Int J Artif Intell Educ (2019) 29:89–121
https://doi.org/10.1007/s40593-018-0167-2

ARTICLE

Instructing a Teachable Agent with Low or High Self-Efficacy – Does Similarity Attract?

Betty Tärning[1] · **Annika Silvervarg**[2] · **Agneta Gulz**[1,2] · **Magnus Haake**[1]

Published online: 2 May 2018

Abstract This study examines the effects of teachable agents' expressed self-efficacy on students. A total of 166 students, 10- to 11-years-old, used a teachable agent-based math game focusing on the base-ten number system. By means of data logging and questionnaires, the study compared the effects of high vs. low agent self-efficacy on the students' in-game performance, their own math self-efficacy, and their attitude towards their agent. The study further explored the effects of matching vs. mismatching between student and agent with respect to self-efficacy. Overall, students who interacted with an agent with low self-efficacy performed better than students interacting with an agent with high self-efficacy. This was especially apparent for students who had reported low self-efficacy themselves, who performed on par with students with high self-efficacy when interacting with a digital tutee with low self-efficacy. Furthermore, students with low self-efficacy significantly increased their self-efficacy in the matched condition, i.e. when instructing a teachable agent with low self-efficacy. They also increased their self-efficacy when instructing a teachable agent with high self-efficacy, but to a smaller extent and not significantly. For students with high self-efficacy, a potential corresponding effect on a self-efficacy change due to matching

✉ Betty Tärning
betty.tarning@lucs.lu.se

Annika Silvervarg
annika.silvervarg@liu.se

Agneta Gulz
agneta.gulz@lucs.lu.se; agneta.gulz@liu.se

Magnus Haake
magnus.haake@lucs.lu.se

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

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

may be hidden behind a ceiling effect. As a preliminary conclusion, on the basis of the results of this study, we propose that teachable agents should preferably be designed to have low self-efficacy.

Keywords Self-efficacy · Teachable agent · Similarity attraction · Math game · Educational technology

Introduction

Digital agents are becoming increasingly common in educational software. They can be used to simulate different pedagogical roles, such as teachers, mentors, instructors, coaches, learning companions, and tutees. Once the role is decided, there are a number of design choices that must be made: the agent's age, gender, and ethnicity (indicated through visual and behavior markers), range and level of knowledge, communicative style, etc. These choices have been shown to influence students' performance and learning (Veletsianos 2009; Arroyo et al. 2009), motivation (Plant et al. 2009), and self-efficacy (SE), i.e. the belief in one's capacity to succeed with a task or in a domain (Ebbers 2007).

Furthermore, research shows that design choices can have different impact on different groups of students (Kim and Wei 2011; Arroyo et al. 2011). In other words, evidence recommends against one-size-fits-all solutions when designing digital pedagogical agents.

A corresponding recommendation with respect to real human teachers would be discouraging and, in a sense, meaningless. A human teacher entering a classroom cannot simultaneously be of different ethnicities, have different genders, use several different communicative styles and use a whole set of different ways to provide feedback. For the domain of educational technology, including agent-based educational software the situation is very different. Here a potential strength is precisely that more than one approach, such as more than one design of a digital pedagogical agent, can co-exist in the same software. In educational software lies an inherent potential to meet and cater for variation. This is also the reason why research that contributes to a mapping out of which design choices have more impact than others, and how impacts vary between groups of students, is needed. The more such knowledge that the research community develops, the more useful and powerful future's educational software can become.

The subject area targeted by the instructional software used in our study is mathematics, more specifically place value, frequently identified as a bottleneck in mathematics instruction (Sherman et al. 2015). Mathematics is, furthermore, known as a subject where students show a large variation in performance and towards which many have a negative attitude. It is also an area where (too) many students have little confidence in their own ability to succeed or even make progress, i.e. they have low SE (Bandura 1997). It is therefore of interest to explore possibilities to influence performance and SE in the area of mathematics, and two of the outcomes we focus on in the present study are precisely these: students' in-game performance in math, specifically place value, and a possible change in the students' SE in this domain. The third outcome addressed is students' attitude towards the digital pedagogical agent. The reason for including this is that the attitude towards someone that communicates or represents a certain domain tends to spill over to one's attitude towards the domain as such (Plant et al. 2009; Rosenberg-Kima et al. 2010). If the student likes the digital agent, she will probably also tend to be motivated to like the subject matter.

The agent in the present study takes the role of a *digital tutee* or *teachable agent* (Biswas et al. 2005) in relation to whom the real student takes a teacher role. In this way, teachable agent-based software implements the pedagogical idea that teaching someone else is a good way to learn for oneself, which has been repeatedly demonstrated by researchers (Annis 1983; Papert 1993; Fiorella and Mayer 2013.) It should be pointed out that even though the students here take a teacher role, educational researchers and designers look upon the students as learners and are interested in their learning.

A central question with respect to our study was the following: Which characteristics of a digital tutee have particular impact on students' learning? Previously studied characteristics of teachable agents include visual markers with respect to gender (Silvervarg et al. 2012, 2013) and communicative style with respect to how much the tutee challenges the student (Kirkegaard 2016). Two studies on learning companions that incorporate elements of a digital tutee (Uresti 2000; Uresti and du Boulay 2004) have addressed effects of the agent's competence level on students' learning – something that has been well examined also for digital agent taking other pedagogical roles.

