual elements on the screen, but their behavior is not directed by goals and knowledge. However, children without a sufficiently developed understanding of agency will not perceive the categorical difference between the TA and the clouds.

during a learning process. A central potential of artificial intelligence for education is the possibility to individually adapt software to learners at different levels of understanding. Potentially, eye-tracking measurements on learners who use an educational game may provide real-time data to suggest what is going on – or not – in terms of processing and understanding on the part of the learners. Such data can be useful both for adaptive educational systems and for user modeling.[3]

Both data-logging of game behavior and eye-tracking have obvious limitations given the goal of linking the data to cognitive phenomena, such as levels of understanding. A gaze fixation measure, as such, cannot differentiate between someone who stares blankly at an informative interface area versus someone who carefully processes the information in question. Log data measuring a correct response (as a series of clicks) cannot, as such, tell for sure if the response was a result of chance or of a series of well-informed choices. Intelligent combinations of both methods can, however, as showed already by Gluck et al. (2000), be used to disambiguate between interpretations, such as different strategies that in principle can lead to the same responses.

Below follows a brief introduction on the area of eye-tracking, followed by a presentation of some previous related studies that address mathematical understanding and understanding of teachable agents, in particular studies that make use of eye-tracking.

Background

Gaze Behaviors and Eye-Tracking

The development of eye-tracking has advanced the possibility to study cognitive processes as reflected by gaze behaviors. Important elements of eye-movements are fixations and saccades. A fixation is the act of maintaining one's gaze on a single location, whereas a saccade is the rapid eye-movement between fixations to move the eye-gaze from one point to another. Such elements of eye-movements or gaze behaviors can be quantitatively and objectively assessed (Holmqvist et al. 2011) and to some extent be coupled to cognitive behaviors like visual attention (Deubel and Schneider 1996; Corbetta et al. 1998; Holmqvist et al. 2011; Smith 2012). It should be kept in mind that the detection of specific eye-movements is never a *direct* assessment of a certain cognitive process – but can suggest the presence of a certain cognitive process. Importantly it is known that eye-movements are influenced both by visual input (bottom-up) and by top-down constraints based on previous knowledge (Henderson 2003). As a consequence, there are situations in which the likelihood that certain eye-movements will occur is strongly related to which interpretation and understanding a person has of the situation. This opens for experimental designs where the likelihood that someone will fixate certain areas at certain stages of a process will depend strongly on the persons' previous knowledge and contextual understanding of the visual input.

[3] Additional knowledge of how target groups of learners seem to understand or not understand what goes on in a game can also be used for refinement of the specific game – whether adaptive or not – and can also be used for the generation of more general knowledge on which to base novel educational software.

That is, the gaze-behavior data may allow inferences about participants' understanding of the situation.

Regarding understanding, *anticipation* is a both central and prevalent cognitive phenomenon. Anticipation is a future oriented action based on previous knowledge in order to make a prediction (Pezzulo et al. 2008). Looking ahead at a spot where some relevant information is *going to turn up* can, on the one hand, be triggered by bottom-up visual input, such as light, motion, etc. On the other hand, it can also be influenced by an understanding of the present context with its unfolding of information. Overall, eye-tracking is well suited to identify *visual look-aheads*. For young children the method is particularly interesting since also children who cannot or do not respond verbally may still respond by gaze. For example, one study that measured visual anticipation showed that 2-year-olds seem to have a budding capacity to predict which out of two objects an individual will reach for after having shown beforehand which of the objects the individual likes and which of them she dislikes (Vaish et al. 2018).

A number of eye-tracking studies targeting students' attention to pedagogical agents have been conducted over the past decade. For example, Louwerse et al. (2008) investigated the distribution of attention to different agents that took turn speaking. Conati et al. (2013) studied attention patterns to adaptive hints provided by a pedagogical agent in an educational game.

Number Sense and Representations of Numbers

To advance through the math education in school, a well-developed number sense (Griffin 2004; Griffin and Case 1997) is paramount. A well-developed number sense involves an understanding of the following: (1) numbers indicate quantity; (2) the relevance of the concepts 'larger'/'smaller', 'higher'/'lower', 'more'/'less' in the context of sets and numbers; (3) numbers occupy fixed positions in the counting sequence; (4) numbers that come later in the sequence correspond to a larger quantity; (5) each incremental number corresponds to an increase of one; (6) the meaning of different representations and cultural metaphors and how all those meanings are connected to a symbolical number. In essence, the construct 'number sense' refers to the conceptual prerequisites for successfully learning mathematics. To support the development of number sense it is, according to Griffin and Case (1997), important to present numbers and their magnitude using a variety of representations of numbers, such as groups of objects, dot patterns, positions on horizontal or vertical lines, and connecting them to concepts such as taller/shorter, higher/lower, larger/smaller, etc. Although some of these concepts may seem obvious and intuitive to an adult, they are not to a child; they must be learned (Griffin and Case 1997; Griffin et al. 1994). Longitudinal studies show that preschool children, who have not learned this, rarely catch up (Jordan et al. 2009; Hannula et al. 2007; Hannula et al. 2010; Griffin and Case 1997; Griffin et al. 1994).

To support children's number sense development it is, according to Griffin and Case (1997), central to thoroughly introduce crucial concepts such as higher/lower, upwards/downwards, more/less, larger/smaller, and to practice these concepts in relation to sets of different sizes. Along with such practice of fundamental (pre-)mathematical concepts a gradual introduction of representations of sets and numbers should take place. Here children should be step-wise introduced to different representations and their relations, starting with concrete (iconic) representations of objects (e.g., virtual balloons representing balloons), over gradually less iconic and more symbolic representations (e.g., fingers, stripes, and dots), to end with the entirely abstract symbols of Arabic numbers (i.e., 1, 2, 3, …).

But how can we know whether children provided with pedagogical materials designed to support their development of number sense, will indeed process the material as intended, namely in ways that help them to conceptualize and understand the mappings and relations between sets and symbolic numbers? Consider, for example, a mapping between: (i) three virtual balloons (prepared for flight), (ii) the third step or level on a vertical virtual layout, (iii) a treasure placed at step/level five, (iv) the need for two more virtual balloons (to reach two levels higher) and (v) the symbolic representation of these two additional balloons in the form of the Arabic number two – ‘2’. How can we, as designers of educational software, know whether a child is truly working out these mappings, thus building an understanding of number, rather than just acting on a ‘mechanical’ basis, using trial-&-error strategies together with a memorizing of successful combinations? How can we as researchers get closer to children’s actual cognitive processing and understanding? Specifically, can a study of children’s gaze behaviors during game use contribute?

A first observation is that whereas eye-tracking has been extensively used to investigate the development of reading abilities (Jones et al. 2008; Rau et al. 2016; Evans and Saint-Aubin 2005; Benfatto et al. 2016), there is considerably less eye-tracking based research in the domain of learning mathematics, and specifically regarding children and mathematics. However, one of these few studies (Schneider et al. 2008) actually presented a validation of eye-movements as a measure for developing number sense in children. The authors examined an exercise that involved a ‘mental number line’ and investigated to what degree 7- to 9-year-old’s fixational accuracy was related to math proficiency, finding that fixation accuracy did increase with proficiency. Furthermore, they concluded that the use of eye-tracking as a measure of developing number sense has both validity and utility.

Number line estimations were also used by Heine et al. (2010) in a study where 6- to 9-year-olds were presented with a number line in the middle of a computer screen, with start and end points labelled with 0 and 100, respectively. Numerical stimuli in the form of one number at a time were presented on the screen. Children were instructed to actively search for and focus their gaze on the correct number line position for each number. A marker appeared after 4000 ms. Children had then to decide as fast as possible, by clicking a button, whether the marker indicated the correct position on the number line or not. The eye-movement data revealed that compared with children from lower grades, older children shifted their gaze significantly more often and for a significantly longer time to the respective *correct* positions on the number line in trials where the answers they provided were *in-correct*. The authors interpret this to mean that the older children understood more about these numbers than the younger ones, even though the actual answers between the groups did not differ. The older children’s eye-movements might be manifestations of a growing, or partial, understanding at a transition stage. In other words, they might reflect an aspect of the transition from not understanding to fully understanding.

More recently, Bolden et al. (2015) did a pilot study with nine 9- to 10-year-olds with the aim to explore the feasibility of eye-tracking for investigating children’s mathematical knowledge and understanding. The specific focus was on how children interpret mathematical representations. Results from the study were that representational forms like equal groups and arrays were more successful than number line representations when testing multiplicative reasoning and that the success rate was related to children's general mathematical ability. The authors conclude that the study demonstrates the usability of modern eye-tracking technology in the information rich environment of a school classroom. However, the visual information provided to the students in the study of Bolden et al. (2015) was static. What they looked at were a set of static slides, each including a symbolic and a picture representation of a multiplication problem. Likewise, the visual stimuli used in the study by Schneider et al. (2008) were composed of simple and primarily static visual materials.

Turning to the area of mathematics and visual anticipation, Hintz and Meyer (2015) conducted a study with adult subjects. On alternating trials, the participants heard a complete equation, e.g., "Three plus eight is eleven." or heard the first part, e.g., "Three plus four is…" and had to produce the result ("seven") themselves. During all trials they saw a clock face before them, featuring the numbers 1 to 12, and they were encouraged to look at the relevant numbers throughout the trials while their eye-movements were measured. Results were that in both kinds of trials participants fixated the first and the second number of the equations shortly after they were mentioned and fixated the result number well before they themselves named it and well before the recorded speaker named it.

Teachable Agents

In addition to shedding light on children's processing of number in the game we were interested in knowing more about their understanding of the *teachable agent* in the game. In essence, educational software using teachable agents (TAs) is a digital implementation of the pedagogical approach *learning by teaching* (Bargh and Schul 1980). The human student takes the role as teacher and instructs the teachable agent and learns herself by doing this (Brophy et al. 1999).

There are many studies involving school children who use teachable agent-based games, that show pedagogical power of TA-based games both in terms of learning outcomes and motivational effects (Biswas, Leelawong, Schwartz, Vye, & TAG-V, 2005; Wagster et al. 2007; Chase et al. 2009; Gulz et al. 2011; Pareto et al. 2011; Pareto et al. 2012; Lindström et al. 2011). Using a TA-based game can boost a variety of aspects, e.g., time spent, willingness to revise, willingness to search for more information, conceptual understanding, metacognitive processing, number of solved tasks, and so on.

It is, however, uncertain if these benefits apply to preschoolers as well, with their less developed *theory of mind* (Perner 1991) i.e., the ability to recognize that other people can differ in their understanding, feelings, and knowledge of various things compared to your own understanding, feelings, and knowledge of the same things. To teach someone else, you need to be able to recognize that the person being taught does not know everything you know yourself, that she has her own knowledge that she acts upon, and that you can influence her knowledge. According to literature, many 4- to 5-year-olds do not pass standardized theory of mind tests in the way older children and adults do (Perner & Roessler, 2012; Wellman and Liu 2004). This is the reason why there is a question mark as to whether benefits of TA-based games showed for school children will also apply to preschoolers. *It is not obvious to what extent preschoolers can understand the teaching metaphor in a TA-based game.*

Haake et al. (2015) tried to address the question in a previous study where 3- to 6-year-olds played the Magical Garden game. More specifically the researchers used an interview method to explore to what extent children could reason about their teachable agent's actions in ways that reflected an understanding of the teaching metaphor in the game. Central questions were: "What is this game about?"; "What does Panders [the name of the TA] do in the game?"; and "Why is Panders in the game?" Examples of answers were: "Panders wants to join and he wants to help the baby birds too"; "Panders thinks about which button he thinks is right and which is wrong." According to the results, many of the participants were capable of reasoning about their teachable agent as an agent with beliefs – correct and erroneous ones. In contrast, the children's results on standardized theory of mind tests often predicted that they would not be able to reason in the ways they did.

