Int J Artif Intell Educ (2014) 24:251–283
DOI 10.1007/s40593-014-0018-8

RESEARCH ARTICLE

A Teachable Agent Game Engaging Primary School Children to Learn Arithmetic Concepts and Reasoning

Lena Pareto

Published online: 28 May 2014

Abstract In this paper we will describe a learning environment designed to foster conceptual understanding and reasoning in mathematics among younger school children. The learning environment consists of 48 2-player game variants based on a graphical model of arithmetic where the mathematical content is intrinsically interwoven with the game idea. The environment also features teachable agents, which are computer programs that can be taught and behave according to their knowledge. Thus, the environment provides both learning-by-doing (playing the game) and learning-by-teaching (teaching the agent to play). It differs from other learning-by-teaching systems 1) by targeting basic mathematics and primary grade students; 2) by using teachable agents as an extension to educational games in order to leverage engagement, reflection and learning; and 3) by using an agent-driven question dialogue to challenge students' mathematical thinking, to role-model learner behaviour and to transfer game knowledge to out-of-game mathematics. The teachable agent game is described and evaluated in an authentic classroom study enrolling 443 students from 22 classes in 9 schools. Students range from 2nd to 6th grade of mainstream classes and 7th to 8th grade for students with difficulties in mathematics. Part of the study was designed as a quasi-experimental study with controls; part was designed to examine students' change in mental models of arithmetic before and after game play. All students took pre- and post mathematics tests. The 314 playing students used the game and taught their agents during regular math-classes for three months, whereas the control classes attended standard instruction and took the tests. A questionnaire was distributed at the end of the study to investigate students' perceptions and performances of the agent-tutoring task. Results show that 1) there is a significant learning gain for playing students compared to controls, 2) the learning environment can engage children in advanced mathematical thinking in early education, 3) young primary students can act as successful tutors. Thus, we conclude that teachable agents in educational games can help achieve deeper levels of learning that transfer outside the game. This idea combines the motivational power of games with the reflective power of a teachable agent asking though-provoking, deep questions on the learning material during game play.

L. Pareto (✉)
Division of Media & Design, University West, SE-461 86, Trollhättan, Sweden
e-mail: lena.Pareto@HV.SE

Keywords Technology-enhanced learning · Mathematics game · Teachable agent · Arithmetic · Reasoning · Primary school children · Classroom study · Authentic use

Introduction

The question of how to teach mathematics in a way that encourages understanding has been an issue for mathematics educators for a long time. Understanding is characterized by the strength and number of connections a learner makes among important mathematical ideas (Gutstein and Mack 1999). Hence, related concepts should be studied together and not in isolation (Case et al. 1997). Arithmetic concepts, such as understanding that the "2" in 20 means two sets of ten, is critical to all future calculus learning; yet it is a major stumbling block for many primary school children (Carpenter et al. 1993). Causal reasoning is also fundamental to mathematics. According to Jonassen and Ionas (2008), it is one of the most basic and important cognitive processes in developing conceptual understanding and problem solving. The game-based learning environment presented here targets arithmetic understanding and casual reasoning and is designed to be appropriate for young primary school children. How to design a game environment so that in-depth mathematical understanding is fostered is however far from evident according to Moreno and Mayer (2005).

Questioning can be effective for understanding. The research community seems to agree on the importance of questioning for learning, in particular for promoting deep-level reasoning about complex material. Researchers like Papert (1980) and Piaget (1952) have, consequently, long advocated learning environments that support inquiry learning. The problem remains, however, as not all students are yet vigorous question askers (Graesser et al. 2009). In a 10-year longitudinal study in primary school mathematical classrooms, Martino and Maher (1999) showed that posing timely questions that challenge learners' understanding fosters individual cognitive growth in mathematical understanding. Hunter (2008) argues that using questioning in the mathematics classroom allows students to make discoveries about and connections among concepts, and that questioning is more effective than traditional methods of teaching. Learners can be fostered into adopting the practice of questioning (Lave and Wenger 1991) and can learn to think mathematically by participating in mathematical inquiry (Goos 2004). However, Graesser, McNamara and VanLehn (2005) have shown that it is not sufficient to expose learners to good questions; they need to engage with the inquiry. Several conditions can influence a learner's willingness to engage with questions: the learner's level of perplexity, her needs, commitment or courage to challenge her understanding of the issue at hand (Dillon 1986; Van der Meij 1994). One main concern when designing the learning environment has been to engage young students with meaningful mathematical inquiry that supports understanding.

An ideal learner is an active, self-motivated, creative, inquisitive person who asks and searches for answers to deep questions (Otero and Graesser 2001). *Deep questions* (such as why, why not, how, what-if, what-if-not) target difficult scientific material that taps causal structures, whereas shallow questions (who, what, when, where) more often target simple facts (Graesser et al. 2009). Most students are not frequent question askers: an average student in a classroom asks only 0.17 questions per hour, compared to 26 questions per hour in a one-to-one situation (Dillon 1988; Graesser and Person 1994). Moreover, most questions are shallow (Graesser et al. 1996). The ability to ask timely questions is one reason behind tutoring effectiveness (Martino and Maher 1999), and tutoring has shown effective for young students' mathematical understanding (Gutstein and Mack 1999).

Peer tutoring refers to a one-on-one instruction situation, where the tutor typically is untrained, the roles of tutor and tutee are assigned, and the knowledge gap between tutor and tutee can be minimal (Roscoe and Chi 2007). It is not only tutees who learn from peer-tutoring, tutors also gain in understanding (Chi et al. 2001; Roscoe and Chi 2007). Tutors benefit from the instructional task by preparing material and reviewing their own understanding. However, a potentially more powerful knowledge-building opportunity for the tutor comes from trying to respond to tutee questions or tutee confusions due to inadequate, incomplete or contradictory explanations. Tutees' questions can reveal knowledge deficits or create cognitive disequilibria within the tutor's understanding and since tutors normally want to accomplish their task well, a situation of genuine inquiry arises. Such effects provide a foundation for the learning by teaching paradigm (Palthepu et al. 1991) and thus for teachable agents, which are computer agents that can be taught by humans and are described in Brophy, Biswas, Katzlberger, Bransford and Schwartz (1999) for example.

The teachable agent in our game environment is designed to simulate effective peer-tutoring situations. The human tutor teaches an agent tutee to play the learning game. The agent is designed to mimic an ideal learner. The agent actively asks thought-provoking, deep questions that challenge the human tutor to reflect on and explain her game playing actions under the pretext that the tutee agent "wants to know".

Related Work: Learning by Teaching systems and Teachable Agents

Several systems have been developed that use learning-by-teaching in various domains: the Math Concept Learning System (MCLS) for solving linear equation problems (Michie et al. 1989), DENISE for qualitative causal reasoning in Economics (Nichols 1994), STEPs for quantitative problems in physics (Ur and VanLehn 1995), the reciprocal tutoring system of Chan and Chou (1997) for computer programming with recursion in LisP, Betty's brain for concept understanding and qualitative reasoning (Brophy et al. 1999), the virtual classroom for problem solving in psychophysiology (Obayashi, Shimoda, & Yoshikawa 2000), our teachable agent game for arithmetic understanding (Pareto et al. 2009) and SimStudent also for linear equation problem solving (Matsuda et al. 2010).

All these systems involve some kind of virtual learning companion, but their roles vary. There is a distinction between agents that can be taught and agents that can learn: to be *teachable* the agents need to appear as if they are learning from the user, but they do not need to actually 'learn' (Brophy et al. 1999). Likewise, Chan and Chou (1997) argue that to simulate a tutee there are two objectives. The first objective is to model the behaviour of a tutee whose domain knowledge appears to be advancing despite occasional mistakes. As a second objective, the tutee behaviour and performance must be understandable to the human tutor in order for the tutor to monitor the teaching activity. According to Blair, Schwartz, Biswas and Leelawong (2007), teachable agents provide opportunities to optimize learning-by-teaching interactions by adopting the following four principles: 1) use explicit and well-structured visual representations, 2) enable the agent to take independent actions, 3) model productive learner behaviours, and 4) include environments that support teaching interactions. Teachable agent design thus needs to focus on how the teaching activity can be designed to optimize the benefits of teaching for the human tutor and is different to the design of systems that focus on the agents' learning.

The teaching model or the way agents are taught vary in the different systems. In MCLS and SimStudent the tutor provides the agent tutee with problems and example solutions from which the tutee learns. In STEPs the tutor also provides problems; the underlying problem-solver tries to solve these problems and the tutor monitors the problem-solving procedure by providing process directions or hints. The virtual classroom (Obayashi et al. 2000) uses a system that asks tutors questions and then the tutors' answers are mirrored back in a simulated classroom situation. There are several applications based on Betty's Brain (Biswas et al. 2001; Schwartz et al. 2007; Blair et al. 2007), where students teach Betty by drawing and modifying concept maps, and test her knowledge by asking questions or letting her take a quiz to check her knowledge.

If the main purpose of a learning-by-teaching system is to promote tutor learning then the effects of learning in this way should be studied. To our knowledge, only Betty's brain, our TA-game and SimStudent have gone through real-use trials in classroom situations with enough students to verify learning gains. Some other systems have shown potential for learning through the use of pilot studies that enrol too few subjects to establish convincing evidence of learning (Michie et al. 1989; Ur and VanLehn 1995; Chan and Chou 1997; Obayashi et al. 2000). Betty's Brain has been subjected to a large number of studies, which reveal that students using Betty's Brain developed more integrated knowledge of the material than those who used the same system without teaching the agent Betty (Biswas et al. 2005; Schwartz et al. 2007; Schwartz et al. 2009). We have previously conducted a nine-week authentic quasi-experimental study comparing learning effects with and without our game environment that enrolled 153 students in 3rd and 5th grade. The results show significant learning gains for the playing group compared to the control group in mathematical achievement and self-efficacy, but not for attitude (Pareto et al. 2011). A more detailed analysis of the 3rd grade students' results showed that the game playing students' higher gains mainly derived from problems concerning conceptual understanding of the base-10 system (Pareto et al. 2012). The version of SimStudent that enables learning-by-teaching (Matsuda et al. 2010) was used in a three-day classroom experimental study with 160 high school students (Matsuda et al. 2012). The study compared two versions of the system: with and without a self-explanation module. Results show that the self-explanation group worked on significantly fewer problems to achieve the same learning compared to the group without, indicating that self-explanation can be effective for learning. However, the study showed no evidence of students' gaining conceptual knowledge and only weak trends of students gaining in required skills. The intervention may have been too short to show significant learning effects.

Our Approach

Our objective is three-fold: 1) to engage younger students in mathematical discussions and reasoning, 2) to scaffold students' learning of conceptual understanding of integers and arithmetic operations, and 3) to make the experience enjoyable. Our approach is a digital learning game environment where students play two-player games and learn basic arithmetic through playing. The mathematical learning content is intrinsically interwoven with the game design: it is deliberately designed as a "side-effect" of the game play and not as an explicit goal of the game playing activity. In order to play well the player must make good choices in the card game, which in turn involves predicting and performing mental calculations as well as reasoning about numbers and computations, but this is not an explicitly-stated objective. Therefore, a key issue in designing this learning environment is to motivate the students to challenge themselves and to help them acquire knowledge to play well. For this purpose we have extended the environment with teachable agents (Pareto 2009; Pareto et al. 2009).

Players can decide for each game if they want to play themselves or teach their agent to play the game. Game play consists of making (good) choices and therefore students should teach their agents how to choose. However, to formulate a general explanation of what constitutes good types of choices in the games would be very difficult for primary school students, so a teaching scheme based on direct instruction is inappropriate. Instead, we have adopted a cognitive-apprentice model (Collins et al. 1989) where the student tutor has the option to 1) show how to play while the agent-apprentice observes and learns, or 2) let the agent-apprentice try playing while providing corrective feedback, or 3) let the agent play by itself. In apprenticeship learning, scaffolding refers to the assistance offered to the apprentice where the expert completes those parts of the task that the apprentice has not yet mastered. Scaffolding should be temporary and gradually removed when apprentices learn to handle more of the task on their own (Chan and Chou 1997). These three steps correspond to such withdrawal of support from the tutor: in 1) the tutor plays and the tutee observes; in 2) the agent tutee plays under surveiyence of the tutor, and in 3) the agent tutee plays and the tutor observes without possibility to interfere.