However, the level of competence or knowledge does not lend itself to experimental manipulation for an agent that is strictly a teachable agent. The designer can control only its initial level of knowledge and the rate at which it can learn; once the student, in her role as teacher, starts interacting with the digital tutee, the level of competence is largely out of the designer's hand. If the student teaches the digital tutee well, it learns and makes progress; otherwise, like the student, it flounders. Level of competence and knowledge is in other words not a characteristic that a researcher or designer can experimentally control in a digital tutee.

What *can* be experimentally manipulated, however, is the tutee's *attitude* towards its knowledge and competence, for example whether a digital tutee *believes in its capacity to succeed* in a task or domain, that is its SE. This is what we chose to focus on for our study: We wanted to see if, and how, manipulating the tutee's expression of its SE would influence the student's own learning. Would it matter whether the digital tutee expressed a high or low SE in the domain the student was instructing it on? More specifically, would this influence (i) the real student's in-game performance, (ii) the real student's attitude towards the digital tutee that she taught, (iii) a potential change in the real student's belief in her own capacity to succeed (i.e. her SE) in the mathematic tasks at hand?

As discussed above, previous studies show that the effects from manipulations of agent characteristics often vary between different groups of students. Since the present study manipulates the digital tutee's SE as low or high, it was near at hand to wonder whether students who themselves had high or low SE respectively in this domain (learning and performing in math and problems involving place value) would be differently affected by teaching a high or low SE tutee. Would match or mismatch in SE between tutee and student matter for the three measurements of student in-game performance, potential SE-change and attitude towards their tutee?

The next section develops some central concepts and further examines previous research where agent characteristics have been manipulated and resulted in effects on students' in-game performance, attitude to their agent or SE change. First, we present such studies involving teachable agents/digital tutees and thereafter studies involving pedagogical agents more broadly. The section concludes with a discussion on *similarity attraction* (Newcomb 1956; Byrne and Nelson 1965; Byrne et al. 1967).

Background and Related Work

The Phenomenon and Concept of Self-Efficacy

Self-efficacy (SE) plays an importantly dual role in our study, manipulated with respect to the digital tutee and (mis)matched with the students' SE, and then measured as one of three learning outcomes for the students.

As developed by Bandura (1977), SE refers to subject-specific confidence in one's ability to succeed in this subject. There are some things surrounding the concept that need to be sorted out. First, one needs to hold apart 'a belief that I can succeed in a task' and 'a belief that I can succeed in *learning* to succeed in this task' (Panton et al. 2014). Second, it is important to distinguish SE from more general self-attitudes, such as self-confidence or self-esteem. The subject-specific nature of SE is key; a person can think highly of her ability to perform and make progress in, say, ice hockey but not in programming, or in Spanish but not in math.

Low SE in a given subject area, in the sense of a disbelief that one can succeed in this area, is clearly educationally undesirable; as Bandura (1997) writes, this "assuredly spawns failure" (p. 77). Low SE in an area, namely, is accompanied with setting low aspirations for oneself in the area, being weakly motivated to try work on it, and tending to give up quickly rather than persist on tasks in the area. With this, a self-fulfilling prophecy is easily created; the student will indeed also not succeed in the area.

There is, as well, a relation between low SE in an area to what Dweck (2000) has termed having 'a fixed mindset' rather than 'a growth mindset'. Having a fixed mindset means holding that intelligence or intellectual capacities are innately fixed rather than learnable. Some people are good at math, some are not – and nothing can be done to change this. That is, having a fixed mindset and performing below average, usually equals a low SE. A student who does not perform well in an area and does not believe that making an effort will help her improve (since intellectual abilities are fixed), will also not make an effort, probably not succeed and have her weak beliefs in her own ability to succeed reinforced. To make matters worse, studies show that math teachers – more than teachers in other subjects – tend to use a language that encourages a fixed mindset (e.g. Rattan et al. 2012).

For all that SE and performance in an area are connected – experiences of success in an area is one of the primary factors (Bandura 1997) that promotes increased SE, and people with high SE tend to perform well – the relation between them is not just as simple as it may seem, and for various reasons the two are not always well aligned. First, there is one group of students that tend to overestimate their capacity. Their SE in an area then (repeatedly) misaligns with their actual level of performance or learning. These students are certainly not helped by an increased SE in the area. As Bandura (1997) puts it: "The objective of education is not the production of self-confident fools." (p. 65). Second, and of importance for our study, there are students with a really strong belief that they can perform and make progress in an area, and a corresponding high level of performance and learning. It is not obvious that they have anything to gain by increasing their SE even more – and put otherwise, it is not clear from a pedagogical point of view that it makes sense to increase their already high SE in an area further.

In sum, SE is not something that one for each and every student wishes to increase. This is in contrast to performance. Whatever level of performance a student has in an area, it is meaningful to set a goal of reaching an even higher level of performance. This applies to low-, mid-, and high-performers alike, also including top performers.

Characteristics of Digital Tutees in Relation to Student Learning

A great many studies have compared the effects of having students teach a digital tutee to students learning for themselves, while using the same underlying digital material and tasks. In these studies, no agent characteristics have been varied or evaluated, but it is the very idea of using teachable agent-based software that has been evaluated. The majority of these studies show that teaching a digital tutee can have a clearly positive impact on learning and performance (e.g. Roscoe et al. 2008; Chase et al. 2009; Sjödén et al. 2011, Okita and Schwartz 2013). One study (Pareto et al. 2011) also shows that teaching a digital tutee can affect SE positively. Over nine weeks, third graders who played a math game where they taught a digital tutee showed significantly higher increase in SE compared to students in the control condition who had their regular math classes.