There are two reasons to seek a complementary measure for children's understanding of their TAs in the form of gazes. One is that not all preschool children will answer interview questions; for instance, in the study by Haake et al. (2015) some children talked much, some very little. Gaze behaviors, on the other hand, are equally measurable in more and less verbal children, and in more and less talkative children. The other reason is that when addressing cognitive phenomena that by their nature can only be indirectly inferred via measurements, triangulation of methods is recommendable (Altrichter et al. 1996; Lincoln and Guba 1985). Therefore, we wanted to see whether the understanding of a teachable agent – that the study by Haake et al. (2015) indicates that many preschool children have – would also manifest itself in children's gaze behaviors during game-play.

A sufficiently developed Theory of Mind can be seen as a precondition for benefiting from the learning-by-teaching pedagogy in a TA-based game. Another precondition, where there likewise are grounds for doubt as to whether the precondition is at hand for preschoolers, is the ability of maintaining attention and focus on the teachable agent when this is required in order to understand and learn via the pedagogy of learning by teaching. Are 4- to 6-year-olds sufficiently capable of maintaining focus on a teachable agent in an educational game when, one the one hand, there are things happening in the physical environment (as is often the case at a preschool) and, on the other hand, there is a variety of visual, both static and dynamic, elements, on the screen in addition to the teachable agent? This question was addressed in an eye-tracking study by Axelsson et al. (2016) using an experimental version of Magical Garden, which included experimentally introduced visual distractions on the screen such as a ball rolling by or an airplane flying in the air. Before the experimental part of the study the children had been using the standard game, without any distractive animations, to familiarize themselves with the game. An additional rationale was to make sure that we did not measure novelty effects. Results were that a majority of the 36 participants were quite capable of maintaining focus on their teachable agent in spite of experimentally introduced visual distractions on the screen. Many children who did not succeed well in an inhibition pre-test, nonetheless managed to inhibit distractions during gameplay. In addition, the children were significantly less distracted when the teachable agent was solving an early math task than they were when the agent had solved a task and only other things were going on. This selective inhibition of distractions in effect increased the likelihood that the children would attend to educationally important features of the game. In sum, the study made use of eye-tracking measurements but neither did it address children's cognitive processing of number, nor did it address children's understanding of their teachable agent (other than in an indirect way).

The Main Difference between this Study and Related Previous Studies

The purpose of our study – conducted March–April 2015 – was to examine whether an analysis of children's gaze behaviors during game-use could shed light on whether:

- a child processes the information meant to support number sense development and connects the different representations to one another,
- children perceive their teachable agent as an entity that acts goal-directedly and in relation to its own understanding or knowledge.

In relation to previous studies on how preschoolers understand teachable agents, our study has methodological similarities with that of Axelsson et al. (2016), but the central topic in that study was attention and focus in these young children, not their understanding of a teachable agent. The study of Haake et al. (2015) on the other hand addressed precisely that. It did not, however, involve any eye-tracking measurements and can, as discussed above, be seen as a complementary study. We wished to examine whether preschoolers' understanding of their TAs suggested by this previous study would also be manifested in their gaze behaviors during game-play. Thus, with the present study we aimed at complementing the previous study with measurements of a more objective kind.

Turning to the (early) mathematics area, our study is, as far as we know, the first eye-tracking study on number sense or number processing with preschoolers, and the first that captures eye-movements while participants look at dynamically unfolding information in a digital game context. The children in the study received no instructions about what to do, or where to look, and the visual stimuli they were exposed to were rich and included dynamical animated sequences partly unfolding in response to their own actions. Previous eye-tracking studies (with older children and adults) as those described above have used screen scenarios with few and primarily static visual elements. The screens have been presented, one after the other, together with an explicit task, such a number line task, where participants have been provided an instruction on what to do and sometimes where to look.

Why bother to involve dynamical visual representations? If we are interested in how children learn mathematics and solve mathematical tasks, we need to recognize that children are increasingly using early math apps, both in pre-school and at home. These apps are typically visually rich and involve dynamical visual representations. Data-logging methods can provide insights in the unfolding of children's behavior and responses in these kinds of digital environments, and the possibility of combining this with parallel data on gaze behaviors is at least potentially interesting.

It should be pointed out that our study in this sense should be seen as a methodological precursor, investigating methodological possibilities to complement the data logs associated with the step-by-step unfolding of game play with eye-tracking data to provide additional indications of a player's processing, understanding or non-understanding.[4] In addition, there is a methodological reason for exploring the potential use of eye-tracking analyses with children as young as in our study, as the collecting of valid eye-tracking data does not rely on abilities to verbalize or discuss what one is doing, nor on subjects' willingness to follow instructions from an experimental leader.

Methods

Participants

Participants were 36 children (21 girls, 21 boys), 4- to 6-years-old from three different preschools in middle-class SES environments in the south of Sweden. Two children (girls) were excluded from the study, one due to sickness leave and one due to eye issues that caused erroneous calibration of the eye-tracker, and an additional four children (two girls and two boys) were excluded due to partial loss of eye-tracking data during the experimental session. Thus, the final data set consisted of 36 participants ($N = 36$; 17 girls, 19 boys) with the age distribution presented in fig. 1 ($Med = 4$, $Min = 4$, $Max = 6$).

The study was a part of a research program approved by the Regional Ethical Review Board of Lund.

The Digital Play-&-Learn Game

Next we describe the full, standard game used in the first, nonexperimental part of the study, where children played the game over three weeks in order to get accustomed to it. Thereafter we describe the modified version of one of the sub-games that was used when the children then participated in the second, experimental part of the study.

The Regular Game

Magical Garden (Gulz and Haake 2019; Haake et al. 2015a; Haake et al. 2015b; Husain et al. 2015) builds on work by Griffin et al. (1994) and is designed for 4- to 6-year-olds to learn and practice number sense and basic math skills within the number range 1–9. On the most basic levels the game involves neither symbolic nor iconic representations of numbers, beyond spoken number words. All actions are performed on virtual objects in virtual space. Children are introduced to crucial concepts such as higher/lower, longer/shorter, upwards/downwards, larger/smaller, too few/too many, and more/less, along with relations between them. Thereafter they practice these concepts and their relations to sets of different sizes as *represented* in a variety of ways. Number words are thus related to magnitude and a variety of visual number representations, starting with iconic representations (e.g., a hand with fingers held up), progressing via the semi-symbolic (e.g., slashes as found in Roman numerals) to fully symbolic representations

[4] In turn, such indications might be used for: diagnosis, game evaluation, user modeling and real-time adaptation in terms of scaffolding and feedback (Wikholm et al. 2019).

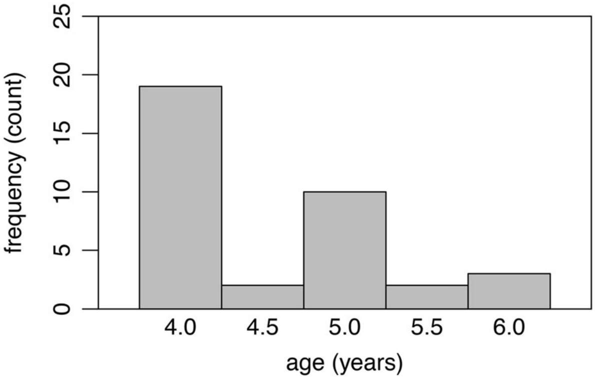

Fig. 1 The age distribution of the participating children ($N = 36$)

(Arabic numbers). The overall goal is to ensure that number concepts are well grounded and integrated with each other through a variety of forms of representation.

Magical Garden makes use of the pedagogical principle of *learning-by-teaching* with the child taking on the instructor's role, helping a digital tutee (or *teachable agent* [TA]) (Biswas et al. 2001) solve progressively more difficult tasks. The child is introduced to three characters – a mouse, panda, and hedgehog – whose garden is barren and in desperate need of watering. The child chooses one as her friend, whom she will help collect water drops to bring the garden back to life (see Fig. 2).

The game comprises 60 scenarios, ordered by difficulty defined by number range (1–4, 1–6, or 1–9), representation (none, fingers, dots, dice, number symbols) and method (counting, proto-addition/subtraction, 'true' addition/subtraction). All scenarios can be presented through several sub-games that feature distinct narratives: e.g., a nearsighted bumblebee needing help to find the right flower or a treasure hunter needing help to reach one of several caves in a cliff by attaching balloons to her basket.

Regardless of sub-game, any given scenario is always repeated in three successive pedagogical modes (Fig. 3): after having been introduced to the task the child practices on her own (mode 1); then the child shows her the TA how to do the task (mode 2), and, finally, the child supervises the TA who attempts the task (mode 3). Supervising involves accepting the TA's answer (when judged correct) by pressing the happy smiley, or otherwise pressing the unhappy smiley and correcting the TA. As an example: When the child, in *mode one*, has helped three

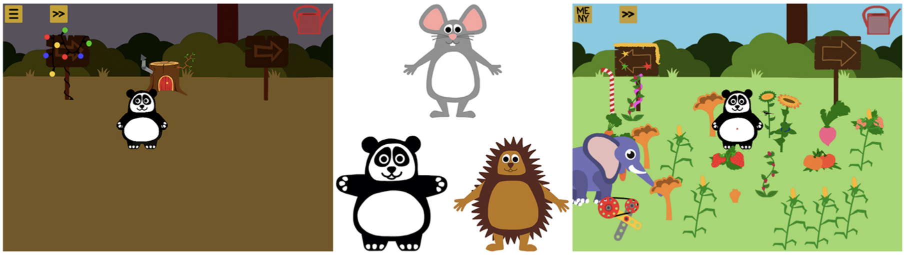

Fig. 2 Screenshots showing the initial garden (left); the three characters Panders the panda, Mille the mouse, and Igis the hedgehog (center); and the thriving garden (right)

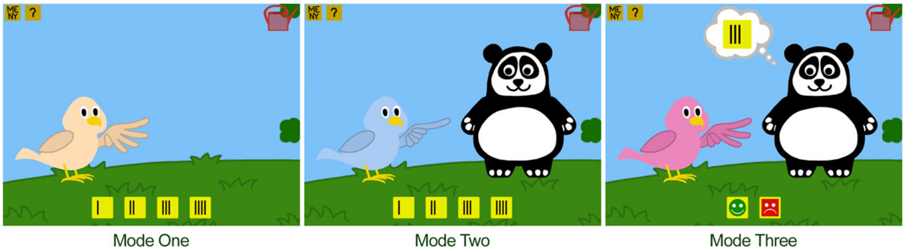

Fig. 3 The three pictures display the three different modes of gameplay in Magical Garden. In *mode one* the child chooses which button to press (the row at the bottom of the screen) without the TA present. In *mode two* the child chooses which button to press with the TA watching. In *mode three* the TA proposes a button (displayed as an image of the proposed button in a thought bubble) and asks the child if the button it thinks about is the correct one. The child answers by pressing either the happy or the sad face. If the child answers that the TA's proposal is wrong (pressing the sad face), she has to show the TA the correct answer by pressing the correct button

baby birds, the TA enters the game and says s/he wants to help the baby birds as well, which initiates mode two. In *mode two*, the TA stands in the background and watches the player continue helping three baby birds, sometimes commenting upon what the child is doing and what happens. Furthermore, the TA's eyes are programmed to follow the pointer at the screen. In *mode three*, the TA says s/he wants to try, and the child helps the TA. Having completed the three modes, the child and TA receive three water droplets that can be used to water the magical whereupon various plants – some strange and some 'magical' – grow, piece by piece. The TAs as well as other characters in the game communicate via their own personalized voices and they show some different facial expressions, such as happiness, bewilderment and disappointment, that correspond to outcomes in the game.