When being taught, the agent will ask the student tutor situation-specific questions to explain and justify the tutor's action. For example, the agent can ask the student to justify a particular choice, to compare their respective choices when in disagreement, or to justify a disapproval of a previous choice. Hence, the agent acts as "an ideal learner" and poses numerous deep, explanatory questions prompting the student-tutor to reflect on and explain her playing behaviour. Recently, Matsuda et al. (2012) also added explanatory questions to their TA to prompt for self-explanation. The purpose of the teachable agent dialogue is three-fold; to challenge students to reflect on why they play as they do; to support transfer of tacit game playing knowledge to articulate language; and to provide a role model of questioning to imitate.

Our teachable agent game differs from the abovementioned learning environments in the following aspects:

1) *Learning content*: Our game targets basic arithmetic understanding such as the base-10 number system, whereas other learning-by-teaching systems target more

advanced content such as linear equation solving, recursive programming and causal reasoning.

2) *Student age:* Consequently, our user group consists of younger students grade 1–6. Age plays a major role for children's ability to perform cognitive tasks, as evident from a developmental review by Zimmerman (2007). Examples of abilities that changed during primary school age concerned predicting from hypotheses (Ruffman et al. 1993), testing ideas through experimentation (Penner and Klahr 1996), learning from new, contradictory evidence (Chinn and Malhotra 2002), and verifying ideas (Bullock and Ziegler 1999). These abilities are involved in productive tutoring, so it is important to examine whether children of such a young age can manage to tutor the agent.
3) *TA-enhanced game:* We have extended an existing game environment with teachable agents in order to i) improve the games' learning effect, ii) motivate students to higher engagement and elaboration levels, and iii) facilitate the connection between the game world and the learning content. This role of a TA as a reflective and content-focused game-playing partner is novel and may be useful for other learning games.
4) *The TA questioning system*. The way the TA is designed as an inquisitive, active learner that takes the initiative and challenges the student to explain and justify her actions and thinking. Roscoe and Chi (2007) showed that successful learners use self-explanation as a learning strategy. Most students do not spontaneously self-explain but they do so when prompted and can learn to self-explain effectively (Mitrovic 2005). The level of difficulty in the agent's questions follows the game performance level of the human tutor to challenge the student without being de-motivating.

Research Questions

In this paper we will present our learning environment with a focus on the teachable agent questioning system, which has not been described in previous publications. The learning environment has been used and studied in authentic classroom situations for a decade. The environment is web based and currently over 60 Swedish schools are registered in the user database. Here, we will report empirical findings from a study enrolling 443 students ranging from grade 2−6 in mainstream classes and grade 7−8 in special needs classes. Part of the study was designed as a quasi-experimental study with controls; part was designed to examine students' change in mental models of arithmetic before and after game play. There were 314 game-playing students and 129 students in control classes in grade 2, 4 and 6. The experimental part is a follow-up study from a previous experiment with grades 1, 3, and 5. Here we will present findings from the experiment as well as findings of the agent-tutoring activity based on the 314 game-playing students. We address the following research questions:

1) Are there differences in learning effects in mathematical achievement for the playing classes compared to the control classes in the grades 2, 4 and 6?

2) How do students perceive the agent questioning system with respect to a) *importance*, b) *difficulty*, c) *role division: tutor-tutee* and d) *enjoyment*? Are there differences in perception related to grade level?
3) How do students perform with respect to answering the agent's questions? Are there differences in performance related to grade level?
4) How do students' in-game performance and mathematics performance (as revealed in paper and pencil math tests) relate to each other?

The paper is organized as follows: First the learning environment is described with a focus on the teachable agent design and the questioning system. Then the study design is explained, followed by study results, discussion and conclusion.

The Learning Environment

Educational games have well-documented effects on general learning and motivation (Rieber 1996; Fisch 2005; Vogel et al. 2006; Ke 2008; Schwartz and Arena 2009), but games' effect on specific skills and competencies depend on how successful a game design is to stimulate such learning (Moreno and Mayer 2005). Educational content must be represented in appropriate form, be closely intertwined with game play, and feedback and hint structures must be provided to scaffold children as they deal with challenging content (Fisch 2005).

The Mathematics Games

All games in the environment are 2-player digital card- and board games (Pareto 2004). The games can be played collaboratively or competitively. Players take turns choosing a card from their respective hand and put it on a common game board. Cards and the game board represent positive or negative integers, which are modelled by groups of coloured squares. Each player performs an arithmetic operation, which is modelled by animated operations on such graphical numbers. For example moving squares from a card to the game board represents addition, the opposite direction subtraction. A game proceeds as follows: each player receives a hand of 10 cards. Part of the hand (normally 4) is face-up and visible to both players whereas the remaining cards are facedown and unknown to both players. The players take turns to select a card from their respective hands until all cards are played. The selected card will act on the current game board resulting in a modified board, and this constitutes a computation.

For example, if a player performs addition and chooses a card with a value of 3 to add to a board with a value of 5, 3 squares will be added to the board resulting in 8. If the player instead performs subtraction, 3 squares will be removed resulting in 2. One turn for each player constitutes a round, 10 rounds make a game. Each turn (in turn-based games) or each round (in round-based games) may award points to the player who acted (if players compete) or to both players if they collaborate. The objective is always to maximize points, either competitively or collaboratively. Each of the game variants has a different goal awarding points, but all goals are based on the last played card or the new state on the game board after playing the card, so each choice has the potential to award points. The challenge is thus to choose a good, preferably the best, card at each turn. What constitutes a good choice depends on the goal of the game, the current game board and both players' cards. This way, making good choices involves predicting arithmetic computations and making the best choice involves reasoning about current alternatives as well as future game states. McNay and Melville (1993) showed that children in grades 1–6 are both aware of what predicting means, and are able to generate predictions for a number of domains.

The game environment currently offers 4 categories of games: *Find Pair, Pack Many, Remove All* and *Divide*. There are 48 game variants which are combinations of operations addition/subtraction or multiplication/division and number ranges from 0… 10 up to 0…1000 for games with only positive integers, and numbers ranges −10…10 up to −1000…1000 for games with both positive and negative integers.

The Find-Pair category is the easiest type of game, a collaborative round-based game where the players' challenge is to find a sum (addition-addition), a difference as in Fig. 1 (addition-subtraction), or a subtraction of the goal value (subtraction-subtraction). At least one such pair is guaranteed to exist per round. In Fig. 1 we have the difference 96−23 = 73, which matches the goal. The graphical number 96 is represented by 9 orange and 6 red squares (see corresponding card on the screenshot). The cards of player Y playing subtraction illustrates "positions to be filled with squares", so the card 23 on the top has 2 places for orange squares and 3 places or red squares. The game goal 73 is illustrated on the game board in the middle by marking 7 positions in the turquoise compartment for tens and 2 positions in the blue compartment for ones. Nine squares are encircled in each compartment to illustrate that there is only "room for 9 squares" before packing is needed. The players receive a point for each pair they find, so the maximum score is 10 points. The players need to collaborate and discuss choices before playing since if a played card has no matching card from the other hand, the round is lost without points.

The Pack-Many category comprises turn-based games that can be played both collaboratively or competitively. All turn-based games are strategic since the result of a turn affects the rest of the game, as opposed to round-based games where each new round is essentially a new start. The goal is to pack as many square-boxes as possible for addition, or unpack for subtraction. Both players play the same operation. Fig. 2 illustrates such a game.

It is player Y's turn to choose a card from the alternatives 2, 9, 4 or 7 that preferably makes the game board "pack" (i.e., results in 10 or more squares). The player receives one point for each packed square box during the turn. Since there are already 3 squares on the game board and the players compete, the choice 7 is optimal. It is better than 9 (which also would yield a point since 3+9>10), but the 9 will leave 2 red squares on the game board which allows the opponent to score on the next turn if 8 is chosen. The

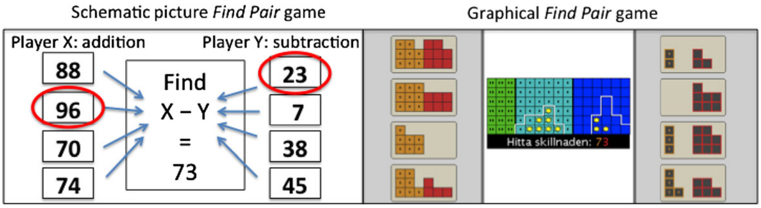


Fig. 1 A game from the collaborative Find-Pair game category, schematic view and screenshot

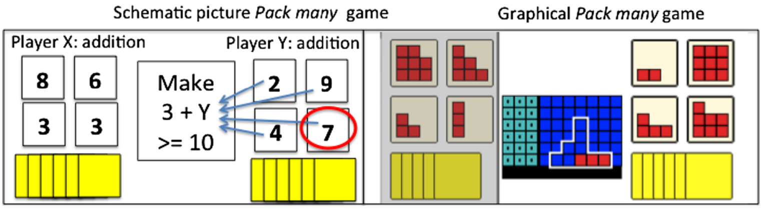

Schematic picture *Pack many* game
Graphical *Pack many* game
Player X: addition
Player Y: addition
8
6
3
3
Make
3 + Y
>= 10
2
9
4
7

Fig. 2 A competitive game in the Pack-Many game category, schematic view and screenshot.=

cards 2 or 4 would not score, but are possible to play. The best choice, 3+7=10, leaves no red squares which makes it impossible to score during the next turn. Thinking further ahead, player X's smallest card is 3, which guarantees that player Y can score next turn as well since there is a 9. In this way, the players can reason about which cards will score and which card is best, what the options are for the opponent given a particular choice, or, for the more advanced, what happens two or more steps ahead. The teachable agent's questioning system poses questions to try to scaffold such a progression towards more and more advanced reasoning.

The player's task is to choose which card to play, and then the system performs the corresponding computation, as illustrated in Fig. 3.

In this scenario, the left player has added 6 to the game board and it is the right player's turn (Fig. 3A). The right player has the cards 5, 7, 3; and 4 to choose from, and chooses card 5, which is then placed above the centrally located game board (Fig. 3B). Adding 4 squares to 6 triggers a packing operation and the 10 squares are packed into an orange "square-box" which is then placed in the compartment one position to the left, just like the decimal positioning system. Such explicit packing provides children with a visual key to understand that there "exist" red squares within the orange "box", it is just that they are not visible until the box is unpacked. Since the goal of this game is to "pack square-boxes", the right player is awarded a point indicated by a flashing star (Fig. 3C). The remaining square on the card is added to the board completing the computation 6+5=11, and the turn is shifted to the player on the left (Fig. 3D).

All parts of an operation are visual and animated by the system, including packing and unpacking of squares into and out of square boxes that correspond to the carrying and borrowing operations, respectively. The metaphor of the base-10 system is that 10 squares can be packed or unpacked into small ten-square-boxes and 10 ten-square-boxes can be packed into a taller "100-box". However, square-boxes are viewed from above so they look just like squares, to mimic the visual similarity of the two 4's in 446. Yet, each time a carrying operation occurs, the packing is made explicit: the students see 10 red squares transformed to 1 orange square-box.. Such low-stress algorithms have been shown to help students with mathematical difficulties (Lock 1996), and graphical representations can help them to visualize complex objects and grasp their properties (Avigad 2008).