A seminal article by Chase et al. (2009) proposes a set of mechanisms to explain the following educational effect of teaching a digital tutee compared to learn for oneself: students teaching a digital tutee put more effort into the task and spend more time on the activities. They propose that this effect, that they call *the protégé effect*, originates from: a feeling of responsibility on the part of the students; approaching the digital tutee as a socio-cognitive entity and from the possibility to share responsibility for failures with the tutee (even when students are aware that the tutee's weak performance comes by because they have not taught it well).

Knowing, thus, that interacting with and teaching a digital tutee can positively influence students learning and SE, our aim was to dig deeper and explore whether a digital tutee's belief in its capacity to succeed (i.e. its SE) would possibly have any further effects on students' in-game performance, attitude and/or SE. More broadly: could the positive effects from teaching a digital tutee be amplified (or the opposite) with certain design choices?

There are a few previous studies on digital tutees that relate to what we set out to do. Uresti (2000) let students collaborate with a digital learning companion (communicating via text, without embodiment or physical appearance) in the domain of Boolean algebra. There were two types of learning companions: one with a little less knowledge than the student (weak) and one with a little more expertise (strong). In order for the learning companion's performance to increase the student had to teach it. Although the effect was not statistically significant, students interacting with the weak learning companion tended to learn more than the students interacting with the stronger companion.

Uresti and du Boulay (2004) conducted a follow-up study with similarly strong vs. weak learning companions under two conditions. In one condition, students were regularly reminded by the system to collaborate with the learning companion and encouraged to work for a high score; in the other condition they were reminded only a few times that it could be good to collaborate with the learning companion. No statistically significant differences in learning were found between the four conditions, but the learning behavior varied between groups. Students that collaborated with and guided a weak companion, and were reminded regularly to collaborate, spent more time teaching their companion and worked harder than the students in the other experimental conditions.

In studies by Kirkegaard (2016) middle-school students instructed a digital tutee in the area of history. The tutee had one of two alternative communicative styles. Either it was a more traditional, compliant tutee that accepted everything the student (as teacher) proposed, or it was a more independently minded tutee, who would now and again

question or challenge the student's answers or explanations. The sample was balanced with respect to students' level of SE in history. Results were that students with high SE performed better when teaching the 'challenging' agent, whereas students with low SE performed better with the traditional teachable agent.

In a separate pilot study, Kirkegaard et al. (2014) let students teach history to a digital tutee, with the tutee designed to look androgynous. After two game sessions the students were asked how they perceived the agent: "*Absolutely like a girl (boy)*", "*A little like a girl (boy)*", or "*Neither like a girl nor a boy*". They were then asked to look at a list and circle all the adjectives – positive and negative ones – they associated with the tutee. When the digital tutee was perceived as a boy it was assigned more positive words: when perceived as a girl, it was assigned more negative words.

Silvervarg et al. (2013) made use of three visually gendered digital tutees, one girl-like, one boy-like and one androgynous. Each of 108 students interacted with two of the three agents during two 45-min sessions. Girls had a more positive attitude towards the androgynous tutee than towards the two other tutees, whereas boys equally favored the boy-like and the androgynous tutees over the girl-like tutee.

Next, we turn to studies on characteristics in other kinds of pedagogical agents than digital tutees and how they can affect students' SE-change, in-game performance, and attitude towards the agent.

Characteristics of Pedagogical Agents in Relation to Students' SE-Change, In-Game Performance, and Attitude Towards Their Agent

Baylor and Kim have conducted a series of studies exploring whether certain characteristics in pedagogical agents can affect *SE-change* in students. One study (Baylor and Kim 2004) found that pedagogical agents perceived by learners as less intelligent, had a more positive effect on the learners' SE growth than agents perceived as more intelligent; another (Kim and Baylor 2006) found that students whose learning companion had low competence increased their SE more than students collaborating with a learning companion with high competence. Yet another (Baylor and Kim 2005) found that students interacting with an agent who offered verbal encouragement increased their SE more compared to students who interacted with an agent that provided less verbal encouragement. Finally, Rosenberg-Kima et al. (2008) studied possible effects of pedagogical agents' perceived gender, age and 'coolness' (equated with having a cool hairstyle and cool clothes) on female students' SE in engineering related fields. Students interacting with a young and cool agent had higher SE at the end of the experiment.

Numerous studies have examined agent characteristics with respect to *students' performance.* Lee et al. (2007) found that a digital learning companion expressing empathy and providing encouraging feedback lead to higher performance (measured as recall) than an emotionally neutral digital co-learner that did not provide encouraging feedback. Veletsianos (2009) found, along a similar line, that a digital tutor who made pauses and varied its voice loudness, led to a better recall of the material learned compared to a less expressive digital tutor. Wang et al. (2008) found that a more polite agent had a more positive impact on student's learning outcomes than a less polite agent.

Mayer and DaPra (2012) showed that an agent using social cues in the form of gestures, facial expressions and eye-gaze led to better learning outcomes than the same agent lacking these social cues. Likewise, Johnson et al. (2015) found that an agent

who pointed agent compared to an agent that did not point had a beneficial impact on learning outcomes – for students with low prior knowledge.

Fewer studies have examined agent characteristics in relation to students' *attitude towards the agent.* That said, some have addressed attitude together with one or another of the other student outcomes addressed in our study.

Kim et al. (2006) compared effects of four kinds of peer agents: (i) low-competence peers with a proactive interaction style, (ii) low-competence peers with a responsive interaction style, (iii) high-competence peers with a proactive interaction style, or (iv) high-competence peers with a responsive interaction style. Students interacted with their peer agent in order to learn about instructional planning. Those who interacted with high-competence agents were better at applying what they had learned and showed a more positive attitude towards their peer agents, while those who interacted with a low-competence agent, on the other hand, showed an increase in SE. The authors speculate that students evaluated their competence higher when they compared themselves to a pedagogical agent with lower competence and thus felt more confident. However, it was not only the agent's competence that influenced SE. Students who collaborated with a responsive, but not with a non-responsive agent, showed a significant increase in their SE in the domain.