Magical Garden is implemented with a state machine that keeps track of the progress of each child, steering each child by different paths through the game. The amount of repetition and the places at which repetition takes place differ will vary among individual children, as well as the kind of feedback provided When a child masters a scenario, she moves on; when she encounters difficulty, she repeats and practices the same scenario – through varying sub-games – until her performance improves. If the child's performance remains low, the system goes back a half or entire level in difficulty so that the child can practice and prepare for a new attempt on the higher level. In this way some children get a substantial amount of training with a certain kind of task, whereas others quickly leave the same task behind. The basic logic is that a child who masters a pedagogical scenario moves to the next, while a child that has trouble repeats the same scenario (with different sub-games) until her performance improves. From the child's perspective, the mix of sub-games presents variety, even when the pedagogical challenge does not vary. This makes it possible to repeat a given task without too much of monotony. In pedagogical terms Magical Garden relies on *dynamic assessment*, i.e., integrating assessment and instruction (Vygotskij 1978), to address a 'potential for learning' rather than a 'static levels of achievement'. The goal is to make each child work at a level just beyond current mastery (*the zone of proximal development*) to maximize learning (Vygotskij 1978).

'The Bird Rescue' Sub-Game – The Standard and the Experimental Versions

The goal of the sub-game Bird Rescue is to help baby birds get to their parents that are on a specific tree branch, by sending up the baby bird in an elevator carriage made of a bucket (see Fig. 4). The baby bird holds up a number of feathers to indicate which branch it wants to go to, and the task is to press the button representing the number in question. The baby bird jumps into the tree-elevator and the elevator displays on the front of the bucket, in symbolic number, which branch it is currently at. If the elevator stops at the correct branch, the bird jumps off and is welcomed by its parent and they cheer and tweet. If an incorrect button was pressed, the baby bird says: "This is not my parent! I live further down / higher up." Thereafter it is returned to ground floor and the player may try again.

In addition to the symbolic number on the bucket, there are three different number representations that need to be connected. First, there is the set of feathers that the baby bird shows that represents an amount. Second, there are the iconic representations of amounts on the elevator buttons (which kind of iconic representation that is used depends on how far the child has advanced in the game). Third, the tree with its branches represents a vertical sequence from bottom to top, from the first to the ninth branch. The combination of these three representations in the game make up an operationalized exercise of: number identification, ordering of representations and relating these representations and concepts to one another.

In more general terms a full-blown number sense involves the ability to connect and map the following to one another: (i) different iconic representations, such as a set of tallies or a dice with a number of spots, (ii) steps on vertical and horizontal lines of different kinds, (iii) a variety of groups of objects. It is only when these mappings or connections can be made correctly, that a child can refer them all to the respective mathematical symbols (1, 2, 3, …) and have an understanding what these arbitrary symbols stand for. With this as rationale we set out to implement an experimental version of Bird Rescue that could serve in accessing aspects of children's (non)understanding regarding the relations of: the set of feathers that the baby bird

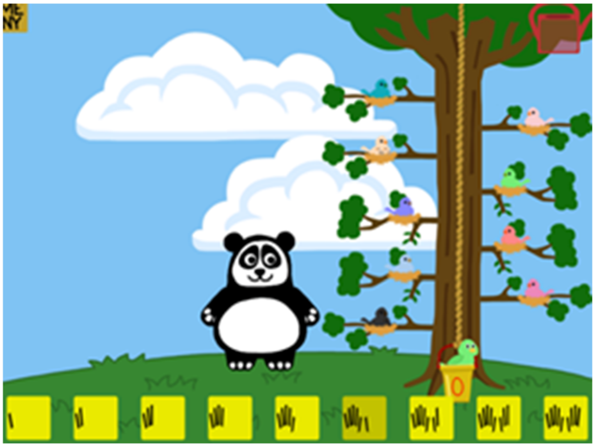

Fig. 4 The sub-game Bird Rescue with the components the tree with nine branches, the baby bird, the elevator, the buttons, and the teachable agent Panders

holds up, the branch where the baby bird's parent is waiting for it (first, second, third, …; higher than…, lower than…) and the elevator button that represents the same number.

For an experimental version of the sub-game we, thus, implemented an event that we name a *malfunction*, where the elevator will not end up at the branch corresponding to the button the child or the TA has chosen. Instead, the elevator continues *past* this branch and eventually gets stuck in the treetop where a 'service bird' repairs the elevator before returning to the chosen) branch. Responses to such malfunction would, we argue, differ between children who are correctly (or almost correctly) connecting the different representations (numbers of feathers and representations on buttons with the corresponding vertically ordered branch) and children who are not processing this information in a helpful or adequate way. In other words, we predict differences due to whether a child has or has not a more developed number sense in this respect. Only with a sufficiently developed number sense will a child identify which number the baby bird shows with its feathers, which lift button to press, and the ordering of the branches in the tree – and understand how all these representations of number relate to one another. And only then can a child have an *expectation* that the elevator will move to the branch at the position that matches the representation on the elevator button and the number of feathers held up by the baby bird. Indeed, all children are likely to understand that something strange has happened at the moment when the *elevator gets stuck in the treetop*. However, to realize or understand that the elevator *passes the correct branch*, one must be aware of which branch is the correct branch. Importantly, there are no bottom-up salient stimuli present that can indicate malfunction when the elevator passes the branch that corresponds to the button pressed, only previous knowledge or understanding on the part of the child – in this case understanding that relates to number sense. If a child does not expect the elevator to stop at a specific branch, there is nothing unusual or unexpected in its continuing to move, and there are no bottom-up salient stimuli that can influence her into gazing back towards the branch that was 'erroneously' passed by.

The experimental version of Bird Rescue was designed to contain a mixture of tasks or trials where the elevator was functioning all well (as in the regular version of Bird Rescue) with tasks where the elevator was malfunctioning. We refer to them as *standard trials* and *manipulated trials.* A second modification of the experimental version of the Bird Rescue sub-game served to address our second research topic on whether children understand that the TA is an agent with its own knowledge that it acts upon. A description of this modification follows: In the regular version of the sub-game there are always three consecutive modes (see above). For this version we wanted to compare situations where the child and the TA, respectively, is 'in charge of' what happens in the game. Therefore, we chose to combine the regular mode two, where the child decides which button should be pressed while the TA is watching, with a modified mode three where the TA suggests which button to press (which can be seen in the TA's thought-bubble) and then also presses the button while the child observes. (This differs from the regular game's mode three where the TA suggests which button to press but the child gets to decide whether the TA's suggestion should be effected or not.)

In the *standard trials* the elevator moves to the branch indicated by the button pressed by either the child or the TA. If the correct button is pressed, the baby bird reaches the correct branch. If the incorrect button is pressed the baby bird lands on the chosen branch and says: “This is not my parent! I live further down / higher up.”, whereupon the bird is returned to the ground floor for a novel try. This continues until the correct button is pressed, with no maximum number of trials. For the standard trials, the correct branch – the one the baby bird wants to go to – is in the number range of 1–9.

In the *experimental trials* the malfunction occurs when the child or TA presses the correct button. If the child choses the wrong button, the elevator goes to the incorrectly chosen level, the baby bird says it is the wrong level and the player needs to try again. When the correct button is pressed by the child or TA, the elevator passes it and goes to the treetop. Then, when stuck in the treetop a service bird enters the screen and comes to rescue, saying: “Oh dear, oh dear, the elevator is broken! Don’t worry, I will fix it.” (The service bird then repairs the elevator and the elevator moves down to the correct branch.) The motive for having a service bird taking care of the event was to avoid that children would get upset about the baby bird being stuck in the treetop or wonder if the game was broken. (The TA is programmed in the experimental trials in the manipulated version of Bird Rescue – but of course not in the regular game – to always chose correct buttons. The reason is to expose children to sufficiently many events of malfunction.) For the experimental trials the correct branch is in the number range of 1–6. The reason for this is that children should have an opportunity to notice the malfunction before the elevator reaches the treetop and the manipulation is disclosed as the elevator stops and the service bird enters. Note that the visual scenes with the buttons, the tree, etc. are identical for standard and experimental trials. The only effect on part of the children is that some standard trials can be more demanding (in terms of number range) than experimental trials.

Before each new trial the bird parents are randomly assigned to the different branches in the tree and the bird’s colors are randomized. The color of the baby birds is also randomized since the correct branch for each task is randomized. Therefore, the colors of the birds cannot be used as cues for noticing a malfunction. Which branch each baby bird wanted to go to is randomized within its number range.

A round of the experimental Bird Rescue sub-game consists of the following four *trials*; (1a) standard trial with child in charge and TA watching, (1b) manipulated trial with child in charge and TA watching, (2a) standard trial with TA in charge and child watching, and (2b) manipulated trial with TA in charge and child watching. The trials within a round always come in the same order (1a, 1b, 2a, 2b).

Technical Set-up

The eye-tracking experiment was performed on a Dell laptop (Intel Core i7 CPU 2.67 GHz, 2.98 GB RAM) running Windows XP Professional using Experiment Center version 3.5 (SensoMotoric Instruments). Magical Garden was played on a separate screen a 22″ widescreen monitor (1680 × 1050 pixels). Eye-movements were measured using an iViewX SMI RED250 remote eye-tracker (SensoMotoric Instruments) that recorded binocularly at 250 Hz. The eye-tracking data were aligned with a screen recording of the gameplay and analyzed in BeGaze version 3.5 (SensoMotoric Instruments). An additional laptop was used for the researcher to write down comments and utterances from the children during the eye-tracking experiment. Figure 5 displays the set-up for the eye-tracking experiment at the preschool. JavaScript and HTML5 were used to program the experimental version of Magical Garden.

The use of eye-tracking with a remote eye-tracker enabled us to conduct the experiment at the children's preschools and not in a laboratory setting, which increases the ecological validity of the study.

Calibration

The calibration method was a 9-point routine with a black circle on light grey background. The calibrations points were accepted by the experimenter hitting space bar. The validation was the standard 4-point with a white dot on a light grey background and accepting calibration points automatic when fixation was stable. In this study a re-calibration was performed if a child deviated more than 1.7° on either eye.