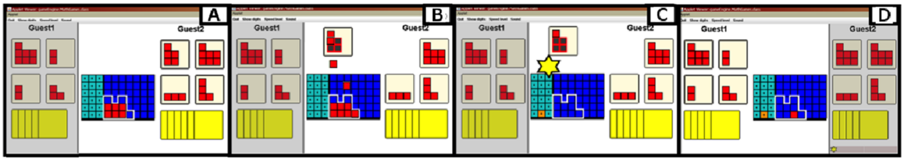


Fig. 3 One scoring turn corresponding to the computation 6 + 5 = 11 in a Pack-Many game

The Remove-All category also consists of turn-based games that can be played collaboratively or competitively, and it involves negative integers as well as positive ones. The graphical model conveys, and the games practice, the idea that negative numbers are the opposite of positive numbers, just as subtraction is the reverse operation to addition. The last category, Divide-Even, consists of collaboratively or competitively played round-based games, where one player performs multiplication and the other division. These games are designed to convey the connection between area and multiplication by automatically rearranging the squares in a rectangular area corresponding to the factors of the product. These games also convey the idea of multiplication as repeated addition, division as repeated subtraction and multiplication and division as opposite operations.

Mathematical Learning Content of the Game Environment

The aim of the learning content is primarily to understand the structure of the base-10 positioning system, how the arithmetic operations are connected to each other, and how integers behave under these arithmetic operations. A graphical number shares the structural properties of a symbolic number through the concept of digits in a sequence and the meaning of the position. The graphical representation, however, emphasizes the position by colour and by making the transitions between positions explicit, and digits have concrete representations and can be pointed at and counted. The graphic numbers provides children with visual cues but, more importantly, with more familiar vocabulary to use in mathematical discussions. The graphical model enforces all mathematical rules so that all possible actions represent valid computations. Such simulation can help us understand the behaviour of dynamical systems (Avigad 2008). This way, properties and relations can be explored and discovered by learners without jeopardizing mathematical validity.

Another part of playing the game well is thinking strategically and inventing computational strategies, for example clever strategies to find the matching pair or strategies to decide if packing will occur without doing more calculations than necessary. Student invention as learning strategy was examined in multiple studies using the Betty's Brain system (Schwartz and Bransford 1998; Schwartz and Martin 2004) The studies showed that students' learned in a deeper and more lasting way and were better prepared for future learning compared to tell-and-practice instruction. We anticipate that our system will have a similar effect. The games provide ample opportunities to reason, to predict computations, and to make choices. According to Jonassen and Ionas (2008), empirical research on instructional methods for supporting causal reasoning is scarce. By contrast, in our games, students constantly practice making choices; making good choices is the only way to perform well in the game. Schwartz and Arena (2009) argue that choice-making will be an important skill in the 21st century and that it should be practiced and assessed in education. This is the way mathematical reasoning is promoted.

Teaching the Teachable Agent

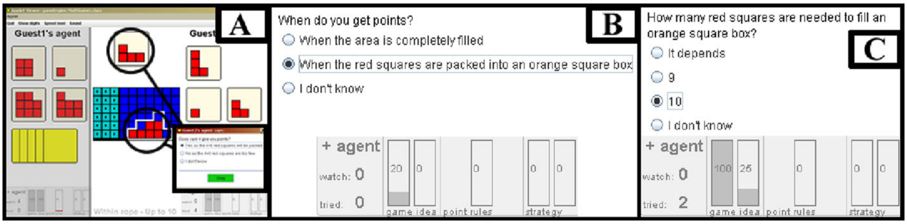
Besides playing the games themselves, students can teach an agent how to play or watch the agent play. The teaching is done while playing, so that the student and the agent "work together" as one of the players. The other player in the game can be a student, the computer, or a student together with his or her agent. There are two ways to teach: by showing the agent how to play (show-mode) and by letting the agent try and then accept or reject its choice (try-mode). In try-mode the agent selects a card according to its knowledge and the card is moved above the game board, but the tutor has the final decision whether the card should be chosen or not (a go and a stop sign appear on the screen). If the student acknowledges the choice of the agent, the card is played. If the student rejects the choice, the card is moved back to the player's hand and the tutor must provide an alternative card. In either case, the agent may pose a question afterwards.

The agent learns in two ways: 1) by observing the tutor (i.e., the student) play, and 2) by asking the tutor questions. The agent can also play without without help from its tutor. At the start, the agent has no knowledge and can only "guess" (choose a random card). Subsequently, it plays according to its current knowledge level, which depends on how much and how well it has been taught. How the agent learns is described below. There are three ways for the student tutor to know if the agent is learning:

1. By the type of questions the agent asks; the further the agent progresses in its knowledge, the more difficult the questions it will ask. This is explained below.
2. By explicit knowledge meters illustrating the agent's current knowledge level compared to "full mastery". This is an intuitive way of describing knowledge progression, common in games. See Fig. 4A bottom left and right on the screen and Fig. 4B-C.
3. By how well the agent selects cards when it plays in try- or play-mode.

The agent asks the tutor questions while being taught, at the time when a card is chosen but before the effect of the choice is executed. The questions concern game play; they always relate to the current situation and most often to the card choice (s) just made. The questions are text-based and appear in a dialogue window on the tutor's side of the game area. See Fig. 4A where a question just is posed. The tutor then reads the multiple-choice question and the provided response alternatives including a 'don't-know' option, before selecting the preferred response, and the chosen card's effect on the game board is executed.

The timing of the question is important; it is when the choice is made but the computation not yet performed, to encourage the student tutor to (again) do the mental calculation and reason about the choice compared to the alternatives. It is also a reasonable situation for a tutee to ask why-questions, i.e., just after the tutor declared:


When do you get points?
○ When the area is completely filled
● When the red squares are packed into an orange square box
○ I don't know

How many red squares are needed to fill an orange square box?
○ It depends
○ 9
● 10
○ I don't know

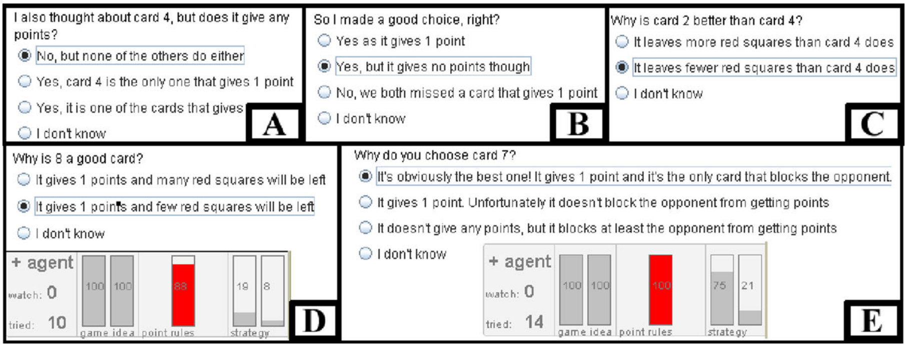
Fig. 4 Game situation when the agent just posed a question (A), agent question example from game idea category (B), and agent question example from graphic model category (C)

"let's do this". Undoing a choice is not an option for the tutor, based on the idea that realizing that mistakes cannot be undone will encourage tutors to *think before* they choose in the future. However, this game feature could be changed. We are also experimenting with question frequency; currently questions are posed every 3rd turn on average, which means about three questions per game. These parameters could be part of a future agent personality profile.

The agent's questions are organized into five categories, reflecting stages in becoming a skillful game player, which coincide with progression in mathematical thinking. The categories are:

1) *Game idea*: understanding what the game is about (see Fig. 4B for an example).
2) *Graphical model*: understanding the graphical model and how it relates to mathematics (see Fig. 4C for an example)
3) *Scoring*: knowing how points are awarded in the game and predicting the outcome of a choice (predicting the calculation and evaluating the result in relation to the game goal).
4) *Basic strategic thinking*: how to choose the best card considering their own cards only which involves predicting each of the four cards' effects and choosing between these (see Fig. 5A-D for examples)
5) *Advanced strategic thinking*: how to choose the best card considering both players' cards and predicting two steps ahead. Normally this means considering and discriminating between 16 alternatives that are two arithmetic computations ahead (see Fig. 5E for an example).

Just as a human learner, the agent asks questions depending on how much it already knows. Questions are chosen to be slightly above the agent's knowledge level to allow progression towards the tutor's level when the tutor knows more. When the tutor's level is reached (that is, when the questions are becoming too difficult for the tutor to answer satisfactorily), the agent still challenges the tutor by asking slightly more advanced questions. If progression stops, however, so will the advancement of questions. This follows the idea of Vygotskij's (2001) zone of proximal development since the question should reflect the tutor's approximate level and the environment prompts the tutor with ideas, namely, the response alternatives..


I also thought about card 4, but does it give any points?
● No, but none of the others do either
○ Yes, card 4 is the only one that gives 1 point
○ Yes, it is one of the cards that gives
○ I don't know

So I made a good choice, right?
○ Yes as it gives 1 point
● Yes, but it gives no points though
○ No, we both missed a card that gives 1 point
○ I don't know

Why is card 2 better than card 4?
○ It leaves more red squares than card 4 does
● It leaves fewer red squares than card 4 does
○ I don't know

Why is 8 a good card?
○ It gives 1 points and many red squares will be left
● It gives 1 points and few red squares will be left
○ I don't know

Why do you choose card 7?
● It's obviously the best one! It gives 1 point and it's the only card that blocks the opponent.
○ It gives 1 point. Unfortunately it doesn't block the opponent from getting points
○ It doesn't give any points, but it blocks at least the opponent from getting points
○ I don't know

Fig. 5 Different examples of strategic agent questions: in show-mode (A), try-mode (B), either mode (C), advanced level (D), most challenging type of question (E)

How the agent learns and its knowledge representation (described below) is not visible to the students. Students can see only how much the agent knows compared to an "expert player", which is illustrated in the agent's knowledge meters. The five knowledge meters correspond to the five categories above. Thus, they provide performance feedback just like Kay's (2008) open learner model. The amount of training that should be required before an agent qualifies as an expert (and, therefore, plays optimally) is a pedagogical issue. It requires balancing the effort and challenge with the motivational power of reaching the expert level. The right balance can only be found by studying authentic long-term use of the game.

The progression of question difficulty according to the agent's knowledge level is illustrated by the questions in Fig. 4 and Fig. 5. The simple game idea question, "*When do you get points?"* (4B) is asked when the agent knows very little and the meters are almost empty. If the tutor selects the correct response "*When the red squares are packed into an orange square box*" then the agent learns this. Simultaneously the system has tested the student tutor's knowledge. Important questions are checked several times with different surface level formulations to ensure that a correct response was more than a lucky guess. If the tutor's answer is "don't know", the question will reappear with a prefix like "*Perhaps you know now?*". This indicates that the agent remembers previous questions. There is a penalty scheme for wrong answers; different formulations of the question will be asked repeatedly until enough correct answers are provided and the agent is finally "convinced". This is how we try to ensure that what the agent knows, the student tutor knows as well.

In the third question category about how to score, all questions are about predicting if a card will score or not. The last two categories consist of strategic questions; examples are illustrated in Fig. 5. Question A is from show-mode, where the tutor has chosen the card 4 and the agent has made a hypothetical choice which was the same as the tutor's: "*I also thought of card 4,…"*. Question B is from try-mode, where the agent's choice was accepted. Try-mode also prompts the tutor for a justification of their agreed-upon choice. Question C occurs in either show- or try-mode, and compares the tutor's and the agent's different choices: "*Why is card 2 better than card 4?*" Question D concerns the strategic aspect of how many squares to leave on the game board for the next player. Question E is the most challenging type of question that involves scoring as well as blocking the opponent in the next turn. To be sure of the correct response: "*It's obviously the best one! It gives 1 point and it's the only card that blocks the opponent*", the tutor must predict and distinguish between up to 16 alternative paths two turns ahead: each of their own cards paired with any of the opponent's cards in the next turn.