Baylor and Kim (2005) studied all three outcomes that we address in our study: students' performance, attitude towards the agent, and potential SE change. They used three types of agents. The 'expert' and 'mentor' agents had more expertise than the 'motivator' agent, whereas the 'motivator' and 'mentor' agents were more motivational than the 'expert' agent. Results showed that the 'expert' and 'mentor' agents led to improved learning and a more positive attitude towards the agent. The motivator and mentor agents, who were more like coaches, led to an increase in SE for the students.

Ebbers (2007) finally, compared the effects of pedagogical agents demonstrating either mastery or good coping strategies. The first type of agent showed positive attitudes towards the task and learned the requisite information with ease, enthusiasm and confidence. The second type of agent learned with more difficulty and expressed discouragement but did not give up – succeeding in the end. This second type of agent had more positive impact on students' SE and attitude towards them. The 'mastery' agent, though, had more positive effect on student learning.

Similarity Attraction

Human beings often tend to like people they perceive as similar to themselves; a phenomenon known as *similarity attraction* (Newcomb 1956; Byrne and Nelson 1965; Byrne et al. 1967; Nass and Lee 2001). Similarity attraction, thus, provides a potential mechanism for influencing attitudes towards others.

In addition, similarity can be the basis for increasing a learner's SE in a domain. As mentioned earlier, a learner's SE on a domain tends to be affected over time by her own (non)success in the domain. Experiencing success tends to boost one's SE. But another mechanism is what is called vicarious experience: Observing *someone else* succeed in the domain can generate an expectation in the *observer* that she too can succeed (Bandura 1977). However, this is more likely to work if the learner sees herself as sufficient similar to whom she is observing. Bandura (1997) claims that three characteristics are central to these similarity judgments: gender, ethnicity, and competence. Someone who is similar to me in one or several of these respects also has the best chance of influencing my belief in my own

ability. Schunk (1987) argues along similar lines but focuses only on similarity of competence. Especially, Schunk argues, this applies for unfamiliar tasks where the learner has little information to base her SE judgements on.

Previous Research on Matching Vs. Not Matching Characteristics Between Actors in an Educational Context

Similarity attraction does not only apply between humans. Reeves and Nass (1996) provide considerable evidence for the so-called *Media Equation Hypothesis*: the way humans treat media, including digital media, parallels how they treat their fellow human beings. Similarity attraction mechanisms have been shown in a number of studies. Humans tend to be more positive towards computers that are similar to themselves more than computers that are dissimilar to themselves (Nass et al. 1995; Nass and Lee 2000).

In this section we present some studies from educational context and pedagogical agents, where gender, ethnicity and competence, the characteristics Bandura (1997) lifts forth, have been matched or mismatched between pedagogical agent and student.

Plant et al. (2009) found that a female compared to a male pedagogical agent had a larger positive influence on female students' attitude towards engineering-related fields, as well as on their SE towards these fields. Behrend and Thompson (2011) on the other hand, found no similarity attraction effects with respect to gender between participants and their digital trainer that supported them in an Excel training activity.

In an early study Lee and Nass (1998) found that people rated agents of the same apparent ethnicity as themselves as more attractive and more trustworthy than agents of different ethnicity than themselves. Pratt et al. (2007) found that learners changed their opinion to be consistent with agent advice to a higher degree when matched with a same-ethnicity agent.

Rosenberg-Kima et al. (2010) explored the potential for digital agents to encourage female students – white and black – to pursue an engineering career. When the agent's ethnicity matched the student's, the student expressed a more positive attitude towards and more interest in engineering. Behrend and Thompson (2011) found no matching effects for ethnicity in the study where participants had a digital trainer that supported them in an Excel activity. However, they *did* find similarity-attraction effects with respect to the style of providing feedback: directive versus non-directive. When feedback styles of student and agent were matched, students showed an increase in declarative knowledge and a more positive attitude (measured as affective responses) to the agent.

Finally, some studies have examined matched or mismatched *levels of competency,* the third of the characteristics lifted forth by Bandura (1997). Hietala and Niemirepo (1998) looked for similarity effects for competence in math; Kim (2007) did the same for competence in instructional planning. Both studies showed that low-performing students benefited most, in terms of performance, when interacting with a low-performing agent and high-performing students when interacting with a high-performing agent.

Hietala and Niemirepo (1998) measured attitude towards the learning companion as 'preferred choice' since the students could freely change their learning companion. For this measurement results were that high-performing students over time chose increasingly to collaborate with a high-performing companion and that low-performing student over time chose increasingly to collaborate with a low-performing companion.

However, Hietala and Niemirepo (1998) actually varied *a combination of two characteristics* in their learning companions. The high-performing companion was not merely

high-performing but also expressed certainty in its suggestions and took command, whereas the low-performing companion was not merely low-performing but also less certain, expressing itself more hesitating. For example, the high-performing agent might say "*The answer is x = 5 and I know it's right.*" while the low-performing agent, for example, might say "*I suggest x = 5 as the answer but I might be wrong.*" (Hietala and Niemirepo 1998, p. 182). In other words, there is a high-performing companion with high SE versus a low-performing companion with low SE – which, as the authors themselves conclude, makes it hard to "dissociate the two factors: the actual level of expertise and the way the agent expresses itself" (Hietala and Niemirepo 1998, p. 191).