Procedure

The first part of the study had a duration of 3 weeks, where the children played Magical Garden on tablets at preschool 1–2 times per week. They played individually on their own tablet and with their personal log-in but sat together in small groups. Each session lasted 15–20 min. The play-time during this part of the study ranged between 50 and 90 min in total, with a mean play time of 70 min. At the first occasion a researcher accompanied the children, one at a time, to make sure each child was able to use the game and had an overall understanding of the idea of the garden, solving tasks and receiving droplets. The game narrative is introduced by game characters and, thus, the researcher was mainly checking that the child listened and acted accordingly. Thereafter the playing sessions were directed by the teachers at the preschools with whom researchers had close communication, and the researchers visited at several occasions. Each time the children were logged in on their individual account with their own 'magical garden' and digital friend (the TA). Each preschool had several tablets

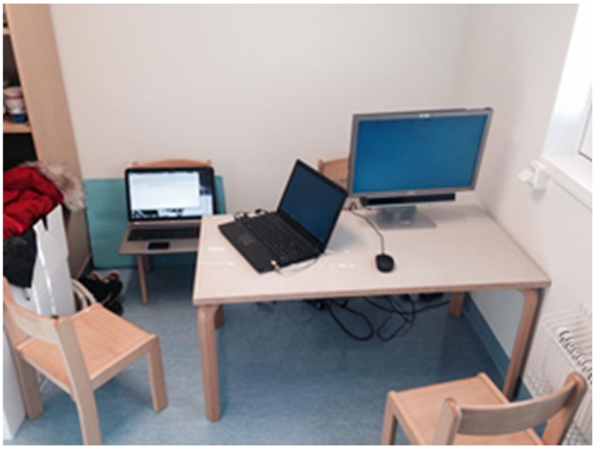

Fig. 5 Picture of the eye-tracking set-up at the preschool. Left: The laptop for notetaking; Middle: The laptop running the experimental control center (Version 3.5; SensoMotoric Instruments); Right: The 22″ monitor with the iViewX SMI RED250 remote eye-tracker

available, and standard was that two to four children played during a session – each with their own tablet and focusing on that. All the game data from the tablets was automatically logged on a server. At the occasions where researchers visited, they took notes on spontaneous utterances from the children during game-play. There were two rationales behind the non-experimental part of the study. One was that that the children when they participated in the experimental part should be familiar with the researchers and understand the narrative of the game, its structure and interactive rules. This to make sure that we would measure understanding related to number sense and the teachable agent as aimed for, and not understanding related to the narrative, the structure or how to interact. The other, that we during the non-experimental part of the study collected game data and spontaneous utterances for research purposes beyond the topics of this study and article.

The second, experimental, part of the study took place at preschool as well, in one of the smaller rooms that was familiar to the children and often used for drawing activities in smaller groups. During the session a teacher was present if children wanted this, sitting on a chair a few meters away. The average duration for a participant including calibration was 30 min. The set up (see Fig. 5) consisted of a laptop controlling the experiment. The laptop was, furthermore, connected to a second external monitor (screen) where the child was presented with the experimental stimuli (the educational game with malfunction events). The external monitor was, in turn, equipped with a remote eye-tracker and a mouse (for interacting with the game). A second laptop was used for researchers' notetaking. In this part of the study children participated individually, one at a time.

The eye-tracking equipment we had available was a remote eye-tracker connected to the computer screen that the game was run on. The decision to use a computer screen and not a tablet came from the following considerations. To record accurate data with a remote eye-tracker of this kind, one needs the participants to stay within a rather confined tracking space defined by the angle to the screen and the integrated, remote eye-tracker as well as the distance. Children's standard use of tablets involves moving (and tilting) both the tablets and themselves around. Thus, another solution was necessary. The option of fixating a tablet within a frame on the table would solve some of the challenges; however, using a tablet means that hands and fingers will interfere with the documentation of the central eye-tracking measurements (AOI hits).

Since most children lacked experience with handling a computer mouse, the experimenter conducted the clicking most of the time. The children instead pointed at the screen to indicate which elevator button should be pressed. During their previous tablet gameplay, the children had pressed the tablet's screen to indicate buttons, so this was a natural solution to the mouse handling problem. The experimenters' handling and clicking the mouse is likely to affect children's gaze. However, and importantly, this did not take place during the periods of time that the crucial eye gaze behaviors were measured.

To start the experiment a calibration of both eyes was made. The children were told to keep their head still and look at a dot when it moved across the screen. To convey the importance of calibrating the eyes a cover story was told saying that looking at the dot and following it was the password to unlock the game and to start playing. With a too extended calibration we would risk that children would lose interest in participating. Therefore, a maximum of four re-calibrations were performed and children who still had a deviation of more than 1.7° were allowed to continue the experiment anyway and data was collected and used. This was a trade-off between an accurate calibration and the children's compliance and willingness to participate.

The children were asked which friend (Panda, Mouse, or Hedgehog) they wanted to play with. Each child played a minimum of five rounds and maximum ten round of the experimental version of the sub-game Bird Rescue. One round consisted of two standard trials (one with themselves in charge, one with the TA in charge) and two manipulated trials (one with themselves in charge, one with the TA in charge). The number of rounds each child undertook depended on the willingness to participate. Some gladly played more than five rounds and some needed motivation just to complete five rounds. Five rounds were set as the minimum for obtaining a sufficient amount of data. After completing the tasks, the TA and the birds cheered and provided water drops to the child as a reward. In between each round the children were given the choice to use the water drops they had collected to water their garden, as is done in the regular version of Magical Garden that they had been playing during the preceding three weeks. This off-task served to keep children motivated and was also useful as a break. After completing at least five rounds the children were thanked for participating and given a diploma.

Measurements

Game Performance

During the experimental part of the study a measure of 'game performance' was calculated from the tasks where the child was in charge of choosing buttons, i.e., trials '1a' and '1b'. There was no maximum number of tries for each trial, but the child was allowed to play until the correct level (branch) was chosen. Performance was calculated from the total number of correct answers divided by the total number of attempts on each trial. This performance measure is thus not a value of clearance rate as all children play until they have completed each trial – but it is an indication of how well (on average) each child did complete a round.

Areas of Interest as Basis for Measuring Gaze Behaviors

The eye-tracking measure we chose for measuring gaze behaviors, specifically attention towards specific areas, was 'area of interest hit' (AOI hit). Holmqvist et al. (2011) defined an AOI hit as follows: "AOI hit, which states for a raw sample or a fixation that its coordinate value is inside the AOI" (Holmqvist et al. 2011, p. 189). In accordance with previous text passages, we do not claim to measure attention in any straightforward way. 'Gaze fixations' do not equal 'attention', but measurements of fixations are tools for identifying possible instances of attention. The reason we chose AOI hit over other measurements, such as dwell time, number of fixation and attentional shifts, was because we were interested in whether children looked at specific AOIs during certain time intervals. There is no consensus for how to set the lower level of fixation duration as indications of attention. For instance, fixation durations used by Rötting (2001) ranged from 60 to 120 ms while Granka et al. (2008) used 200 ms as cut off. In this study the lower level of fixation duration for an AOI hit was set to 150 ms. In other words, to count as an AOI hit children had to fixate inside the AOI for longer than 150 ms; fixations shorter than 150 ms were not recorded as AOI hits even though they were located on the AOI.

For the study we constructed four AOIs: *the correct branch*, *the pressed elevator button*, *the TA's thought bubble* showing which elevator button to press; and *the TA*. These were defined in order to quantitatively measure AOI hits (fixations within the AOI).

Look-Back at Passed Correct Branch

What we term a 'look back' was delimited as follows. The child fixates the AOI 'correct branch' during the specific timeframe given by: From when the elevator passes the correct branch until the elevator reaches the treetop. (See Table 1). We wanted to shed light on whether a child would direct attention towards the correct branch when the elevator was malfunctioning. Note the technical definition of 'look back'. It focuses on what from an objective standpoint is the correct branch, which a moving object (an elevator) passes by, and measures participant's looks at the branch after the elevator has passed the branch and until the elevator reaches the treetop. If we had constructed AOIs for each branch and the top of the tree and measured transitions between AOIs we could have identified, for instance, gaze transitions between branches higher up in the tree and (back) to the correct branch. However, it is not clear that such analyses would have revealed more with respect to our research questions.

Look at Choice

This refers to a look at 'the pressed elevator button' and/or 'the TA's thought bubble showing which elevator button to press' from when the elevator passes the correct branch until it reaches the treetop. Both areas hold information about the actual chosen 'correct branch', i.e., the thought bubble contains an image of the selected button (cf. Figure 3).

Look Ahead

In the eye-tracking domain, 'anticipating eye-movements' mean eye-movements towards an area before an anticipated event/object/item/entity has appeared within that

Table 1 Display of the different measures of AOIs: measurement name, location on the screen, and specified timeframe

Measure	AOI	Timeframe
look back	The correct branch	From when the elevator passes the correct branch until the elevator reaches the treetop.
look at choice	The pressed button *alt.* The thought bubble	From when the elevator passes the correct branch until it reaches the treetop.
look ahead	The correct branch	From the choice of the correct branch (pressing the elevator button) until the elevator is one branch below the correct branch.
look at TA after malfunction	The TA	From when the elevator passes the correct branch until the elevator (after intervention by service bird) arrives at the correct branch.

area. ‘Looks ahead’ refer to looks at the correct branch from the moment the button is pressed until the elevator is one branch below the correct branch.

Look at TA after Malfunction

‘Looks at TA after malfunction’ refer to looks at the AOI for the TA from the moment the elevator passes the correct branch, until the elevator (after intervention by service bird) arrives at the correct branch (see Table 1).

Descriptions of Measures

All measures described above were recorded under the same protocol (fixation criteria and coding) with respective timeframes defined in Table 1.

Data Collection and Coding

The eye-tracking data was collected during the experimental trials (not the standard trials) using the following protocol: If there was at least one fixation inside the AOI for a period longer than 150 ms, the exercise was registered as a *hit* (coded as a 1), otherwise the exercise was registered as *miss* (coded as a 0). Additionally, a ‘look back’ hit was not registered if the child was looking at the correct branch at the time the elevator was at the correct branch and continued to look at this correct branch after that the elevator had passed (i.e., when entering the specified timeframe for measurement). It was, in other words, required that the child first had moved her eyes away from the correct branch before entering the measurement time frame and have a ‘look back’ event count as a hit. The coding was done manually.

Furthermore, data was coded as *NA* (missing data) if the recorded gaze was flickering or moving over the screen and thus not suitable for a reliable evaluation. For the specific case of no attainable data due to the child not looking at the screen and thus no recorded gaze cursor – the data was coded as a miss (0), i.e., ‘not looking at the AOI’. If this case of lost data was coded as NA, there would be a loss of information affecting the results as the relative proportion of hits would increase, giving a higher level of attention to AOIs.

The gaze behavior measures were then expressed as “the mean probability that an AOI was attended to within the critical period”, that is (for each respective measure) *the sum of* ‘number of recorded hits recorded during experimental trials for each child’ *divided by* ‘the total number of played experimental trials for each child’.

Research Questions and Hypotheses

Research Question 1

Will eye-gaze responses during the experimental trials with the malfunctioning elevator vary according to whether a child connects the different representations in the game to one another or not? If so, eye-gaze measurements could be used to evaluate a participant’s number sense proficiency.