We chose the multiple-choice response format for several reasons. First, the response alternatives prompt young tutors for ideas as to how to connect their game actions with mathematics and how to make more sophisticated choices. Kim and Baylor (2006) found that novice learners, especially, learnt better when they worked with pedagogical agents that provided ideas proactively, compared to agents that only responded to their requests. Secondly, to model good learner behaviour the agent always wants to know the reason behind actions. Thirdly, based on many hours of game-playing observations, we know that young children's ability to play well is often ahead of (i.e., more advanced than) their ability to explain their actions. Providing multiple-choice explanations is a way to explore if they are able to *identify proper explanations* even if they could not formulate them. Finally, it is a practical reason since the youngest children are not good at writing and it is technically much simpler.

The Teachable Agent – Technical Description

As mentioned, the teachable agent is an extension to an existing game environment. The agent was added to the game environment as a special type of player, and the environment was extended with the agent's knowledge representation, methods for learning, and behaviour.

The Teachable Agent's Knowledge Representation

The agent has 3 types of knowledge: score rules knowledge, strategy rule knowledge, and question knowledge. The score rule and the strategy knowledge represent the two components of playing well: being able to evaluate if a card will score or not and being able to choose the best card among the alternative cards. All score rules represent properties of either the computation in the turn or round, or properties of the resulting board value, so the player needs to predict the computational effect a chosen card will have on the game board.

Each game has a set of score rules, which specify when players receive points. The idea is to draw students' attention to important properties by making these properties yield points since most children want to score and get points. For example, the Pack-Many games is designed to attract attention to the fact that a unit can not contain more than 9 squares, the basic idea behind the base 10 system. The Remove-All games are designed to attract attention to sign rules of addition and subtraction with negative integers. There are 57 different score rules for the 48 game variants. For the Find-pair games, there is only one rule stating that the chosen cards should compute to the goal value. For the Pack-Many games, i.e., when the players act addition and the goal is to get as many carry-overs when adding the card value to the board value, there is only one way the right-most digits can cause a carry-over: if the ones adds up to 10 or more. The corresponding production rule is *rule (addition, red squares, none,* ≥10), which says that for the operation addition, if the sum of the red squares on the board game and the red squares on the chosen cards is equal to 10 or more, then the production rule matches and 1 point is scored. Similarly for the orange squares (the tens), if the tens add up to 10 or more, they also cause a carry over, now to the hundreds. However, there is one more situation causing the orange squares to carry over: if the red squares carry over and the orange squares add up to 9, since the place value positions affect each other by carrying and borrowing. Hence, we have the three production rules for 2-digit additions yielding a point each if matched:

rule (addition, *red squares, none,* ≥10)
rule (addition, *orange squares, none,* ≥10)
rule (addition, *orange squares, one carry-over from red squares, =9)*

The addition 46+57 thus yields 2 points, since the first and last rule match. Such patterns are fundamental to the base-10 system, and therefore we draw the students' attention to the packing patterns by making them scoring rules.

The strategic knowledge needed to play well in the different games is also represented by rules. There are basic strategy rules and advanced strategy rules. Basic strategy rules take into account the player's own hand and the game board; they do a 1-step look ahead. Advanced strategy rules also consider the other player's hand and look 2 steps ahead. There are about 30 different rules, some generic to all games, some game-specific, and some specific to a certain playing mode (collaborative or competitive). Examples of generic rules are *Best_1_Only,* which represents choosing the only card scoring 1 point that is the best, and *Chosen_1_Of_Best_2,* which means choosing a 1-point card when there is a better 2-points card, that is, choosing a card that that yields points but is not the best choice. An example of a game-specific strategy is a rule for the Pack Many game in competitive mode where it is good to leave few squares on the game board for the opponent. In competitive mode there are generic rules for blocking the opponent from scoring, and corresponding rules for allowing the partner to score in collaborative mode.

The question knowledge is represented by text templates containing placeholders. A question template consists of id, criterion, game state description, question text and a set of response alternatives. The question criterion is a label that indicates which learning objective the question concerns. This information is used to select relevant questions from the question bank. The game state description is used to check if the question is applicable in a given situation. Important criteria have several associated questions. Each question has several different formulations with the same content, so questions can be rephrased. Finally, the placeholders in the template can be used to refer to specific information from the current game state. For example, the placeholder *$AV* will be replaced by the Agent's card Value when instantiated. Below is a template about scoring of cards, which involves the agent's and the child's choices of cards (*$AV* and *$CV*):

<*408*: ***compare two choices with respect to points*** >
Q: *As you saw, I chose the card $AV. Is $CV better?*
> *No, both are just as good but none of them give any points*
> *Yes, $CV gives 1 point but $AV gives nothing*
> *Yes, but both $CV and $AV give 1 point*
> *Don't know*

There are normally three response alternatives besides the "don't know" option. One response is correct, the others wrong, but which is correct depends on the game state description associated with the question template. This means that students cannot know from the formulation of the question which response that is correct; it is *situation dependent*. For generating questions, we have a bank of question templates from which a set of questions is extracted for a given game. When there is time for the agent to ask a question, a set of applicable questions is filtered out from the game's question bank that matches the current game state and is within the development zone of the agent's knowledge level. The set of questions constitute the knowledge base to be learned for that particular game.

How the Teachable Agent Learns from the Student Tutor

Agents' learning comes from two sources; from tutors' playing behaviours and from tutors' responses to the agents' questions. Following the apprenticeship teaching model, agents should learn from observing tutors' playing behaviours and try to mimic such behaviour. This means that the agent keeps track of when the student tutor scores a point with a chosen card, and when the tutor accepts a scoring card chosen by the agent in try-mode. These are signs of good game playing behaviour. However, the agent also records when the tutor misses cards scoring better than the chosen card and when the tutor rejects and replaces the agent's choice with a worse card. These are signs of less good game playing behaviour. This behaviour learning is achieved in two steps: 1) the tutor's game-playing actions are observed and logged in the database, and 2) these logs are related to the specified knowledge requirements to achieve the agent's current knowledge level. Each score rule or strategy rule, therefore, has a quadruple of variables recording what the tutor does during teaching sessions. The idea is to keep track of the actions that support the idea that the tutor understands the rule, as well as the actions indicating the opposite. Recall that we can only observe which card the tutor chooses, we do not know if the choice is because of the scoring or strategy rules associated with the choice, or a lucky coincidence. However, if a tutor consistently chooses cards associated with a rule she probably knows the rule. Supporting actions are when the tutor chooses a card that matches the rule, or acknowledges the agent's choice of cards that match the rule. Indications that the tutor is not aware of the rule arise when the rule supports playing a better card than the one actually chosen (where better means either that the card scores more points or is strategically better). Hence, the quadruple of observed playing behaviour keeps track of the total number of occurrences, the two positive indication parameters, and the negative indication variable:

Rule :: *[nrOccurred, nrChosenChild, nrAcceptedAgent, nrMissedbetter]*

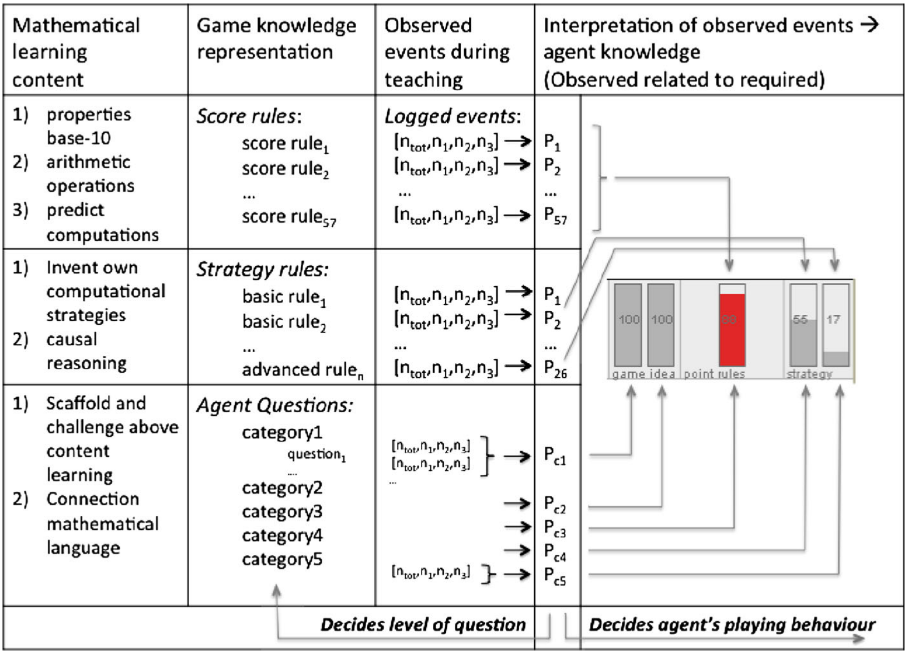
In order to reason about card choices in this way, the system needs to analyse and evaluate each card in every turn during game play. A card is good if it scores, that is if one or more score rules will be invoked when the card is played (this is deterministic since it is part of a computation). After the scoring is decided, the alternatives are compared to each other in order to evaluate their basic strategic value (i.e., a 1-point card can be good if the others do not score but bad if other cards give 2 or more points). Then, the algorithm looks ahead one step by computing the result of each choice and evaluates the situation for the other player. This is part of a standard procedure in the alpha-beta algorithm commonly used for machine playing of two-player games. The game strategy rules are only used as part of the evaluation function of the alpha-beta search, and should not be confused with the minimax algorithm that is always optimal with respect to the evaluation values. To achieve a non-optimal playing behaviour from the agent, the evaluation function will be adopted to the agent's knowledge level, but the minimax algorithm is the same. We have decided not to look ahead more than two steps since the purpose is to simulate human behaviour (and we have observed students do that) and not to make the best machine playing system. Finally, the alternatives are compared against each other, resulting in a "goodness value" for each alternative, taking into account scoring and strategic value of the card in relation to the alternatives.

The other learning sources for the agents, the agents' questions, are treated in a similar manner. Quadruples are used to log all the tutors' responses to the questions, comprising the number of occurrences asked, as well as number of occurrences in which the questions were answered by the correct, the wrong and the 'don't-know' option. Finally, these logging variables are then used to interpret the logged behaviour and compute a measure associated with each rule reflecting all of the tutors' previous playing and question actions. We use these measures as probable or estimated rule knowledge levels of the tutor, and the agent adopts these knowledge levels. The measure is computed as the ratio of positive indications and negative indications of the rule's observed usage. All positive indications are considered, but we allow a smaller part of the rules to be missed, meaning that some negative indications are ignored. The measures (denoted by the $p_{i\ in}$ Fig. 6) are then combined and related to the required knowledge specification to calculate the agent's knowledge levels shown to the tutor in the skills meters.

How the Agent Behaves According to its Knowledge

The agent's knowledge level affects its behaviour in two aspects; how well cards are chosen and which questions are asked.