Matching Vs. Mismatching Self-Efficacy Between Student and Digital Tutee

Our study explores potential effects of matching vs. mismatching SE between a student and the digital tutee she is teaching. This sets it apart from previous related research on (mis)matching characteristics, where the digital agent is most often taking the role of teacher, mentor, or coach, and the goal is to see how much a match assists in making the agent a better liked, more powerful role model. But where the student becomes the teacher and the agent the student, who is meant to be role modeling whom? In effect, role modeling and observational learning mechanisms are less straightforward in this case. A digital tutee or teachable agent becomes a reflection of the student's learning and performance: in some sense, the student is observing herself. The effects of a match or mismatch in SE between student and digital tutee – on the student's in-game performance, potential SE change and attitude toward the tutee – become difficult to predict.

Research Questions and Predictions

The main purpose of in this study was to investigate the effects on students of manipulating a digital tutee's SE. We looked specifically for effects on students' in-game performance, SE, and attitude towards their digital tutee. In addition, we examined whether it mattered, for the outcomes mentioned, whether the student's and the digital tutee's SE – low or high – were matched or mismatched.

Given the novelty of the study – to our knowledge, no previous studies have manipulated a digital tutee's SE in this way – the study was essentially exploratory.

How Do Students Respond to a Digital Tutee with High Vs. Low Self-Efficacy?

Q1: How do students respond to a digital tutee with high versus *low SE?*

- *Q1a: How will the digital tutee's SE (high/low) affect students' attitude toward the tutee?*
- *Q1b: How will the digital tutee's SE (high/low) affect a potential change in students' own SE?*
- *Q1c: How will a digital tutee's SE (high/low) affect students' in-game performance?*

According to previous studies we know that digital tutees as such, regardless of whether or how they express their SE, can – when compared to an equivalent learning

situation without teaching a digital tutee – have positive effects on students' performance as well as boost students' SE. But this does not provide any basis to make predictions in relation to the three sub-questions above.

Does Match Vs. Mismatch of SE (High/Low) Between Student and Digital Tutee Affect Student Responses?

The same three research questions as above are repeated with respect to SE match or mismatch between the digital tutee and the student.

> *Q2: How do students with high or low SE (measured at the pre-test) respond to a digital tutee with high or low SE?*

- *Q2a: Does match/mismatch in SE between student and digital tutee have effects on students' attitude to the digital tutee?*

According to the Similarity Attraction Hypothesis people tend to like people they perceive as similar to themselves with respect to a variety of personal characteristics; while the Media Equation Hypothesis claims that this holds for artefactual agents as well (Reeves and Nass 1996). Therefore, we hypothesized that students who taught a digital tutee who appeared similar to them in terms of SE would show a more positive attitude towards the tutee compared to students who taught a digital tutee that appeared dissimilar to them in terms of SE.

- *Q2b: Does match/mismatch in SE between student and digital tutee have effects on students' potential SE change?*
- *Q2c: Does match/mismatch in SE between student and digital tutee have effects on students' in-game performance?*

Some studies have shown (cf. section "Previous research on matching vs. not matching characteristics between actors in an educational context") that matching the *level of competence* between *student and digital companion* tends to be positive for the students' performance. High-performing students tend to perform better when collaborating with a high-performing digital companion, and vice versa for low-performing students and low-performing digital companions. However, this does not provide us with firm basis to make predictions with respect to match/non-match in SE – neither for digital companions, nor for digital tutees.

Method

The study comprised one pre-test session, seven game-playing sessions, and one post-test session – all sessions lasting 30–40 min. During the pre-test session students took a math test and filled out an SE questionnaire (Appendix 2). Students completed the same SE questionnaire at the very end of the study, with an additional questionnaire probing their experiences and their attitude towards the tutee (Appendix 1).

Participants

A total of 166 fourth graders (83 girls and 83 boys) took part, recruited from nine classes in four schools in Southern Sweden in areas with median to low socio-economic status. Students were randomly assigned to one of the two conditions: teaching a digital tutee who expressed high SE or one who expressed low SE. In preparing data for analyses, 24 participants were excluded due to missing data or poor attendance, leaving 142. These were categorized as low or high SE based on results on the pre-test SE questionnaire. Approximately two fifths were assigned to the low SE group and two-fifths to the high SE group. The rational for excluding the middle fifth was to increase the contrast when comparing low and high SE students.

The result was two data sets: one with 142 participants for addressing Q1 and one with 113 participants for addressing Q2. The latter was divided into four groups based on (mis)match of SE (see Table 1).

Material

The Math Game

The math game was developed by Lena Pareto (2014). It targets the basic addition/subtraction skills of place and value, including carrying and borrowing, with squares and boxes as spatial representations of numbers. For example, ten red squares can be packed into one orange box, ten orange boxes into one yellow box, (representing carry-overs during addition). Sub-games tackle different kinds of mathematical problems: addition up to 10, up to 100 and up to 1000 and subtraction up to 10, up to 100 and up to 1000. For this study, students played only the addition sub-games. They were encouraged to start with addition up to 10 and progress from there.

All sub-games use the same playing board and cards depicting various constellations of squares and boxes, representing different numbers. Each player begins by having a set of cards. They take turns choosing one of their cards, the content of which is added to (or subtracted from) the board. A star is awarded for each carry-over, and the player with the most stars at the end wins the game.