Below we describe our hypotheses for how three different gaze behaviors measured in the experimental part of the study will differ between children in relation to their number sense proficiency.

H1: Look back Children who do not connect the number of feathers, the representation on the correct elevator button and the corresponding branch in the three, cannot know which branch is the correct one, i.e., the branch where the elevator will go. They will therefore not notice that the elevator passes the correct branch and have no reason to 'look back' to that branch. Indeed, if a child does not expect the elevator to stop at a specific branch, there is nothing unusual or unexpected in its continuing to move. In effect the very movement of the elevator is the most bottom-up salient object in the scene that as default will attract attention (Corbetta and Shulman 2002), and children are likely to follow the elevator with their gaze and wait for it to stop. *Therefore, we hypothesized that look-backs would not occur unless a child correctly associates the set of feathers shown by the baby bird, the representation on the elevator button and the branch (as vertically ordered)*. In terms of AOI-hits, we hypothesized that there will be a significant correlation between the probability of 'look back' and 'game performance'.

H2: Look at choice This refers to looks at 'the pressed elevator button' and/or 'the TA's thought bubble showing which elevator button to press'. Both are areas that (in an identical format) represent the actual chosen 'correct branch'. *We hypothesized that looks at choice would not be likely to occur unless the child has an understanding, as described above, of which branch is the correct one*. Only with such understanding does it make sense to make a check: "Oh, maybe the wrong button was chosen?" In terms of AOI-hits, we hypothesized that there will be a significant correlation between the probability of 'look at choice' and 'game performance'.

H3: Look ahead An anticipatory eye-movement of an event can be regarded as an indication of knowledge of/about the event since anticipation is a future oriented action based on previous knowledge in order to make a prediction (Pezzulo et al. 2008). *Again, our hypothesis was that looking ahead at the correct branch would unlikely occur unless the child has a sufficiently well-developed number sense, exhibited in an understanding of the number representations being used in the trial*. In terms of AOI-hits, we hypothesized that there will be a significant correlation between the probability of 'look ahead' and 'game performance'.

Research Question 2

Will there be differences in eye-gaze behaviors during two of the modes of game-play that indicate to what extent children of this age have an understanding of their TAs' agency? More specifically, will there be differences in children's attention to their TA in situations of elevator malfunctioning between (i) the mode where the child is in charge of choosing a button and the TA watches, and (ii) the mode where the TA is in charge of choosing a button and the child watches?

H4: Look at TA after malfunction There is no difference between the two modes with regard to the TA's behavior or degree of activity during the time range where the gaze behavior in question (*look at TA after malfunction*) is measured. In other words, there are no bottom-up visual stimuli that may explain potential differences in children's inclinations to look at the TA in the two modes. A child who understands the similarity between herself as an actor who, based on her knowledge, decides on and presses a button, and the TA deciding on and pressing a button based on her/his knowledge, is, we hold more likely than a child who does not have such an understanding of the TA to look more at the TA in the malfunction situation when the TA rather than the child herself is in charge. A child who, in contrast, perceives their TA as a visual moving element on a par with the clouds that move in the sky in the game will look equally much or equally little at the TA in the two different modes. *Our hypothesis was therefore that if children understand the TA as an entity that acts in a goal-directed manner on the basis of own knowledge they would pay more attention to the TA when it is indeed the TA and not the child that controls whereto the elevator with the baby bird is being sent.* In terms of AOI-hits, we hypothesized that if the participants on a group level have an understanding of TA agency, there will be a significant difference in the probability of 'look at TA after malfunction' between the case when the TA is in charge versus the case when the child is in charge.

Results

All statistical analyses were performed using *R version 3.4.3* (R Core Team, 2017). All *p* values are evaluated at an alpha-level of .05 (and adjusted for multiple comparisons when appropriate). All effect sizes are interpreted against the guidelines of Cohen (1988). The analyses were conducted on the final dataset of 36 children (see Method section).

Among the 36 children represented in the data set, 29 children played the required 5 rounds and 7 children played some additional rounds (with one child playing up to 10 rounds). For the following analyses, the data set were restricted to the 5 mandatory rounds (ex~~c~~luding any additional rounds for the 7 participants playing more than 5 rounds).

Eye-Tracking Performance

Table 2 presents eye-tracking performance statistics for the whole experimental session. During the unexpected (malfunction) events the children attended the monitor event to a larger extent than they did during the other parts of the experiment.

Overall Results

For the analyses, four gaze behavior measures were used ('look back', 'look choice', 'look ahead', and 'look at TA after malfunction') together with log data of game performance during the experimental session. The definitions of these measurement variables are presented in Table 1 in the Method section.

Table 2 Eye-tracking statistics (mean and standard deviation)

Performance (eye-tracking)	*M* (*SD*)
Tracking ratio (%)	50.4 (10.7)
Deviation X (°)	0.82 (0.55)
Deviation Y (°)	0.93 (0.62)

Overall descriptive statistics for the gaze behavior measures and game performance during the experimental session are presented in Table 3. As seen from the Shapiro-Wilk's tests, only 'look ahead' followed a normal distribution.

Research Question 1

Game Performance

The participants 'game performance' showed a non-normal distribution (Fig. 6: Skew = −0.37; Kurtosis = −1.19; *p* (Shapiro-Wilk) = 0.002) not yielding to any normalization method (e.g., Box Cox transformation and Yeo-Johnson transformation). Consequently, we chose a more qualitative approach using scatterplots and Spearman's rank correlations for research question 1.

Scatterplots and Correlations

To assess the relations between the three behavioral gaze measures ('look back', 'look at choice', and 'look ahead') and participants' 'game performance', three corresponding scatterplots were produced with correlation lines and 'lowess' smoothing lines (Fig. 7).

Next, corresponding Spearman's rank correlations were calculated (Table 4) to evaluate the correlations between the three gaze behaviors ('look back', 'look at choice', and 'look ahead') and participant's 'game performance'.

Table 3 Descriptive statistics and test of normality (Shapiro-Wilk's test) for overall game performance and overall behavioral gaze measures (AOI hit probabilities) evaluated on participant level (*N* = 36); corresponding variable names in italics within parentheses

Game performance (%)	*Median*	*Min – Max*	*M*	*SD*	*p* (SW) [1]
experimental version	90.9	62.5–100.0	85.2	12.4	0.002
AOI hit probability (%)					
look back	20	0–50	18.3	12.1	0.013
look at choice	10	0–50	16.7	16.7	< 0.001
look ahead	40	0–80	44.4	18.1	0.103
look at TA after malfunction	20	0–80	23.1	22.3	0.001

[1] *p* (SW) = *p* value of the Shapiro-Wilk's test of normality

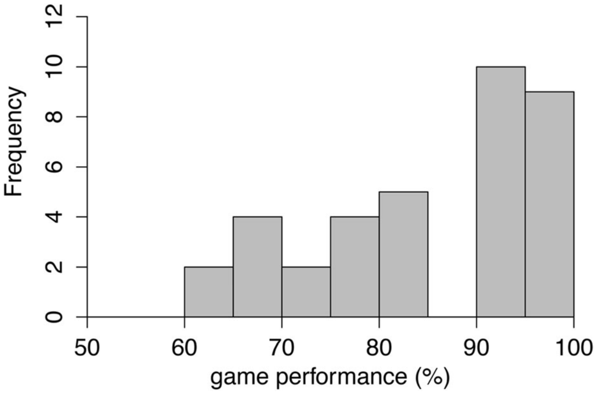

Fig. 6 Histogram showing the distribution for participants game performance during the experimental sessions

Look Back

The scatterplot for the ‘look back’ gaze behavior measure indicates a weak correlation (fig. 7) and the Spearman rank correlation test showed a significant medium correlation (Table 3: 휍 = .379, *p* (adj) = .034). Thus, the ‘look back’ gaze behavior, in support of hypothesis H1, to some degree may reflect the child’s concurrent number processing.

Look at Choice

For the ‘look at choice’ gaze behavior measure, the scatterplot indicates a relatively stronger correlation with a sharp knee at the 90% ‘game performance’ mark (suggesting an underlying split in the correlation strength depending on the concurrent number sense proficiency of the child). The Spearman rank correlation test revealed a large significant correlation (Table 3: 휍 = .588, *p* (adj) < .001) with ‘game performance’. The result supports our hypothesis H2 that the ‘look at choice’ behavioral gaze measure reflects the child’s concurrent number processing with regard to mapping of number representations, i.e., the child’s number sense. In other words, looking at the button that had been chosen correlated strongly with high performance. Indeed, this makes sense, since the chosen button contains information about which branch the elevator was supposed to go to. Looking at the button

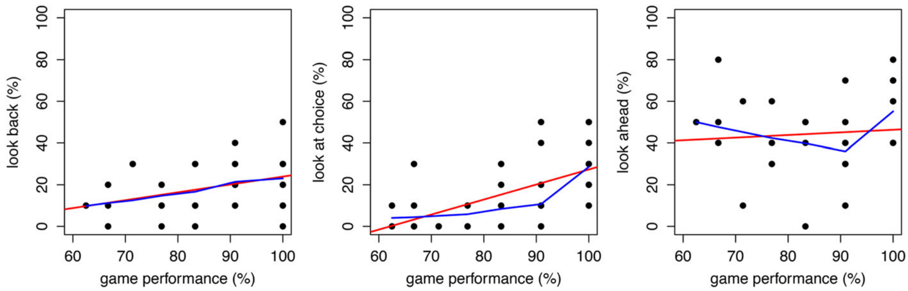

Fig. 7 Scatterplots for ‘game performance’ vs. gaze behavior hit probabilities (‘look back’, ‘look at choice’, and ‘look ahead’); ‘red lines’ represent correlation lines (linear regression lines); ‘blue lines’ represent lowess smoothing (locally weighted scatterplot smoothing) lines. *NB: Data points may overlap*

Table 4 Spearman's 휌 (rho) rank correlation coefficients for the three behavioral gaze measures ('look back', 'look at choice', and 'look ahead') vs. 'game performance'. NB: *p*-values adjusted for multiple comparisons using the *fdr*-method (control for false discovery rates)

game performance vs.	*look back*		*look at choice*		*look ahead*
Rho	0.379		0.588		0.083
p (adj)	0.034	*	< 0.001	***	0.629

p (fdr):. $p < 0.1$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$.

when one finds a mismatch between what one expected to happen and what is happening can be interpreted as a way to orient oneself in the malfunction situation, comparing what the button says with what – unexpectedly – did not fall out as it should.

Look Ahead

The scatterplot for 'look ahead' vs. 'game performance' presents a dispersed distribution of the data points (fig. 7) and the Spearman rank correlation test showed a non-significant small correlation (Table 3: 휌 = .083, *p* (adj) = .629). Thus, the 'look ahead' gaze behavior measure – in contrast to our hypothesis (H3) – does not correlate with number sense proficiency.

On a side note, the lowess smoothing line in the scatterplot for 'look ahead' against 'game performance' – in parallel to 'look at choice' vs. 'game performance' – presents a sharp knee at the 90% 'game performance' mark.