The agent's playing behaviour is determined as follows: When the agent plays alone or is tutored in try-mode, it has to make its own card choices. The system analyses the situation and identifies which score and strategy rules are present for each card alternative. An optimal player would base their choice on *all* of these rules, but the agent should not play optimally if not trained to full mastery. Therefore, for each of the identified rules, the system will decide whether the agent "sees" the rule or not depending on the agent's current knowledge of that rule (the p_i in Fig. 6). This is how the evaluation function of the alpha-beta search is altered to achive non-optimal behaviour. The knowledge level is based on the measure for the rule as well as



Mathematical learning content	Game knowledge representation	Observed events during teaching	Interpretation of observed events → agent knowledge (Observed related to required)
1) properties base-10 2) arithmetic operations 3) predict computations	*Score rules*: score rule$_1$ score rule$_2$ ... score rule$_{57}$	*Logged events*: $[n_{tot}, n_1, n_2, n_3] \rightarrow$ $[n_{tot}, n_1, n_2, n_3] \rightarrow$ ... $[n_{tot}, n_1, n_2, n_3] \rightarrow$	P_1 P_2 ... P_{57}
1) Invent own computational strategies 2) causal reasoning	*Strategy rules:* basic rule$_1$ basic rule$_2$ ... advanced rule$_n$	$[n_{tot}, n_1, n_2, n_3] \rightarrow$ $[n_{tot}, n_1, n_2, n_3] \rightarrow$ ... $[n_{tot}, n_1, n_2, n_3] \rightarrow$	P_1 P_2 ... P_{26}
1) Scaffold and challenge above content learning 2) Connection mathematical language	*Agent Questions:* category1 question$_1$ category2 category3 category4 category5	$[n_{tot}, n_1, n_2, n_3]$ $[n_{tot}, n_1, n_2, n_3]$ } → ... → → → $[n_{tot}, n_1, n_2, n_3]$ } →	P_{c1} P_{c2} P_{c3} P_{c4} P_{c5}
		Decides level of question	***Decides agent's playing behaviour***

Fig. 6 Overview of the teachable agent's learning process

questions associated with the rule. The level is used as a probability measure meaning that if a rule has knowledge level 86 %, the agent "will see" the rule with a probability of 86 %. Then the agent makes a rational decision based on what it "sees" in each situation. Hence, no knowledge means no rules identified and hence a guess is the only option. When all rules are seen, the agent will make the optimal choice. This approach is more flexible and continuous compared to the previous knowledge level estimation with a pre-set threshold that should be passed to use the rule, which is described in Pareto (2009).

The agent's question asking behaviour is determined as follows. The question templates are organized in a progression tree, where each question template is a node in the tree. The tree is ordered at two levels: between categories and within category. There is one question template progression tree per game and arithmetic operation, including strategy questions for both collaborative and competitive play. There are conditions regulating progressions from one category to the next, as well as conditions regulating progression between nodes in the tree. Currently, progression from one category to the next is allowed when 75 % of the previous questions are answered satisfactory, as a simple condition to be able to start "peeking" at the next question category before completing the previous. Within a category, there can be conditions between nodes so that some questions are asked before others, as illustrated Fig. 7 below where the "predict scoring" category progression tree is shown. (For the full progression tree, see appendix A.)

For the Predict scoring category shown in Fig. 7, there are only two types of question templates; for the scoring and non-scoring situation, respectively. The progression tree states that the tutor must answer at least 6 scoring and 4 of the non-scoring type questions satisfactorily, before the questioning can advance to the next question, the verification question 399. A question is satisfactorily answered when there are more correct than wrong answers, or at least 3 correct answers (to avoid going on for ever). If the tutor does not manage to answer these questions satisfactorily, the agent will not proceed with more advanced questions, but stay at that level until the condition is fulfilled. Verification-questions thus function as gatekeepers into the next category. The combination of the wrong-answer penalty schema and requiring several questions to be satisfactorily answered for each knowledge criterion, makes it unlikely that students can progress solely by guessing or trying all alternatives. The requirements for questions are based on the designer's judgment of how many questions suffice to demonstrate understanding of particular criteria, but these numbers can be altered or be part of a future agent profile.

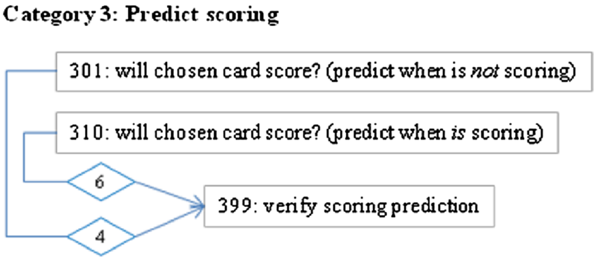

Category 3: Predict scoring
301: will chosen card score? (predict when is *not* scoring)
310: will chosen card score? (predict when *is* scoring)
6
4
399: verify scoring prediction

Fig. 7 Question progression tree within the scoring category

To set a goal for the overall tutoring task, we need to specify how much knowledge is required for the agent to be fully trained and to provide feedback to the tutor. This is specified as a layer on top of the other modules, so it can be altered. The knowledge level meters for the first two categories (game idea and graphical model knowledge) are currently based entirely on the student tutor's answers to questions, since it hard to judge such knowledge by observing playing behaviour. The remaining three categories are based for 50 % on the correctness of answers to questions and for 50 % on the observed game-playing performance as explained above.

Finally, the mechanism implemented for the teachable agent has been useful for other purposes: we can define computer players at many different levels by pre-setting knowledge levels and treat them as non-teachable agents. We have also used the play evaluation module, e.g., goodness values, to register the students' playing performance in the game log data and we have used it to perform in-game performance analysis.

The Study

In this section we will report findings from an evaluation study of the game. This study is the third step in a dissemination procedure to integrate the game in mathematics teaching in a Swedish municipality. The first step was a pilot study; the second step was a study enrolling 1st, 3rd and 5th graders; and this study, being the third, involves about 500 students from grade 2 to 8. The diffusion model was based on the prior successful diffusion of technology in the region's schools, where teachers who were already involved (rather than management or researchers) recruited new teachers via social networks. (The school management approved the project and the diffusion procedure.) The teachers' interest in the learning game was crucial to the study, since they had to learn the game in order to use it in their teaching without support. This diffusion model was designed to reflect how new learning materials are spread in Swedish schools.

One of the goals of the study (and the previous studies) is to determine the most appropriate target groups for the mathematics content included in the game. Negative numbers and computation-strategy invention are not concepts normally taught in early primary school. Also, we have the ambition to reach low-performing, at-risk students as early as possible in education. Therefore, a reasonable recruitment base was set to grade 2-6 for mainstream students, and any grade for students with special needs in mathematics. In total, 22 classes from 9 schools were recruited through the teacher network.

The study is fully authentic, and by this we mean that:

1) Lessons are planned by the teachers (e.g., organization in class, choices of games, student play constellations; instructions provided; goals of the game play);
2) Game sessions are instructed and lead by the teachers without support;
3) The game is used instead of other subject activities during ordinary classroom lessons;
4) No additional resources are provided (e.g., time, personnel, researchers, hardware);
5) The class situation is normal (e.g. often full class 20-25 students, occasionally half-class).

The advantage of an authentic setting is high external validity; the disadvantages are lack of control of many parameters related to teachers' interest and competence as well as usage of the game in the classroom. However, these are differences that exist in schools. All teachers received two half-days of training together with the instruction to use the game in class preferably once a week for the remaining 3 months of the semester. The duration of the study is so that the students go beyond the initial excitement of playing a novel game. The teachers received no directions on how to integrate the game in classroom instruction, since another aim was to explore different approaches. From post-study teacher interviews and game logs we learned that the level of guidance the teachers gave and the amount of time the students ended up playing varied substantially between classes.

The study enrolled 22 playing classes in total: 20 mainstream classes in 2nd to 6th grade and 2 small classes of students with special needs in mathematics in 7th and 8th grade. The enrolled classes were diverse in many aspects; they came from 9 different schools including one Montessori-school and one international school; they came from different socio-economic backgrounds and there were single-age classes as well as age-integrated classes. However, the diversity of classes is not used for comparison; it reflects normal variations in schools. The students played the game approximately once a week for 3 months, and they took paper- and pencil mathematics pre- and post-tests. There were two types of tests; one traditional test with arithmetic problems and one mental model test. The traditional test was used for grades 2, 4, 6 as a follow up study from the year before when grade 3 and 5 were examined. The 2th, 4th, and 6th grade classes were assigned to a play condition and a control condition based on when the offer to participate in the study was accepted. Control classes participated in their ordinary mathematics lessons during the study intervention. The material covered in the game, that is the base-10 system, the relations between arithmetic operations, and problem solving, is a major and fundamental part of mathematics curricula for grade 1 to 3, and is practiced directly or indirectly in about 75 % of the lessons. For grade 4-6 the game content is practiced less directly, but given this content is fundamental to other mathematics topics, it is addressed in at least half of the math lessons. Since game play constituted at most half of the math lessons, it is reasonable that the control classes got comparable amount of time with these topics. Moreover, the school management and the teachers would not have accepted to devote so much time of their lessons to the game unless they considered its content important and aligned with the grades' curricula. We did not enroll controls for the special need students, since the large variation of difficulties within the group makes a control group less meaningful. All grade 3 and 5 students took the mental model test, with no control classes. The researchers administered all tests. There was also a web-based questionnaire concerning the students' perception of and attitude towards the game distributed to all playing classes at the end of the 3-month trial.

Data Collection

For the overall investigation into students' learning gains, the traditional mathematics pre- and post-tests were used. The tests consisted of the same type of math problems as in our previous studies, and were constructed in three difficulty levels for grades 2, 4 and 6, respectively. The problems address conceptual understanding of the base-10 systems, relations of arithmetic operations, and problems that can be solved by reasoning rather than cumbersome computations. The three math tests contained 48, 55 and 66 small problems, respectively. The problems are categorized as computational, conceptual, effective strategy and other problems. All students received the same pre- and post-tests. The special needs teacher decided which of these tests were appropriate for their students. Test results are computed as percentage correct of the maximum value in the respective categories. Gains in learning are computed as the difference between post- and pre-tests. In the mental model test, the students' pictures were evaluated in 7 categories. The score was computed as a percentage of a maximum value.

Students' perceptions of playing the games were collected in a web-based questionnaire. Due to the young age of the respondents, we could neither use standard questionnaires nor ask too many questions. There were 12 questions in total. For students' perception of the agent-tutoring task, there were 4 questions concerning a) importance, b) difficulty, c) role division: tutor-tutee and d) enjoyment of the agent-tutoring task. The questions use 4- and 5-item scales. Students' responses to the four questions are compared in total and with respect to grade level. The importance, difficulty and enjoyment ratings are also compared grade-wise to each other by mapping the response alternatives to a 0-10 scale and comparing averages.

Students' performance in answering the agent questions was computed from the data collected in the game-logging database. This data includes number of correct, wrong and 'don't know' answers per question category and class level. From this data we computed measures of tutors' answering performance in terms of correctness, error and uncertainty ratio per question category and per grade level. This was done in order to explore students' performance with respect to different categories of mathematical thinking and grade. The correctness ratio is converted to a scale 0—10, so students' actual performances can be compared to their subjective ratings of the agent-tutoring task.

We have used test results in combination with game-logs to compare in-game and out-of-game performance. For the in-game performances we have used two measures: *game progression* and *agent knowledge*. Having no control over which games and in which order students play means that our measure of game progression has to be fairly rough. We have defined game progression to be the linear approximation of all goodness-values of the students' card choices over time. The goodness-value measures how close a card choice is to being optimal given the circumstances in the game situation, so it is a general measure of how well a student chooses each card. If the goodness-values are consistently better than a random choice we can assume that the student uses her understanding to choose cards. For the agent knowledge measure, we use the percentage of the student's best-trained agent compared to a fully trained agent. These in-game measures are used to examine if game progression has an effect on maths progression, if agent knowledge has an effect on math progression, and if the total number of played turns correlate with maths progression and game progression.

Results

The results from the study are organized according to the research questions regarding: 1) overall learning gain 2) students' perception of the agent-tutoring task, 3) students' performances in the agent-tutoring task and 4) relations between in-game and out-of-game performances

Learning Gains

Learning gains were examined in a pre-post quasi-experimental study with play and control classes in grades 2, 4 and 6, as a follow up study from the year before when the same experiment was conducted for grades 3 and 5 (Pareto et al. 2011). There were 154 students in the play condition with a pre test mean of 58 % and 129 in the control condition with a pre test mean of 63 %. A pre-treatment test (*T*-test) for between group comparisons showed that there were no significant differences ($p>.05$) between the two condition groups with respect to pre test results.

For the between group comparison for learning gains from pre- to post-test, we found that the play group gained more in mean test scores than the control group for all four problem types: computational, conceptual, effective strategies and other, as shown in Table 1.