Figure 2 depicts a situation where there are six yellow boxes, three orange boxes, and no red squares on the board, which represents the number 630. The student is competing against the computer and has chosen a card representing 79. Playing this results in the calculation $630 + 79 = 709$, yielding one carry-over (from tens to

Table 1 Descriptive statistics of student SE based on the SE pre-questionnaire (min = 7, max = 35), and distributed in the four participant groups separated on agent x student SE

Self-efficacy agent x student	N	n (Girls/Boys)	Range	Median
low.low	27	19 / 8	7–25	23
low.high	30	14 / 16	29–35	32
high.low	28	16 / 12	9–25	22.5
high.high	28	10 / 18	29–35	31

hundreds). The student is to receive one point. The digital tutee, Lo (see Fig. 1), has posed a question to the student about her choice.

Lo is designed to look androgynous, allowing students to form their own opinions on gender.[1] Silvervarg et al. (2013) report positive educational effects for visual androgyny.

Lo knows nothing about the base-ten system at the start. Her knowledge – based on the digital system's knowledge domain (Pareto 2014) – develops entirely on the basis of what the student teaches her: if taught wrong, she learns wrong. She participates through one of three game modes.

In 'Observe' mode, Lo watches the student play, and learns by observation and by posing questions to the student. These might address the student's recent actions or raise more abstract, conceptual issues like "*How many red squares are there in a yellow box?*". All answers are of multiple-choice format, with four, sometimes three, alternatives for the student to select from (see Fig. 2). For each question, there is one correct answer, two (or one) incorrect answers, and one "*I don't know.*" option. The correct answer to the question posed in Fig. 2 is "*Yes, 1 point.*" – which is what the student has selected.

In 'Try and be guided' mode, Lo proposes cards based on what she has learnt in 'Observe'-mode. The student offers feedback by accepting or rejecting the proposed card. If rejecting a card, the student must exchange it for one she finds to be a better choice. In this way, this is both an opportunity for the student to see what Lo has learned so far, and for revising Lo's knowledge. Multiple-choice questions are included in this mode as well. They are of the same type as in 'Observe' mode but also include questions that ask the student to explain why the card the student chose is better than the card that the tutee proposed.

In 'On her own' mode, the student watches Lo play on her own against the computer (at any of five competence levels) or another digital tutee. This gives the student an opportunity to evaluate Lo's performance (which reflects how well she has taught Lo). For further details, see (Pareto 2014).

The Chat

In addition to the scripted 'conversation' via multiple-choice questions, the present version of the game also includes a chat (Silvervarg and Jönsson 2011). The chat window appears after each round in which Lo has been active and closes automatically after one minute. The idea behind the chat is to give students the opportunity to strengthen their relationship with the tutee. The chat is open ended, allowing students to take a break from the game, if they so wish, to talk for example about music or sports, what we call 'off-task topics', i.e. topics that does not directly relate to the game.

The chat is also the primary channel in which the students receive feedback from the digital tutee, including Lo's reflections on the result in the just completed game session and on her own performance and learning – which is where the tutee's SE with respect to the math game and its challenges is expressed (Fig. 3).

Before the chat function was added, students could only receive indirect feedback on their teaching; namely from observing the tutee's actions in the 'Try and be guided' or 'On her own' modes. In the 'On her own' mode the tutee competes with the computer

[1] In the chat, a student might ask about the tutee's gender; in that case Lo answers "*I'm a girl*". We therefore refer to Lo as *she* in this paper.

Fig. 1 The digital tutee Lo

itself (which can be set at five different competence levels), and here it is possible to evaluate how well the tutee has learnt the domain targeted by the game. However, this kind of feedback on the success of their teaching the tutee was infrequent, entered late in the process and only offered information on a very general level about the progress of the tutee. The chat, in contrast, provides more frequent and explicit feedback.

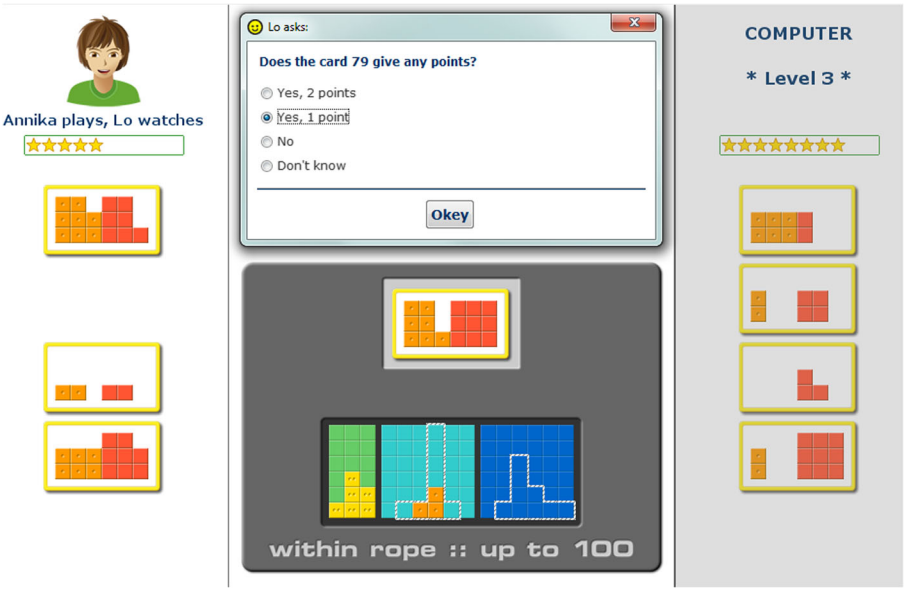

Fig. 2 With the game in observe mode, the digital tutee asks the question "*Does the card 79 give any points?*" The answering options are: "*Yes, 2 points.*", "*Yes, 1 point.*", "*No.*" and "*I don't know*"