Research Question 2

Research question 2 targets the questions whether the children's gaze behaviors relate to their understanding of their TAs' agency or more precisely: "Do participants attend to their TA when encountering a malfunction and is there a difference in attending the TA depending on who is in charge – the child or the TA?"

The overall statistics presented in Table 3 suggest that the participants on average attended to the TA in almost a fourth of the trials ('look at TA after malfunction': $M = 23.1\%$, $SD = 22.3\%$), however with a large individual variation as seen by the standard deviation. Figure 8 shows the mean probability for a 'look at TA after malfunction' event depending on who was in charge of the decision to press the button. When the TA was in charge, the probability for a 'look at TA after malfunction' event ($M = 28.9\%$, $SD = 28.9\%$) was considerably larger than the probability of when the child was in charge ($M = 17.2\%$, $SD = 20.9\%$). This difference between 'TA in charge' and 'Child in charge' is also significant, as showed by a Wilcoxon signed-rank test ($V = 76$, $p = 0.009$).

In order to exclude the possibility of a few children alone increasing the total mean (as indicated by the large standard deviations), the participants' individual difference between looking at the TA when the TA was in charge and when the child was in charge was calculated for each participant (*Median* = 20, *Min* = −40, *Max* = 60). The result was plotted in a histogram (Fig. 9) and negatively tested for normality (Shapiro-Wilks: $W = 0.931$, $p = .027$). A Wilcoxon signed-rank test indicated that the occurrence of a 'look at

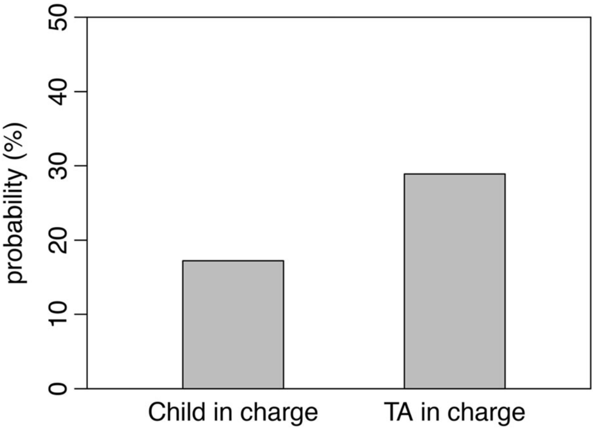

Fig. 8 Bar plot of total mean probabilities of looking at the TA after a malfunction event for all participants ($N = 36$), divided up by who is in charge (child or TA)

TA after malfunction' event when the TA was in charge was significantly higher than when the child was in charge ($V = 275$, $p = 0.009$).

During the experimental session, both verbal reports and non-verbal indications were commonly accompanying the malfunction. Among the verbal reports, utterances like "There is something wrong!"; "No-no-no, wrong!"; and "It broke again!" were frequent during malfunction. The most common non-verbal response to the malfunction was for the child to turn and look at the experimenter. When the child looked at the experimenter, the eye-tracker could not track the eyes, which lead to a decrease in documented detections of AOI-hits.

Discussion

The purpose of our study was to explore whether analyses of preschool children's gaze behaviors during their use of a teachable agent-based early math play-&-learn game

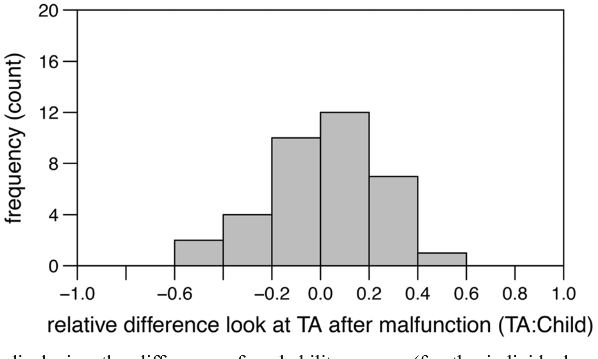

Fig. 9 Histogram displaying the difference of probability means (for the individual participants; $N = 36$) between 'look at TA after malfunction' with 'TA in charge' (TA) and 'look at TA after malfunction' with 'Child in charge' (Child). *Interpretation of histogram:* x-axis < 0: child looks at TA more when child in charge; x-axis > 0: child looks at TA more when TA in charge

could provide insights in whether they: (1) processed the information meant to support number sense development by connecting different number representations to one another; (2) perceived the teachable agent in the game as an entity with *agency* meaning that it is able to act independently and in a goal-directed manner, based on its own knowledge.

RQ1: Do Children Process Number Information during Game Play – Can their Number Sense be Evaluated by Analyses of Gaze Behaviors?

For research question 1, results were that two of the three measured gaze behaviors – 'look back' and 'look at choice' – correlated with number sense proficiency, where a well-developed number sense proficiency entails that that the child processes the number sense information in the game and connects different representations to one another rather than completing tasks on a trial-&-error basis. (In the latter case, the likelihood for high performance is very low.)

The 'look at choice'-measure detects whether a child is looking at the pressed button or at the TA's thought bubble representing the pressed button from the moment the elevator has passed the branch it should have stopped by until it reaches the treetop. This measure strongly correlated with number sense proficiency. In effect, this kind of gaze behavior is extremely unlikely if a child is not actively processing a number of relations between number representations in the game. First, she must connect the vertical number line in the form of the tree to the number representations on the button or the thought bubble. Second, she must connect the number of feathers showed by the bird and the representation on the button or in the thought bubble, which controls the movements of the elevator.

Overall, a 'look at choice' to try to understand what happens in the situation of malfunction does makes sense, but only for a child who understands the malfunction in number sense terms (that is, not just noticing that the elevator gets stuck in the treetop).

A similar explanation applies to the 'look back'-measure which, as well, correlated with number sense proficiency, although less strongly.

We then turn to the 'look ahead'-measure, which as discussed earlier relates to anticipation – a central knowledge- or understanding-driven cognitive activity. In our data, look-aheads to the correct branch is a frequently occurring gaze behavior in the sense that the probability of 'look ahead' behaviors was distinctively larger (by a factor two to four) than all the other measured gaze behaviors. In line with previous research, the 'look ahead' behavior can be interpreted to indicate processes of understanding (in parenthesis, this can be seen as an indication that the early math tasks in the game are on an adequate difficulty level). The measure, however, did not correlate with number sense proficiency.

If we turn to the eye-tracking study by Heine et al. (2010) we may find a possible explanation for this. The older children in their study gazed significantly more (in terms of number and duration of AOI hits) towards the correct answers – but they did not *provide* (by clicking a button) correct answers to a larger extent than the younger children. Yet, the researchers interpreted the older children's increased gazing at the correct answers to indicate a transition from no understanding to full understanding. Correspondingly, we suggest that the common gazing 'ahead at the correct answer' in our study may reflect a transitional phase of understanding – but not mature enough to correspond to behavioral adequacy in terms of a higher number sense proficiency. In other words, the learner increasingly gazes at the correct answer while corresponding behavioral measures still signify ‘incorrect’ or ‘no understanding’. We propose that more studies should investigate the possibility that aspects of gaze can be indicators of processing when understanding is underway, that is of ongoing learning.

A Case for an AI-Model Assessing a child’s Concurrent Number Processing

Based on the findings from our study, gaze measures such as ‘look at choice’ and (possibly) ‘look back’ could be suggested as candidates in a model to assess children’s concurrent number processing. On a speculative note, the ‘look at choice’ and ‘look ahead’ measures both show a knee in their lowess smoothing lines at the 90% ‘game performance’ mark, splitting the participating children in halves; 17 children have a ‘game performance’ below 90% and the remaining 19 children have a ‘game performance’ over 90% (cf. figure 6). This ‘knee’ suggests a segmented (broken-line) model in order to accurately connect tracked behavioral gaze measures to a concurrent number sense proficiency.

Do Children Perceive the Teachable Agent as an Entity with Agency– Can Gaze Behaviors Help us Answer?

The underlying question concerns to what extent children in this target group perceive their TA as an entity able to act independently and in a goal-directed manner on the basis of its own knowledge. Translated to the game context: that their TA acts to help the baby birds to the right tree branches by pressing the buttons with the representations the TA understands to be adequate.

Will there be differences in children’s attention to their TA in situations of elevator malfunctioning between: (i) the mode where the child is in charge of choosing a button and the TA watches, and (ii) the mode where the TA is in charge of choosing a button and the child watches?

Results were as follows. Children attended to their TA after elevator malfunction in about 25% of all trials – that is with either TA or child in charge – but with a significant difference in their tendency to do so between the two modes. The mean probability of looking at the TA was significantly higher when the TA had been in charge compared to when the child herself had been in charge. Importantly, between the two modes there was no difference in the TA’s behavior or degree of activity during the time range where the gaze behavior in question (*look at TA*) was measured, namely after the malfunction has occurred. In other words, no bottom-up visual features can explain different inclinations to look at the TA in the two modes, but such a difference must be explained via knowledge- or understanding-driven factors.

First, the situation is one where ‘something is wrong’. Some children detect that the elevator does not stop where it should, and presumably all children detect that the elevator gets stuck in the treetop. We suggest the following possible top-down influence for why there were significantly more looks at the TA when the TA was in charge when something went wrong. If a child perceives a similarity between her own agency when she decides on and presses a button, and the TA’s agency when the TA decides on and presses a button, it makes sense for her to pay attention to the TA in the situation when the TA had been in charge – but less so in the situation when she herself has been in charge. Specifically, when the TA had been in charge it makes sense to wonder about things like: "What did the TA actually do?"; "Which button did the TA choose?"; "Was it the TA's fault that things went wrong?"; and "Does the TA react to what happens now that it has tried to get the bird to the right branch?" If, on the other hand, it is the child herself who had been in charge, the TA has not been involved as a possible cause or influencer of the problem.

Such differentiated understanding of the two situations requires an understanding that the TA acts independently, on her/his own knowledge and in a goal-directed manner (and is not on par with other visual and dynamic elements in the game, such as the flowers, the helpless baby birds, or the clouds). If the children, as a group, *lacked* such understanding, the difference in gaze behaviors between the two modes would be difficult to explain. Why would children, in that case, pay more attention to the TA when the TA had been in charge?

In sum, the collected results from our study makes likely that at least some children in this group of preschoolers have an understanding of their TA as an entity with knowledge that it acts upon and that is different from the child's own knowledge. While the results do not confirm, they also do not disconfirm the conclusion of Haake et al. (2015a) – conducted with the same game and participants from the same age-group – that many preschoolers have a quite developed theory of mind (more than indicated by performance on standard theory of mind tests). The fact that the present study aligns with this previous study – but uses another method and measurement technique – is an asset from two perspectives. From a research perspective, with respect to triangulation, and from a practice-oriented perspective in that eye-tracking in the future is likely to be built-in in computers and tablets, something that opens for data collection and analyses during game play in the wild.

Methodological Contributions

Collecting Eye-Tracking Data while Learners Use Educational Games 'in the wild'

To our knowledge all previous studies on children's mathematical understanding exploiting eye-tracking have examined children's responses to static visual representations on a screen or a sheet of paper. This study is pioneering in the sense that eye-tracking measures as possible indicators of children's understanding have been collected while they used a fairly complex educational game that involved a variety of visually detailed scenes with dynamical elements, including number visualizations (such as the moving elevator) that respond to the child's (and TA's) actions).