In the conceptual maths problems category there was a significant difference, and a regression model was constructed ($F_{3.279}=8.93$ $R^2=8.8$ %, $p<.0005$). The regression model controls for the effects of pre-test scores. The result of the regression model (see Table 2) shows that there is a statistically significant effect for the play group compared to the control group ($F_{1.279}=8.028$, $p=.0049$). The estimated model for additional learning effect for the play condition=0.435 – 0.610 * pre-test scores. This result supports our previous findings that the game has a learning effect for conceptual understanding, and that low-achieving students gain more. For the other categories of maths problems, the playgroup gained more than the control group, but these gains were not significantly different to the controls.

Students' Perception of the Agent-Tutoring Task

For the students' perception of the agent-tutoring task, we analysed the 201 web-survey responses we received from game-playing students. Unfortunately, not all teachers took the time to have their students fill in the questionnaire. The distribution of responses per class level were in grade 2 (53), 3 (24), 4 (67), 5 (32), 6 (14), 7 (5) and 8 (6). Most students judged answering agents' question as important; 50 % answered very important and 29 % quite important, indicating engagement in the task. The 8 % of students

Table 1 Between group comparison of mean score gain in the different math problem types

Learning gains, mean difference from pre to post test in the groups:

Group	N	Compute	SD	Concept (**)	SD	Strategy	SD	Other	SD
Control	129	1.29 %	.2082	7.75 %	.4160	9.57 %	.3691	6.20 %	.3027
Play	154	1.79 %	.2436	16.74 %	.4456	14.07 %	.3823	9.52 %	.2920

Table 2 The regession model for conceptual math problem gain

Tests of Between-Subjects Effects

Dependent Variable:Diff_bothConcept

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	4.650[a]	3	1.550	8.927	.00001	.088
Intercept	4.826	1	4.826	27.795	.00000	.091
Pre_Math_tot	2.782	1	2.782	16.024	.00008	.054
play_control	1.394	1	1.394	8.028	.00494	.028
play_control * Pre_Math_tot	1.127	1	1.127	6.490	.01139	.023
Error	48.444	279	.174			
Total	57.617	283				
Corrected Total	53.094	282				

a. R Squared=.088 (Adjusted R Squared=.078)

who answered not important at all, almost all came from the same grade 6 class, where 64 % of the students judged the task as being of little importance or not important at all. Regarding perceived difficulty, only 12 % of the students judged the questions to be quite (10 %) or very (2 %) difficult. One must consider that most students only trained their agents less than halfway, and question difficulty increases with advancement. Not more than 19 % of students reached category 4 or 5 during the study, where questions involve choice discrimination and more advanced reasoning.

Regarding students' responses to perceived role division of the agent-tutoring task, that is, the "who teaches who" question, almost all students (96 %) perceived that they teach the agent, which is also what they are told. 35 % also recognized that they learn from the agent, which is one of the purposes of providing response alternatives to choose from. The students enjoyment measures of the three playing modes (student plays, student teaches the agent, agent plays) showed that teaching the agent, which is the only mode when the agent asks the student questions, is most popular for all grades except grade 6, followed by watching the agent play which is most popular in grade 6. For just playing the game, the enjoyment rating is lower for all grades except the youngest. In general students rate teaching the agent as 17 % more enjoyable than only playing the game, and 92 % of the students rated it as fun or okay.

Students' Performances of the Agent-Tutoring Task

We analysed the game logs to examine the students' performance in answering agent questions. In total, 7820 questions were asked by agents and answered by students during the study period. Students were required to respond to all the questions asked. To examine the three response types (Correct, Wrong and Don't know) in relation to student grades, we conducted a z-test to see if differences between proportions are significant (see Table 3).

Table 3 Students' responses to agent questions, result of z-test of differences between proportions=

			Grade							total
			2.00	3.00	4.00	5.00	6.00	7.00	8.00	
Corect_not	Correct	Count	1187a	469b	1435b	738c	77a c	217a	366a	448,9
		% within grade	49.6 %	66.7 %	66.6 %	61.3 %	53.1 %	45.7 %	49.1 %	57.4 %
	Wrong	Count	326a	38b	155b	124c	14a b c	52a c	21d	730
		% within grade	13.6 %	5.4 %	7.2 %	10.3 %	9.7 %	10.9 %	2.8 %	9.3 %
	Dont Know	Count	879a	196b	565b	342b	54a c	206c d	359d	2601
		% within grade	36.7 %	27.9 %	26.2 %	28.4 %	37.2 %	43.4 %	48.1 %	33.3 %
Total		Count	2392	703	2155	1204	145	475	746	7820
		% within grade	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %

Each subscript letter denotes a subse1 of Grade categories whose column proportion do not differ significantly from each other at the 0.5 level

Students in grade 3 and 4 answered the most agent questions correctly; their correctness ratio of 66.7 % is highest and significantly different from the others. Students in grade 2, 6, 7 and 8 have a correctness ratio of about 50 %, are in the lowest performing group. The 7 and 8 grade special needs students have a don't-know ratio close to 50 % and that is significantly higher than the other grades. We also analysed the response correctness per question category with respect to grade level, to investigate if there are differences in performances related to grade (see Fig. 8).

The variation in performance among grade levels is quite small; they all follow similar patterns, which is revealed by z-tests for the difference in proportion per category. The proportion of correct answers is about 55 % for the game idea questions, the graphical metaphor questions and the basic strategy question. The scoring questions had a proportion of 84 % and the most advanced strategy had one of 78 %. We can see that special needs students in grade 7 and 8 start out lowest, but perform at the same levels as the others from category 2 forward. Students seem to learn to avoid guessing;

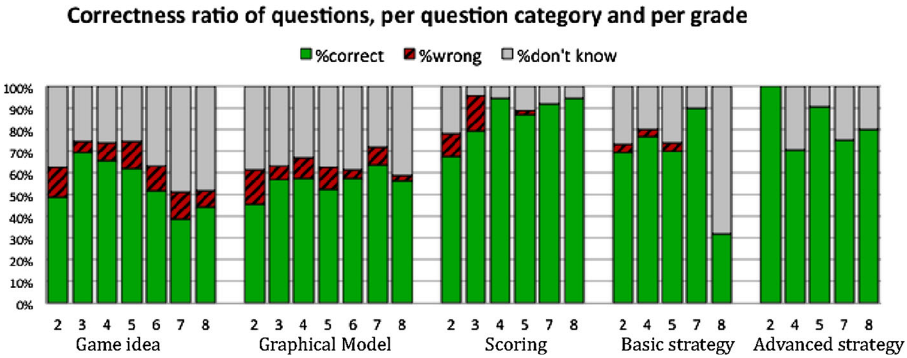


Fig. 8 Response correctness ratio for question category 1 to 5 per grade level (x-axis). Missing bars means that there were no students in the grade who reached that category level.

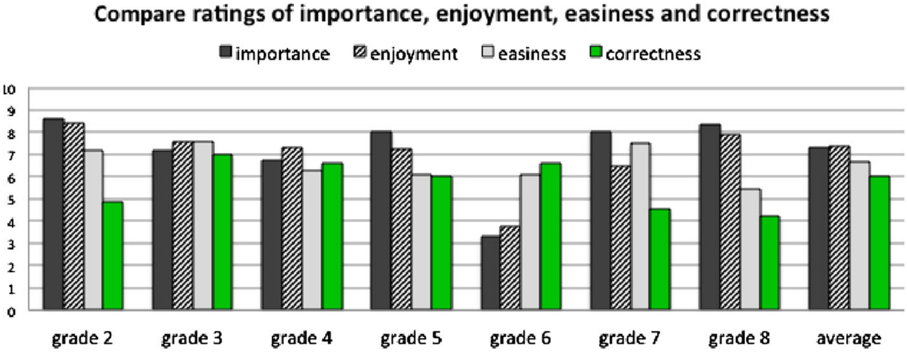


Fig. 9 Comparing students' survey-ratings and students' in-game performances of the agent-tutoring task. All measures are uniformly mapped onto a common scale (0—10).

the percentage of wrong answers declines from 10 % in the easiest category to 0 % in the most difficult, which was our intention with the wrong-answer penalty scheme. The students who reached the more advanced categories are likely the better students; nevertheless these students got more answers wrong in the earlier categories. The question responses in this analysis originate from different game variants, but since the question categories reflect the same type of mathematical thinking across different game variants and levels, they are still meaningful to compare. Finally, we have compared students' ratings with actual in-game performances, as illustrated in Fig. 9.

Analysis at the student level shows that correlations between importance, enjoyment and easiness are all significant at the 0.01 levels, whereas correctness does not correlate with any of the other measures.

Relations Between in-Game and out-of-Game Performances in Mathematics

To examine the relations between in-game measures and out-of-game mathematics tests, we conducted the following analyses:

To examine if game progression is significantly related to maths progression (i.e., the score on the mathematics post-test, controlling for pre-test), a regression model was constructed ($F_{8.264}$=11.582 R^2=26.0 %, p<.0005). The regression model controls for the maths pre-game test score and class level. The result of the regression model shows that game progression has a small but statistically significant positive connection with math progression ($F_{1.264}$= 11.525, p=.001) and the estimated model was maths progression=0.407–0.431*math pre test+0.004*game progression + (class level). This means that the more the student progresses in the game (that is becomes better at choosing the best card), the higher the gain in maths test, when the maths pre-test level is controlled for.

To examine if agent knowledge has a significant relation with maths progression, a regression model was constructed ($F_{11.142}$=6.78, R^2=34.4 %, p<.0005). The regression model controls for the maths pre-test score and class level. The result of the regression model shows that agent knowledge has a statistically significant positive connection with maths progression ($F_{1.141}=12.82$, $p<.0005$), where maths progression=0.089 – 0.585*maths pre test+0.475*agent knowledge + (class level). This means that the better the agent is trained, the higher the gain in maths test, when maths pre-test level is controlled for. What we see here, as in previous results, is that low-performing students gain more.

The total number of played turns has a small but significant correlation with math progression ($r=0.132$, $p=0.039$), but has no significant correlation with game progression ($r=-0.053$, $p=0.375$). This result is aligned with our observations and expectations, and it means that most students learn more if they play more, but it is possible to play much without trying to play well. The motivation to challenge oneself, the *engagement in playing* is crucial for learning.

Discussion

From the comparative part of the study we found that both play and control students learned during the 3-month trial, which is to be expected. There were greater learning gains for the game-playing group than the control group, but the difference was only significant in the category of conceptual understanding. The results are not so strong, but support previous findings that students gain in conceptual understanding (Pareto et al. 2011; Pareto et al. 2012). More importantly, a better conceptual understanding of fundamental topics in mathematics may have effects on future mathematics learning that cannot be revealed in the posttests only, so following the development of the playing classes would be of interest. We did expect that the game-playing students would gain more than control students in the effective strategies category, but that difference was not significant. The reason for the lack of significant differences could be that it is too early for students' learning to be visible in paper and pencil tests, that the problems in the tests are not good at capturing this type of skill, or that the game does not stimulate such learning. For the remaining categories, computational and other problems, we did not expect game-playing students to learn more. Computational problems are of interest since in traditional instruction this is frequently practiced in Swedish schools, and in the game the computations are performed by the system. Other topics such as geometry were included to examine that game-playing students did not drop behind in topics not addressed in the game, since playing took time from other classroom activities. However, the mean gains were higher for the game-playing students in these categories as well, but not significantly so.

The second research question concerned the students' perceptions of the agent-tutoring task. All students, except the class in grade 6, perceived the agent-tutoring task as enjoyable and important. This class played the least and accomplished little in the tutoring task, which is either the result of or the reason for lacking engagement. Reasons could be related to the fact that no teachable agent was yet available for the more advanced games, so perhaps the content was not attractive enough for these students.