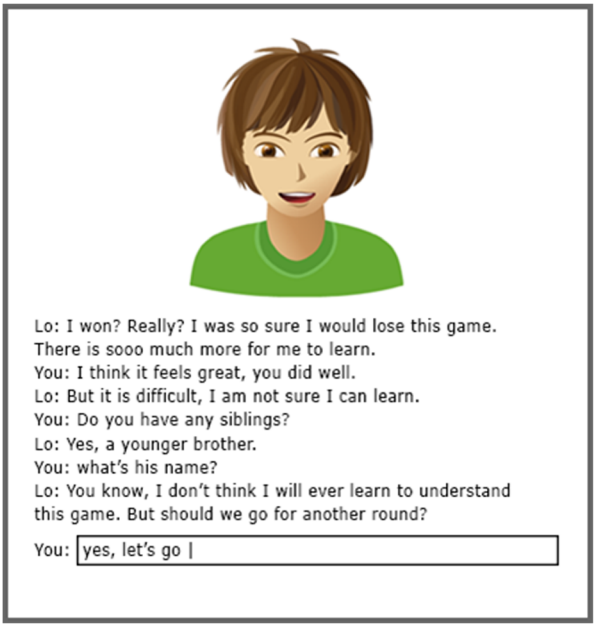

Fig. 3 An example from chatting with Lo when expressing high SE

For this study, we manipulated Lo's feedback in the chat to reflect either high or low SE with respect to her performance and ability to learn to successfully play the math game. Lo always begin a chat by reviewing the result of the previous round (victory, defeat, or draw) saying, for example:

"*I got pretty bad cards now in this last round but I still won. I really play this game brilliantly.*" (high SE upon winning).

"*Wow, we won, did we? Yet I feel so uncertain how to play this game well.*" (low SE upon winning).

"*We didn't win, but that was just bad luck. I at least feel very certain of the game and how it is played.*" (high SE upon loosing).

"*We lost… But that might not be so strange, it feels like I don't remember anything of what you just taught me.*" (low SE upon loosing).

All of Lo's opening comments were pilot tested using 22 fourth graders from a school not participating in the study. They were asked to read the comments and evaluate whether they sounded confident, not confident, or neither. Their ratings resulted in the removal of a few comments and slight modifications of others, yielding a set of 136 comments, 68 reflecting high SE and 68 reflecting low SE.

The comments were adapted as needed to each of the game modes, 'Observe', 'Try and be guided' and 'On her own'. In 'Observe' and 'On her own' modes, Lo talks in first-person singular, for example: "*I'm learning the math game rules slowly. I'm not*

such a brilliant student." (expressing low SE). In 'Try and be guided' mode, she uses both first-person singular and plural, for example: "*That's great! I was sure that we were going to win. I think we played really good.*" (expressing high SE). The subtle changes in pronouns reflect whether Lo is cooperating with the student ('Observe' and 'Try and be guided' modes) or working on her own ('On her own' mode).

Each comment started with a reflection of the actual outcome of the previously played round (ending in victory, defeat, or score even), for example: "*Awesome, we won!*" (high SE) or "*Wow, did we really win?*" (low SE). Thereafter followed a sentence on 'game play', 'learning' or 'knowledge'.

A 'game play' sentence reflects Lo's 'thought' on the game performance "*That's awesome. We won since we choose the best cards the whole time.*" (high SE) or "*Nice to win, but I don't think we played very well this time.*" (low SE). A 'learning' sentence reflects Lo's 'thoughts' on her own learning during the past round "*As expected, I didn't learn much this round. It's too hard for me with tens and hundreds and stuff.*" (low SE) or "*How could we lose? I have learned so many things about how to play this game well.*" (high SE). Finally, a 'knowledge' sentence reflects Lo's 'thoughts' about her general knowledge and learning with respect to math and the math game. "*What! Did we play draw?! I was completely sure that we were going to win. I feel like I know everything about how to play this game.*" (high SE) or "*A draw… Maybe that was good since it feels like I still would need to learn so much more and it is so difficult.*" (low SE).

The chat always ended with Lo presenting her thoughts about the upcoming game, for example: "*I have a feeling that the next round will go really well. Let's play!*" (high SE) or "*It doesn't seem like I understand much really, but let's play another round.*" (low SE).

Measurements

In order to measure the three things that we were focusing on – students' attitude towards their digital tutee, possible increase of students' SE and students' in-game performance – we analyzed the games' data logs together with the questionnaires.

Students' Attitudes Towards their Digital Tutee

The questionnaire on the students' experiences and opinions (Appendix 1) contained three questions targeting students' attitude towards their digital tutee.

"*How has it been to instruct Lo?*"

"*How has it been to chat with Lo?*"

"*Would you like to continue to instruct Lo?*"

The first two questions were accompanied by a five-place Likert scale from 1 = 'very boring' to 5 = 'really fun'. The third offered three options: 'yes', 'no' and 'maybe'. The analysis was performed both for the individual questions and for a total score (*Range* = [3, 15]) of the three questions.

In addition, this questionnaire contained a set of questions about Lo and Lo's knowledge/competence, used in another analysis and another paper, except one where

answers were used to establish that the manipulation of Lo as having high or low SE was successful in terms of the students' judgements of this.

The 'opinions and experiences' questionnaire was distributed at the end of the intervention as a post-test questionnaire.

Self-Efficacy

The pre- and post-test SE questionnaire (Appendix 2) was based on Bandura et al. (1996); adapted for this study and translated into Swedish.