Also, the children were the ones who controlled the game-play (even though the experimenter steered the mouse for them during the experiment) rather than getting instructions on what to do – and on where or how to look.

We hold that we, at least to some extent, have provided a methodological proof of concept that this kind of data collection is feasible and meaningful. Now why is this interesting?

The past years have involved a prominent growth of behavioral data logging and data mining in all kind of areas including the educational domain, where techniques as regression modeling and machine learning are applied to behavioral data. However, it seems clear that additional kinds of data are needed in order to make more progress. Ginsburg et al. (2013) highlight the increasing access to large scale data logs from early math software and the potential of a new step in identifying a variety of learning trajectories, but they also point to the knowledge gap between behavioral data logs and the underlying cognitive processing of individual learners in that more or less identical sets of logged behavioral data can correspond to qualitatively different cognitive processes or vice versa.

A central potential of artificial intelligence in education is the possibility to adapt software to individual students and their degrees of understanding, and there is reason to believe that eye-tracking data may become a relevant source for accomplishing this, maybe even in terms of real-time adaptation (Wikholm et al. 2019). But such eye-tracking data, to be associated with present collection of real-time log data, must be collected during real-world use of real-world educational software – rather than in a laboratory where participants complete experiment specific tasks.

It can be debated to what extent the present study was a study ‘in the wild’, but some features at least contribute to a considerably higher ecological validity than in a laboratory study. The eye-tracking data collection took place in a familiar room at the preschool (where the children often play in smaller groups). The children used an educational game and not a set of disjointed early math tasks. At the time of the eye-tracking data collection they were familiar with both the researcher and the game and all seemed relaxed and clearly satisfied with the situation. For the future, similar data collections can certainly be improved further with respect to ecological validity. Once there are embedded eye-trackers in laptops and tablets, a similar data-collection could be done with small groups of children and without the need for a researcher to interfere in the interaction (e.g., handle a mouse).

The Use of Eye-Tracking with Children of this Young Age

The study extended the knowledge of the usability and feasibility of eye-tracking with young children. Notably, the study involved preschoolers that were considerably younger than participants involved in most eye-tracking studies within the educational domain. The sample consisted of forty children from three preschools from south of Sweden of middle-class SES environments. All children were enrolled in the study with no selection being made. According to our experience the children are typical representatives for their age group.

Overall, we found that with some simple tricks to keep the children interested, conducting an eye-tracking experiment on preschoolers is less hard than it sounds. Using an educational game had the advantage of increasing the children’s motivation and willingness to participate. Not the least, there was the off-task activity of watering their (magical) garden, which was a motivational boost for the children and an effective way of maintaining their willingness to participate and play more.

Conducting the calibration behind a cover story – saying that looking at the calibration dots was necessary to unlock and start the game – turned out to be a successful way of making the calibration an enjoyable activity for the children. At least three factors may have contributed to the unexpectedly small amounts of data losses: the children were relaxed and comfortable since they had learnt to know the experimenter; the mouse was controlled by the researcher; the pace of the game is not as fast as in many other computer games for children.

In sum, we have showed the methodological possibilities for expanding this kind of research even to children as young as 4- to 6-years-old.

Limitations and Concerns

Fixations Vs. Transitions

The study only made use of fixations (AOI hits) as gaze measurements. Potentially, transitions between AOIs could have revealed things that we could/did not access in this study. For example, one could attempt to identify and analyze other AOI hits preceding an AOI hit on the 'correct branch' by adding all branches and other visual elements on the screen as AOIs. However, the amount and complexity of data in such an approach would have been quite large and by far exceed our resources, and it is not certain that we would have had better answers to our research questions.

Fixed Order of Trials

A potential concern is the fact that the order of trials was always the same: standard trial with child in charge; manipulated trial with child in charge; standard trial with TA in charge; manipulated trial with TA in charge. Could the experiment be confounded if a child perceives this pattern and knows when to expect a broken elevator? We do not think this is a major concern. First, the pattern – easily grasped by an adult scientist – is not necessarily a salient pattern for a preschooler. The scenes in the game are visually rich and dynamic. Between trials there is much that happens, not the least in visual terms: there's a novel baby bird, a novel number of feathers, novel colors, and there is motion and talking on the part of characters. After a round (of four trials) children get to water their garden where a butterfly may fly around or maybe an elephant visits and helps watering the flowers. Second, even *if* a child would detect a repetitive pattern of a broken vs. non-broken elevator, it is hard to see that this could prime their gaze behaviors in ways that would confound the study's measurements. For example, a look-back on the correct branch (passed by the broken elevator) requires knowledge of which branch is correct. For a child who does not know which branch is correct, being able to predict an elevator break-down will not help her look back at the correct branch, that is, the risk for false positives is not increased. For the 'look at TA after malfunction' event, noticing the recurring order of trials is not likely to make children look more at the TA after malfunction with TA in charge than look at the TA after malfunction with the child in charge. Similar reasoning applies to the other gaze behaviors: 'look at choice' and 'look ahead'.

Children with Partly Developed Number Sense

The design of the manipulated sub-game together with the construction of measurements sets limitations on the granularity of what could be learned from the study and not. The malfunction event only occurred when a child (or TA) had chosen a button that corresponded to the number of feathers held up by the baby bird. As a result, there could not be any 'look backs' that, potentially, could distinguish between children who were not capable of connecting any of the number representations to one another and children who were able to select the appropriate button to make the elevator go to a certain branch but not able to connect the bird feathers with that number. (In none of these cases a malfunction event was effectuated.) The latter group of children have a partial understanding and are on their way building their number sense. With a differently designed experimental game this could have been addressed. Potentially the latter group of children could make ‘look backs’ to the branch corresponding to the button they chose – if that branch was passed by a malfunctioning elevator. A more detailed charting out of these kinds of nuances in children’s numbers sense proficiency is thus a possible next step for research.

Exploring children’s Understanding of Agency

To use behavioral measures to try to examine what goes on in subjects’ minds – what they think, how they reason, how they understand something – is challenging, and any behavioral measure is nothing but an indication; a basis for – at best – a qualified guess. In addition, it is hard to determine what is required in terms of understanding or cognitive content for a subject to behave in a certain way.

Our conclusion on to what extent and how the children in the study understood their teachable agent cannot be other than tentative. It is also important to point out that the conclusion we draw refers to the group level, saying that *at least some children in this group* of preschoolers have an understanding of their TA as a goal-directed entity with knowledge that it acts upon that is different from the child’s knowledge. Otherwise, it is difficult to explain the large difference at group level between the two modes in the ‘look at TA after malfunction’ gaze behavior.

However, we are not claiming that an individual child who never looked at her TA after malfunction lacks an understanding of TA agency. Some children tended to look at the experimenter after malfunction – we cannot say whether these children have or lack an understanding of TA agency. As for the single gaze metric we used – ‘look at TA after malfunction’) – there is certainly a need for complementary metrics. We hope that future studies will explore other gaze metrics to assess understanding of agency in this or similar context, such as transitions of gaze between relevant AOIs, or length of fixations, that has been proposed to correspond to the depth of information processing (Schwonke et al. 2017). Fixation length was measured, together with rate of fixations, in a study by Lallé et al. (2017) where school children were using an intelligent tutoring system including pedagogical agents that scaffolded the children in various ways.

Eye-Tracking Measurements Only from a Manipulated Sub-Game

The eye-tracking data collection was not done on the standard game but on a customized version of one of the sub-games in the standard game. By doing this, we increased the likelihood that we would be able to measure gaze behaviors that indicated aspects of understanding in children while they used an educational game.

The experimental set-up with a malfunctioning elevator was powerful for setting up AOIs, well-framed time-limits, and predictions – and we wanted to create favorable circumstances since one of our primary goals was to produce a methodological proof of concept that potentially can catch aspects of understanding via measurements of gaze behaviors during real-time real-world game-play.

In addition, since a malfunction is unexpected only for those who have understood the underlying number sense mechanism that is malfunctioning, responses to such events are indications of cognitive processing and understanding. The rational is that the likelihood for some gaze behaviors is very low if not driven by a cognitive, interpretative top-down process. The custom designed version of the sub-game Bird Rescue was in essence a design of a situation where the likelihood for someone to attend to certain areas of information at particular stages of game-play is heavily dependent on their understanding of the context. (Still, it should be pointed out, also the manipulated version had an unfolding narrative, with rich visual sceneries and visual dynamics partly unfolding depending on the child's own decisions.)

Future Research and Applications

In view of the study as a methodological proof of concept it is important to say something on potentials for future research, a topic already touched upon earlier in the text.

It is hardly daring to foresee that combinations of log data with eye-tracking data will lead to more robust and richer AI-models, not the least with the rapid development of embedded eye-tracking technology (in tablets and laptops) that will open for both larger and smoother studies as well as applications. This may enable better predictions of performance and more insight into differences in underlying processes and aspects of understanding, on which to base adaptive systems and other kinds of personalized behaviors of systems and agents.

We hold that eye-tracking set-ups of the kind used in this paper with a customized game exploiting a malfunction of some sort, can be promising for a variety of knowledge domains for those interested in cognition and understanding. In that context also complementing with systematic analyses of verbal responses could be interesting. As an example, verbal responses from the participating children in our study, such as: "There is something wrong!"; "No-no-no, wrong!"; and "It broke again!" were frequent in connection to the malfunctioning elevator. This data could also potentially shed additional light on how participants understand a TA. Compare Ogan et al. (2012) who studied 12- to 15-year-olds, and Pareto et al. (2011) who studied 8- to 9-year-olds and analyzed how students sometimes blamed their TA for non-success in a math game to gain knowledge about the students' understanding of their TA.

Even for more practical purposes, such as diagnosis of learning difficulties and feeding data to adaptive software and to user models, we see room for experimental tasks created to incite anticipation, surprise, and so on. These tasks would not have to be integral parts of an educational game as such, but could be added as diagnostic checkpoints to go through at certain points during game-play. Such measurements could also be feed into software together with data logging (not the least to be used as triangulation).

Finally, with respect to early math, this study should only be seen as a methodological proof of concept. Future studies may accomplish a much more fine-grained charting-out of what is going on when young children develop their number sense. This could be done by constructing a variety of experimental designs with well-chosen AOIs and combining eye-tracking with data logs. Also, more studies should investigate the possibility that aspects of gaze can be indicators of the processes when understanding is underway, that is of ongoing learning.

Acknowledgements This research was funded by the Knut and Alice Wallenberg Foundation. The authors want to thank Professor Daniel Schwartz, Stanford University, and Dr. Richard Andersson, Lund University, for their contributions to the study and its design, and all children and preschool teachers who participated in the study. We also want to express our sincere thanks to the reviewers of the paper.

Funding Information Open access funding provided by Lund University.

References

Altrichter, H., Posch, P., & Somekh, B. (1996). *Teachers investigate their work: An introduction to the methods of action research*. London: Routledge.

Axelsson, A., Andersson, R., & Gulz, A. (2016). Scaffolding executive function capabilities via play-&-learn software for preschoolers. *Journal of Educational Psychology, 108*(7), 969.

Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. *Journal of Educational Psychology, 72*, 593–604.