Our study shows that the subjective ratings and the performances of students do not coincide. The reason we asked students to rate difficulty was to indicate engagement, not to judge their ability, since if questions were perceived as too difficult or too easy, student engagement could be jeopardized. The responses of difficulty ratings do not give such indications. Despite such self-report, perhaps the questions (or having to read the questions) were almost too challenging for the students in second grade, since they were the only group who rated playing by themselves as more enjoyable than teaching the agent. This observation is aligned with the study of Kim and Baylor (2006), which showed that a low competency pedagogical agent was more motivational than a high-competency agent and from this it seems likely that an agent asking questions at a too difficult level may be de-motivating for the young students. Roscoe and Chi (2007) argue that tutors who feel more capable exert more effort toward tutoring. Also, the authors claim that novice peer tutors often feel anxiety about being responsible for another student's learning, but when the tutee is a computer agent, such responsibility is removed. One student who did not perform well in the teaching role at all, claimed that "I didn't want to teach him [the agent] well", which may be a way to hide incompetence behind an attitude as an ego-protective buffer (Chase et al. 2009). Alternatively, the tutoring task simply did not appeal to that student.

The questions about role-division were meant to examine if students accept the fiction of being the tutor, i.e., if the provision of response alternatives changed the perception of the activity from tutoring the agent to being perceived as "the agent telling them what to answer". The result, that 92 % of students claimed that they taught the agent, gives no such indications. In accordance with Chan and Chou (Chan and Chou 1997), we argue that such fiction adoption is more likely to occur if the agent tutee behaves in a way natural to the student tutor. By natural we mean that 1) the questions should be of the form that the students could have asked themselves, 2) the timing of the question should be reasonable, and 3) the agent tutee should not become too clever too soon.

The third research question concerned the students' *performances in the agent-tutoring task*, with respect to how well they managed to answer the agent's questions. The question answering performance varied more between categories of questions than between the different grade levels. The highest rate of correct answers fell within scoring questions and advanced strategy questions. Scoring questions concern predicting a one-step arithmetic computation, and here students perform well. The first two categories of questions concern relating the game graphical metaphor with mathematics. This connection was not explained a priori; students had to discover this. Thus, it is not surprising that students must be accustomed to the agent's questions and to the consequences of answering or ignoring them, before they can act effectively. In general the students improve their performances in answering the agent's questions while progressing through the question categories. Our interpretation is that the students who progressed through the categories became better at identifying the correct explanations, even though they may not have been able to provide such explanations themselves. This is in accordance with results of King (1994) who suggests that knowledge construction is effectively promoted if students are guided in how to explain and how to ask good questions. There were too few students who reached the more advanced questioning levels in order to provide strong evidence of this, but our analysis shows that even students as young as 2nd graders can answer questions on an advanced reasoning level and that it is feasible to train the agent to "full mastery" of a game variant. This indicates that the agent tutee behaviour and performance is understandable to the student tutors, one of two objectives for learning by teaching according to Chan and Chou (1997). The extent to which students learn to ask good questions and to give good explanations outside the agent-questioning system was not part of this study and is a topic for future work.

The study supports a connection between learning in the game and learning outside the game: students, who engage and progress in the game, gain in conceptual understanding. Here in-game performances are measured by how well the student plays the various games over time and to what level they have managed to teach their best agent. Both measures have positive connections with mathematics test performance, which supports that tutoring the agent has a learning effect. The purpose of the agent-questioning system, to engage students with reflective questions, is rather well supported by the amount of questions correctly answered and the students' positive ratings of teaching the agent. The connection between the agents' knowledge level and learning indicate that the questions help in connecting the in-game graphical language with mathematical language. The agents do model good questioning behaviour, but we do not know if such questioning skills are learned by students or transferred to other situations. An effect of providing scaffolding to students to progress their reasoning abilities is not yet established, but is somewhat indicated by students' performances on advanced strategic questions.

There are many possibilities for future studies related to the game environment. A study concerning the effect of the teachable agent as opposed to the game only has already been conducted, but there are other possibilities, including whether the students will imitate the agent's inquisitive behaviour and learn how to ask good questions and whether students learn to self-explain. The vicarious learning environment using overhearing has been shown to improve students' questions (Craig et al. 2006), but these students are older than our target group. So were the students of the self-explanation group in Wong, Lawson and Keeves (2002) study, who showed more frequent use of self-explanation activities after the training. We believe age (or rather mental development) is a key issue, but according to King (1994) primary school students can be trained to generate questions as well as formulating explanations. We would also like to investigate the extent to which students' learning with respect to reasoning ability (as measured by their playing and agent questioning performances) transfer to other situations.

Conclusions

The study shows that the TA-enhanced game environment can engage primary school students in arithmetic concepts and reasoning from 2nd to 6th grade in mainstream education as well as 7th and 8th grade students with special needs in mathematics. Engagement comes from the game activity and from students teaching the agent, which further enhances the game experience. The engagement was lower in 6th grade and the enjoyment of teaching the agent lower in 2nd grade, compared to the other grades. The learning environment is appropriate for such a wide age- and knowledge range of students due to many game variants at varying difficulty levels, and to the main activity of choosing good cards being meaningful and possible to perform at varying skill-levels (from pure guessing to advanced strategic reasoning). Almost all students are engaged, but we also see that engagement is necessary for productive learning. Thus, it is important to design for *engagement* by matching challenges and difficulties with incentives and motivating activities.

The study results provide support that students learn when playing the game. Game-playing students learn more than the control students in problems dealing with conceptual understanding of the base-10 system when playing on their regular mathematics lessons, and they did not drop behind in other mathematical skills being tested but not practiced on in the game. The addressed topics constitute a major part of the lower grades' curricula and are therefore most likely substantially practiced during a three-month period of mathematics lessons even for the control classes.

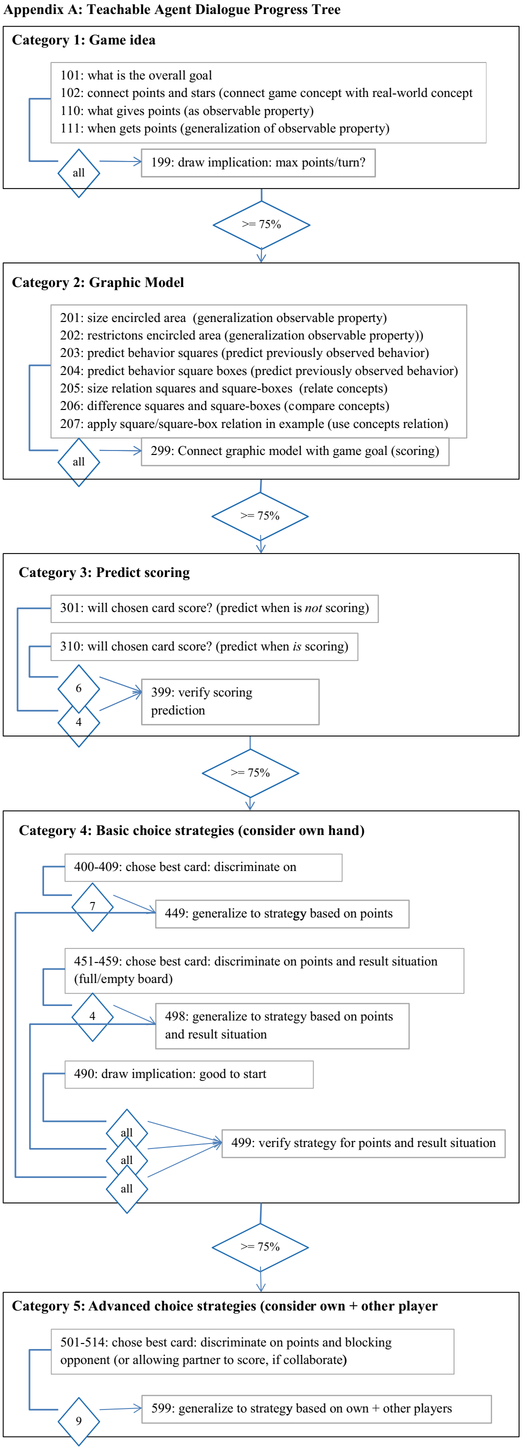
The results show that young primary school students can act as successful tutors, with the additional support of providing response alternatives to prompt for reflection and explanation. Progression patterns of the agent-tutoring task were similar for all grades, and how far students reached in their teaching was related to teaching amount rather than grade level. The agent-tutoring task may have been too challenging for grade 2 students, partly due to their lack of reading. They were the only group who judged playing the game as more enjoyable than teaching the agent. We have indications that the agent questioning activity helped students transfer their in-game knowledge to traditional mathematics. To investigate this further is future work.

It is challenging to study truly authentic classroom usage of educational software. The lack of control over learning situation parameters makes the empirical data complex and cumbersome to analyse. Also, to study learning content, such as deeper conceptual understanding, requires the study to stretch over time since such understanding takes time to acquire. This in turn means that there are other learning activities going on in parallel, which makes it difficult to isolate the learning effect from the studied activities. Yet, to study learning effects of the real use of our systems, these challenges must be handled.

Finally, our system shows that teachable agents can be appropriate for students as young as 8-years-old if the agents are designed to meet their cognitive and social development. An agent that takes the initiative to challenge the student with thought-provoking questions and provides multiple-choice response alternatives is a way to engage and scaffold students of this age. Using teachable agents in the role of a reflective game-playing companion is a promising candidate for leveraging learning in educational games to deeper and more advanced learning. The teachable agent enhances learning by engaging students in reflection and explanation of their game actions under the pretext that they teach an agent to play the game. This idea combines the motivational power of games with the reflective power of a teachable agent asking deep questions on the learning material during game play, and could improve the learning effects of games.

Acknowledgments This work is a result of collaboration with Daniel L. Schwartz, School of Education, Stanford University, USA, and is the financed by the Wallenberg Global Learning Network. The author also wants to thank Tobias Arvemo for statistical guidance, and the reviewers for extremely detailed and insightful comments on previous versions of this paper.

Appendix A: Teachable Agent Dialogue Progress Tree


Appendix A: Teachable Agent Dialogue Progress Tree
Category 1: Game idea
101: what is the overall goal
102: connect points and stars (connect game concept with real-world concept
110: what gives points (as observable property)
111: when gets points (generalization of observable property)
all
199: draw implication: max points/turn?
>= 75%
Category 2: Graphic Model
201: size encircled area (generalization observable property)
202: restrictons encircled area (generalization observable property))
203: predict behavior squares (predict previously observed behavior)
204: predict behavior square boxes (predict previously observed behavior)
205: size relation squares and square-boxes (relate concepts)
206: difference squares and square-boxes (compare concepts)
207: apply square/square-box relation in example (use concepts relation)
all
299: Connect graphic model with game goal (scoring)
>= 75%
Category 3: Predict scoring
301: will chosen card score? (predict when is *not* scoring)
310: will chosen card score? (predict when *is* scoring)
6
4
399: verify scoring prediction
>= 75%
Category 4: Basic choice strategies (consider own hand)
400-409: chose best card: discriminate on
7
449: generalize to strategy based on points
451-459: chose best card: discriminate on points and result situation (full/empty board)
4
498: generalize to strategy based on points and result situation
490: draw implication: good to start
all
all
all
499: verify strategy for points and result situation
>= 75%
Category 5: Advanced choice strategies (consider own + other player
501-514: chose best card: discriminate on points and blocking opponent (or allowing partner to score, if collaborate)
9
599: generalize to strategy based on own + other players

References

Avigad, Jeremy, "Computers in Mathematical Inquiry" (2008). Department of Philosophy. Paper 28. http://repository.cmu.edu/philosophy/28

Biswas, G, T. Katzlberger, J. Brandford, D. Schwartz, D.S. & TAG-V (2001). Extending intelligent learning environments with teachable agents to enhance learning. In *Artificial Intelligence in Education*, J.D. Moore et al. (Eds.) (pp. 389-397). IOS Press

Biswas, G., Leelawong, K., Schwartz, D., & Vye, N. (2005). Learning by Teaching: A new Agent Paradigm for Educational Software. *Applied Artificial Intelligence, 19*, 363–392.