Benfatto, M. N., Seimyr, G. Ö., Ygge, J., Pansell, T., Rydberg, A., & Jacobson, C. (2016). Screening for dyslexia using eye tracking during reading. *PLoS One, 11*(12), e0165508.

Biswas, G., Katzlberger, T., Bransford, J., Schwartz, D., & the teachable agents Group at Vanderbilt (2001). Extending intelligent learning environments with teachable agents to enhance learning. In *Artificial Intelligence in Education* (pp. 389-397). IOS Press.

Bolden, D., Barmby, P., Raine, S., & Gardner, M. (2015). How young children view mathematical representations: A study using eye-tracking technology. *Educational Research, 57*(1), 59–79.

Blair, K., Schwartz, D., Biswas, G., & Leelawong, K. (2006). Pedagogical agents for learning by teaching: Teachable agents. *Educational Technology, 47*(1), 56–61.

Brophy, S., Biswas, G., Katzlberger, T., Bransford, J., & Schwartz, D. (1999). Teachable agents: Combining insights from learning theory and computer science. In S. P. Lajoie & M. Vivet (Eds.), *Proceedings of the 9th international conference on artificial intelligence in education, AIED 1999* (pp. 21–28). Amsterdam: IOS Press.

Chase, C. C., Chin, D. B., Oppezzo, M. A., & Schwartz, D. L. (2009). Teachable agents and the protégé effect: In-creasing the effort towards learning. *Journal of Science Education and Technology, 18*(4), 334–352.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.

Conati, C., Jaques, N., & Mary Muir, M. (2013). Understanding attention to adaptive hints in educational games: An eye-tracking study. *International Journal of Artificial Intelligence in Education, 23*, 136–161.

Corbetta, M., Akbudak, E., Conturo, T., Snyder, A., Ollinger, J., Drury, H., Linenweber, M. R., Petersen, S. E., Raichle, M. E., van Essen, D., & Shulman, G. (1998). A common network of functional areas for attention and eye movements. *Neuron, 21*(4), 761–773.

Corbetta, M., & Shulman, G. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience, 3*(3), 201–215.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*(12), 1827–1837.

Evans, M. A., & Saint-Aubin, J. (2005). What children are looking at during shared storybook reading evidence from eye movement monitoring. *Psychological Science, 16*(11), 913–920.

Gelman, R., & Meck, E. (1983). Preschoolers' counting: Principles before skill. *Cognition, 13*, 343–359.

Gerholm, T., Hörberg, T., Tonér, S., Kallioinen, P., Frankenberg, S., Kjällander, S., et al. (2018). A protocol for a three-arm cluster randomized controlled superiority trial investigating the effects of two pedagogical methodologies in Swedish preschool settings on language and communication, executive functions, auditive selective attention, socioemotional skills and early maths skills. *BMC Psychology, 6*(29), 1–25.

Ginsburg, H., Jamalian, A., & Creighan, S. (2013). Cognitive guidelines for the design and evaluation of early mathematics software: The example of MathemAntics. In L. English & J. Mulligan (Eds.), *Advances in mathematics education: Reconceptualising early mathematics learning* (pp. 83–120). Dordrecht: Springer.

Gluck, K. A., Anderson, J. R., & Douglass, S. A. (2000, June). Broader bandwidth in student modeling: What if ITS Were "Eye" TS?. In International Conference on Intelligent Tutoring Systems (pp. 504-513). Berlin: Springer,

Granka, L., Feusner, M., & Lorigo, L. (2008). Eye monitoring in online search. In Passive eye monitoring (pp. 347–372). Berlin/Heidelberg, Germany: Springer.

Griffin, S., & Case, R. (1997). Re-thinking the primary school math curriculum: An approach based on cognitive science. *Issues in Education, 3*(1), 1–49.

Griffin, S. A., Case, R., & Siegler, R. S. (1994). Rightstart: Providing the central conceptual prerequisites for first formal learning of arithmetic to students at risk for school failure. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 25–49). Cambridge: MIT Press.

Griffin, S. (2004). Building number sense with number worlds: A mathematics program for young children. *Early Childhood Research Quarterly, 19*(1), 173–180.

Gulz, A., Haake, M., & Silvervarg, A. (2011). Extending a teachable agent with a social conversation module –effects on student experiences and learning. In International conference on artificial intelligence in education (pp. 106–114). Berlin/Heidelberg, Germany: Springer.

Gulz, A., & Haake, M. (2019). Can preschoolers learn by teaching a digital character? (in Swedish). In B. Riddersporre & S. Kjällander (Eds.), *Digitalization in preschool on a scientific basis*. Natur och Kultur: Stockholm.

Haake, M., Axelsson, A., Clausen-Bruun, M., & Gulz, A. (2015a). Scaffolding mentalizing via a play-&-learn game for preschoolers. *Computers & Education, 90*, 13–23.

Haake, M., Husain, L., Anderberg, E., & Gulz, A. (2015b). No child behind nor singled out? – Adaptive instruction combined with inclusive pedagogy in early math software. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *LNAI/LNCS: Vol. 9112. Proc. of AIED 2015* (pp. 612–615). Berlin/ Heidelberg: Springer.

Hannula, M. M., Räsänen, P., & Lehtinen, E. (2007). Development of counting skills: Role of spontaneous focusing on numerosity and subitizing-based enumeration. *Mathematical Thinking and Learning, 9*(1), 51–57.

Hannula, M. M., Lepola, J., & Lehtinen, E. (2010). Spontaneous focusing on numerosity as a domain-specific predictor of arithmetical skills. *Journal of Experimental Child Psychology, 107*(4), 394–406.

Heine, A., Thaler, V., Tamm, S., Hawelka, S., Schneider, M., Torbeyns, J., et al. (2010). What the eyes already 'know': Using eye movement measurement to tap into children's implicit numerical magnitude representations. *Infant Child Development, 19*, 175–186.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7*(11), 498–504.

Hintz, F., & Meyer, A. S. (2015). Prediction and production of simple mathematical equations: Evidence from visual world eye-tracking. *PLoS One, 10*(7), e0130766.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Husain, L., Gulz, A., & Haake, M. (2015). Supporting early math – Rationales and requirements for high quality software. *Journal of Computers in Mathematics and Science Teaching, 34*(4), 409–429.

Jordan, N., Kaplan, D., Ramineni, C., & Locuniak, M. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology, 45*, 850–867.

Jones, M. W., Obregón, M., Kelly, M. L., & Branigan, H. P. (2008). Elucidating the component processes involved in dyslexic and non-dyslexic reading fluency: An eye-tracking study. *Cognition, 109*(3), 389–407.

Lallé, S., Taub, M., Mudrick, N. V., Conati, C., & Azevedo, R. (2017). The impact of student individual differences and visual attention to pedagogical agents during learning with MetaTutor. In International conference on artificial intelligence in education (pp. 149–161). Cham: Springer.

Lindström, P., Gulz, A., Haake, M., & Sjödén, B. (2011). Matching and mismatching between the pedagogical design principles of a math game and the actual practices of play. *Journal of Computer Assisted Learning, 27*(1), 90–102.

Lincoln, Y., & Guba, E. (1985). Establishing trustworthiness. *Naturalistic inquiry, 289*, 331.

Louwerse, M. M., Graesser, A. C., McNamara, D. S., & Lu, S. (2008). Embodied conversational agents as conversational partners. *Applied Cognitive Psychology, 23*(9), 1244–1255.

Ogan, A., Finkelstein, S., Mayfield, E., D'Adamo, C., Matsuda, N., & Cassell, J. (2012). Oh dear Stacy! Social interaction, elaboration, and learning with teachable agents. In R. Grinter et al. (Eds.), *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 39–48). New York: ACM.

Pareto, L., Arvemo, T., Dahl, Y., Haake, M., & Gulz, A. (2011, January). A teachable-agent arithmetic game's ef-fects on mathematics understanding, attitude and self-efficacy. In International conference of artificial intelligence in education (pp. 247–255). Berlin/Heidelberg, Germany: Springer.

Pareto, L., Haake, M., Lindström, P., Sjödén, B., & Gulz, A. (2012). A teachable-agent-based game affording collaboration and competition: Evaluating math comprehension and motivation. *Educational Technology Research and Development, 60*(5), 723–751.

Perner, J. (1991). *Understanding the representational mind*. Cambridge: MIT Press.

Perner, J., & Roessler. (2012). From infants' to children's appreciation of belief. *Trends in Cognitive Sciences, 16*, 519–525.

Pezzulo, G., Butz, M. V., Castelfranchi, C., & Falcone, R. (Eds.). (2008). *The challenge of anticipation: A unifying framework for the analysis and design of artificial cognitive systems*. Berlin/Heidelberg: Springer.

Rau, A., Moll, K., Moeller, K., Huber, S., Snowling, M., & Landerl, K. (2016). Same same, but different: Word and sentence reading in German and English. *Scientific Studies of Reading, 20*(3), 203–219.

Rötting, M. (2001). *Parametersystematik der Augen-und Blickbewegungen für arbeitswissenschaftliche Untersuchungen*. Düren: Shaker Verlag GmbH.

R Core Team (2017). R: A language and environment for statistical computing [Computer Software]. Vienna: R Foundation for Statistical Computing.

Schneider, M., Heine, A., Thaler, V., Torbeyns, J., De Smedt, B., Verschaffel, L., et al. (2008). A va-lidation of eye movements as a measure of elementary school children's developing number sense. *Cognitive Development, 23*(3), 409–422.

Schwonke, R., Renkl, A., & Berthold, K. (2017). Knowledge construction with multiple external representations: What eye movements can tell us, *Proceedings of the European Cognitive Science Conference* 2017.

Smith, B. (2012). Eye tracking as a measure of noticing: A study of explicit recasts in SCMC. *Language Learning & Technology, 16*(3), 53–81.

Vygotskij, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge: Harvard University Press.

Vaish, A., Hepach, R., & Grossmann, T. (2018). Desire understanding in 2-year-old children: An eye-tracking study. *Infant Behavior and Development, 52*, 22–31.

Wagster, J., Tan, J., Wu, Y., Biswas, G., & Schwartz, D. (2007). Do learning by teaching environments with metacognitive support help students develop better learning behaviors. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th annual meeting of the cognitive science society* (pp. 695–700). Austin: Cognitive Science Society.

Wellman, H., & Liu, D. (2004). Scaling of theory of mind tasks. *Child Development, 75*, 523–541.

Wikholm, A., Önnered, A., Gulmann, C. M., Egli, S., & Wipp Ekman, V. (2019). Real-time adjustable feedback based on eye tracking algorithms in educational games. In Haake, M., Gulz, a., Balkenius, C., Wallergård, M. (Eds.). (2019). Intelligent, socially oriented technology IV. *LUCS, 175*. Retrieved from https://www.lucs.lu.se/LUCS/175/LUCS_175.pdf

Affiliations

Agneta Gulz[1,2] **· Ludvig Londos**[1] **· Magnus Haake**[1]

Ludvig Londos
laddy_ludde@msn.com

Magnus Haake
magnus.haake@lucs.lu.se

[1] Division of Cognitive Science, Lund University, Helgonavägen 3, 223 62 Lund, Sweden

[2] Department of Computer and Information Science, Linköping University, Mäster Mattias väg, Campus Valla, 581 83 Linköping, Sweden