Blair, K., Schwartz, D. L., Biswas, G., & Leelawong, K. (2007). *Pedagogical Agents for Learning by Teaching: Teachable Agents*. Saddle Brook Then Englewood Cliffs NJ: Educational Technology.

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.), *Artificial Intelligence in Education* (pp. 21–28). Amsterdam: Ios Press.

Bullock, M., & Ziegler, A. (1999). Scientific Reasoning: Developmental and Individual Differences. In F. E. Weinert & W. Schneider (Eds.), *Individual development from 3 to 12: Findings from the Munich Longitudinal Study* (pp. 38–54). Cambridge: Cambridge University Press.

Carpenter, T. P., Fennema, E., & Romberg, T. A. (Eds.). (1993). *Rational Numbers: An Integration of Research*. Hillsdale: Lawrence Erlbaum Associates, Inc.

Case, R., Okamoto, Y., Griffin, S., McKeough, A., Bleiker, C., Henderson, B., et al. (1997). The Role of Central Conceptual Structures in the Development of children's Thought. *Monographs of the Society for Research in Child Development, 61*, 1–295.

Chan, T.-W., & Chou, C.-Y. (1997). Exploring the Design of Computer Supports for Reciprocal Tutoring. *International Journal of Artificial Intelligence in Education, 8*, 1–29.

Chase, C., Chin, D. B., Oppezzo, M., & Schwartz, D. L. (2009). Teachable Agents and the Protégé Effect: Increasing the Effort Towards Learning. *Journal of Science Education and Technology, 18*, 334–352.

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from Human Tutoring. *Cognitive Science, 25*, 471–533.

Chinn, C. A., & Malhotra, B. A. (2002). Children's Responses to Anomalous Scientific Data: How is Conceptual Change Impeded? *Journal of Education and Psychology, 94*, 327–343.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive Apprenticeship: Teaching the Crafts of Reading, Writing, and Mathematics. In L. B. Resnick (Ed.), *Knowing, Learning, and Instructions: Essays in Honor of Robert Glaser* (pp. 453–494). Hillsdale: Lawrence Erlbaum Associates, Inc.

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). Deep-Level Reasoning Questions Effect: The Role of Dialog and Deep-Level Reasoning Questions During Vicarious Learning. *Cognition and Instruction, 24*, 563–589.

Dillon, J. T. (1986). Student Questions and Individual Learning. *Educational Theory, 36*, 333–341.

Dillon, T. J. (1988). *Questioning and Teaching: A Manual of Practice*. New York: Teachers College Press.

Fisch, S. M. (2005). *Making Educational Computer Games "Educational"* (In *Proceedings of the 2005 Conference on Interaction Design and Children* (pp. 56-61)). New York: ACM.

Goos, M. (2004). Learning Mathematics in a Classroom Community of Inquiry. *Journal for Research in Mathematics Education, 35*, 258–291.

Graesser, A. C., & Person, N. K. (1994). Question Asking During Tutoring. *American Education Research Journal, 31*, 104–137.

Graesser, A. C., Baggett, W., & Williams, K. (1996). Question-Driven Explanatory Reasoning. *Applied Cognitive Psychology, 10*, 17–31.

Graesser, A. C., McNamara, D., & VanLehn, K. (2005). Scaffolding Deep Comprehension Strategies Through Point & Query, AutoTutor, and iSTART. *Educational Psychology, 40*, 225–234.

Graesser, A. Otero, J. Corbett, A. Flickinger, D. Joshi, A. & Vanderwende, L. (2009). Chapter 1: Guidelines For Question Generation Shared Task Evaluation Campaigns, In V. Rus and A.C. Graesser (Eds.) *The Question Generation Shared Task and Evaluation Challenge*, http://www.questiongeneration.org.

Gutstein, E., & Mack, N. K. (1999). Learning About Teaching for Understanding Through the Study of Tutoring. *The Journal of Mathematical Behaviour, 17*, 441–465.

Hunter, R. (2008). Facilitating communities of mathematical inquiry. In M. Goos, R. Brown, & K. Makar (Eds.), Navigating currents and charting directions. *Proceedings of the 31st annual conference of the Mathematics Education Research Group of Australasia* (Vol. 1, pp. 31–39). Brisbane: MERGA.

Jonassen, D., & Ionas, I. (2008). Designing Effective Supports for Causal Reasoning. *Educational Technology Research and Development, 56*, 287–308.

Kay, J. (2008). Lifelong Learner Modeling for Lifelong Personalized Pervasive Learning. *IEEE Transactions on Learning Technologies, 1*, 215–228.

Ke, F. (2008). Alternative Goal Structures for Computer Game-Based Learning. *International Journal of Computer-Supported Collaborative Learning, 3*, 429–445.

Kim, Y., & Baylor, A. L. (2006). Pedagogical Agents as Learning Companions: The Role of Agent Competency and Type of Interaction. *Educational Technology Research and Development, 54*, 223–243.

King, A. (1994). Guiding Knowledge Construction in the Classroom: Effects of Teaching Children how to Question and how to Explain. *American Education Research Journal, 31*, 338–368.

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

Lock, R.H. (1996). Adapting Mathematics Instruction in the General Education Classroom for Students with Mathematics Disabilities. In LD Forum: Council for Learning Disabilities. Available on www.ldonline.org/ld_indepth/math_skills

Martino, A. M., & Maher, C. A. (1999). Teacher Questioning to Promote Justification and Generalization in Mathematics: What Research Practice Has Taught Us. *The Journal of Mathematical Behaviour, 18*, 53–78.

Matsuda, N., Cohen, W. W., Koedinger, K. R., Stylianides, G., Keiser, V., & Raizada, R. (2010). Tuning Cognitive Tutors into a Platform for Learning-by-Teaching with SimStudent Technology. In *Proceedings of the International Workshop on Adaptation and Personalization in E-B/Learning using Pedagogic Conversational Agents* (APLeC) (pp.20-25), Hawaii.

Matsuda, N., Cohen, W. W., Koedinger, K. R., Keiser, V., Raizada, R., Yarzebinski, E., et al. (2012). Studying the Effect of Tutor Learning using a Teachable Agent that asks the Student Tutor for Explanations. In M. Sugimoto, V. Aleven, Y. S. Chee, & B. F. Manjon (Eds.), *Proceedings of the International Conference on Digital Game and Intelligent Toy Enhanced Learning (DIGITEL 2012)* (pp. 25–32). Los Alamitos: IEEE Computer Society.

McNay, M., & Melville, K. W. (1993). Children's Skill in Making Predictions and Their Understanding of What Predicting Means: A Developmental Study. *Journal of Research in Science Teaching, 30*, 561–577.

Michie, D., Paterson, A. & Hayes-Michie, J. (1989), Learning by teaching, *2nd Scandinavian Conference on Artificial Intelligence '89,* (pp. 307-331). Tampere, Finland, IOS

Mitrovic, A. (2005). The effect of explaining on learning: A case study with a data normalization tutor. *Proceedings of the International Conference on Artificial Intelligence in Education*, (pp. 499-506).

Moreno, R., & Mayer, R. E. (2005). Role of Guidance, Reflection and Interactivity in an Agent-Based Multimedia Game. *Journal of Education and Psychology, 97*, 117–128.

Nichols, D. (1994). Issues in Designing Learning by Teaching Systems. In *In Proceedings of the East–West International Conference on Computer Technologies in Education (EW-ED'94), 1* (pp. 176–181).

Obayashi, F., Shimoda, H., & Yoshikawa, H. (2000). *Construction and Evaluation of CAI System Based on Learning by Teaching to Virtual Student* (pp. 94–99). Orlando: In Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics.

Otero, J., & Graesser, A. C. (2001). PREG: Elements of a Model of Question Asking. *Cognition and Instruction, 19*, 143–175.

Palthepu, S., Greer, J., & McCalla, G. (1991). Learning by Teaching. In *The Proceedings of the International Conference on the Learning Sciences, AACE* (pp. 357–363).

Papert, S. (1980). *Mindstorms: Children, Computers, and Powerful Ideas*. New York: Basic Books.

Pareto, L. (2004). The Squares Family: A Game and Story Based Microworld for Understanding Arithmetic Concepts Designed to Attract Girls. *World Conference on Educational Multimedia, Hypermedia and Telecommunications, Issue, 1*, 1567–1574.

Pareto, L. (2009): Teachable Agents that Learn by Observing Game Playing Behaviour. In *Proceedings of the Workshop on Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education* (pp. 31-40).

Pareto, L., Schwartz, D.L. & Svensson, L. (2009): Learning by Guiding a Teachable Agent to Play an Educational Game. In *Proceedings of the International Conference on Artificial Intelligence in Education,* (pp. 662-664), IOS press.

Pareto, L., Arvemo, T., Dahl, Y., Haake, M., & Gulz, A. (2011). A Teachable-Agent Arithmetic game's Effects on Mathematics Understanding, Attitude and Self-Efficacy. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 247–255). Heidelberg: 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. *Journal of Educational Technology Research and Development, 60*, 723–751.

Penner, D. E., & Klahr, D. (1996). When to Trust the Data: Further Investigations of System Error in a Scientific Reasoning Task. *Memory and Cognition, 24*, 655–668.

Piaget, J. (1952). *The Origins of Intelligence*. Madison: International Universities Press.

Rieber, L. P. (1996). Seriously Considering Play: Designing Interactive Learning Environments Based on the Blending of Microworlds, Simulations, and Games. *Educational Technology Research and Development, 44*, 43–58.

Roscoe, R. D., & Chi, M. T. (2007). Understanding Tutor Learning: Knowledge-Building and Knowledge-Telling in Peer Tutors' Explanations and Questions. *Review of Educational Research, 77*, 534–574.

Ruffman, T., Perner, J., Olson, D. R., & Doherty, M. (1993). Reflecting on Scientific Thinking: Children's Understanding of the Hypothesis-Evidence Relation. *Child Development, 64*, 1617–1636.

Schwartz, D. L., & Arena, D. A. (2009): Choice-based assessments for the digital age. [White paper]. Retrieved from http://aaalab.stanford.edu.

Schwartz, D. L., & Bransford, J. D. (1998). A Time for Telling. *Cognition and Instruction, 16*, 475–522.

Schwartz, D. L., & Martin, T. (2004). Inventing to Prepare for Learning: The Hidden Efficiency of Original Student Production in Statistics Instruction. *Cognition and Instruction, 22*, 129–184.

Schwartz, D. L., Blair, K. P., Biswas, G., Leelawong, K., & Davis, J. (2007). Animations of Thought: Interactivity in the Teachable Agents Paradigm. In R. Lowe & W. Schnotz (Eds.), *Learning With Animation: Research and Implications for Design* (pp. 114–140). UK: Cambridge Univ. Press.

Schwartz, D. L Chase, C. Wagster, J. Okita, S. Roscoe, R. Chin, D. & Biswas, G. (2009): Interactive meta cognition: Monitoring and regulating a teachable agent. In D. J. Hacker, J. Dunlosky, and A. C. Graesser (Eds.)*, Handbook of Metacognition in Education*, 340-358.

Ur, S., & VanLehn, K. (1995). STEPS: A Simulated, Tutorable Physics Student. *Journal of Artificial Intelligence in Education, 6*, 405–437.

Van der Meij, H. (1994). Student Questioning: A Componential Analysis. *Learning and Individual Differences, 6*, 137–161.

Vogel, J. F., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer Gaming and Interactive Simulations for Learning: A Meta-Analysis. *Journal of Educational Computing Research, 34*, 229–243.

Vygotskij, L. (2001). *Tänkande och språk* (eng. Thought and Language), Göteborg: Daidlos

Wong, R. M., Lawson, M. J., & Keeves, J. (2002). The Effects of Self-Explanation Training on students' Problem Solving in High-School Mathematics. *Learning and Instruction, 12*, 233–262.

Zimmerman, C. (2007). The Development of Scientific Thinking Skills in Elementary and Middle School. *Developmental Review, 27*, 172–223.