Seven items targeting the students' SE with regard to the place-value system are included. The items line up beneath each other with the same starting sentence: "How good are you at solving these types of tasks?"

Item one to five are calculation tasks such as '1136 + 346', whereas item six and seven are about the place value system, for example: "Which number has the largest value in 6275?"). All items are graded in five steps from "Not good at all" to "Very good at" (see Appendix 2), making up a five-level Likert scale equivalent to the one in the 'opinions and experiences' questionnaire.

The SE score for each student was calculated as the sum over all items, resulting in a value ranging from 7 to 35. SE-change was calculated by subtracting the score on the pre-test from the score on the post-test, providing a theoretical range from −28 to 28.

In-Game Performance

Student in-game performance was determined from how well students answered Lo's in-game questions (of which there were three in each game session) designed to reveal how well the students understood place value and the base-ten number system. Example questions are: "*How many orange square boxes are there in the 2 yellow square boxes on the game board?*" and "*How many red square boxes are needed to fill a yellow square box?*"

An in-game performance value was calculated as the percentage of correct answers in relation to incorrect answers and then standardized (formula: [correct answers − incorrect answers +100]/2). A value of 100 means that the student answered all the questions correct, 0 means that all questions were answered incorrectly, and 50 means that the student answered equally many questions correct as incorrect. Pareto (2014) showed that in-game performance in this math game correlated well with standard pen-and-paper tests on the place-value system.

Procedure

The study comprised nine sessions over seven weeks: one pre-test session, seven game-playing sessions, and one post-test session. At the pre-test session, the participants were asked to fill out the SE questionnaire and also to take a math-test on the computer.[2] During the seven game-playing sessions, the participants

[2] The math pre-test was part of a parallel study (Tärning et al. 2017). It was used to identify high- and low-performing students who were targeted in that study.

played the math game and taught their digital tutee. At the post-test session, the participants were once again asked to fill out a SE questionnaire as well as the opinion and experiences questionnaire.

The Pre-Test Session

In their respective home classroom, the students were introduced to the study and the researchers. Thereafter the students were asked to individually take a math test regarding the base-ten system on a computer and thereafter to fill out the pre-questionnaire targeting SE with regard to math and specifically the base-ten system.

The SE pre-questionnaire was used to calculate an SE score for each student, who was then assigned to one of the two experimental conditions – high-SE agent and low-SE agent – for the game sessions, balancing for student SE and gender.

The Seven Game-Playing Sessions

Students started out playing the game on their own without Lo present. As their familiarity with the game increased, they were asked to start instructing Lo. The game-playing sessions presented mathematical content with increasing difficulty. The level of difficulty was partly controlled in that sub-games using numbers in the 1000 were not available during the first three game-sessions but first at session four. Otherwise, we did not control what sub-games they chose to play. Two experimenters were always present at each game-session in order to help the students with technical issues when necessary.

The Post-Session

At this session, the students were asked to fill out the same SE questionnaire as in the pre-test session (Appendix 2). They were also asked to fill out the opinions and experiences questionnaire (Appendix 1). At the end of the post-session one of the researchers debriefed the students about the two versions of Lo and the study in general. Students were thanked for their participation and given a piece of candy.

Results

Statistical analysis was performed using *R* version 3.2.4 (R Core Team 2016) at alpha level 0.05. All effect sizes were interpreted using the guidelines from Cohen (1988). In the case of multiple comparisons, Holm corrections were used to adjust for family-wise error rates. Analysis of the SE manipulation (Q1) was done on the full dataset of 142 participants. Analysis of match-mismatch in SE (Q2) was done on the reduced dataset of 113 participants, excluding one fifth of the participants categorized as middle SE as noted in section "Participants". This second dataset was divided into four groups of agent vs. student low/high SE; self-depending on the student's and agent's SE: high/high, high/low, low/high, and low/low (see Table 1).

Validation of the Experimental Manipulation

To validate the SE manipulation, we used the sixth question in the attitude questionnaire: "What do you think about Lo's confidence in math?" Students could choose from 1 'really uncertain' to 5 'really confident'. A Mann-Whitney's U-test showed a statistically significant difference with medium to large effect size ($Z = 4.56$, $p < .001$, $r = .38$) between the condition with the low-SE tutee ($Mdn = 2$) and that with the high-SE tutee ($Mdn = 4$). Thus, the result supports the intended manipulation, in that the students perceived the manipulation of the digital tutee's SE the way it was intended.

Students' Responses to their High/Low Self-Efficacy Digital Tutee

Previous research relating to our first research questions (Q1a–Q1c), points in different directions. We therefore felt that we could not make any predictions.

Students' Attitude Towards Their Digital Tutee (Q1a)

To determine whether the digital tutee's SE affected students' attitude towards the tutee, we compared scores for question 4, 7, and 8 from the questionnaire regarding students' experiences and opinions (section "Students'attitudes towards their digital tutee"); Question 4: "How has it been to instruct Lo?" Question 7: "How has it been to chat with Lo?", and Question 8: "Would you like to continue to instruct Lo?" Each question, in the form of a 5-level Likert item, was analyzed using a non-parametric Mann-Whitney's U-test. Results showed a marginally significant effect of the digital tutee's SE for Question 8 only (Table 2). Overall, we found no evidence that the digital tutee's SE affected students' attitude towards her.

Students' Self-Efficacy Change (Q1b)

To determine whether the digital tutee's SE affected students' own SE, we compared students' pre- and post-testing SE scores (section "Self-efficacy"). A two sampled *t*-test revealed no statistically significant difference on SE increase between the two student

Table 2 Medians (*Mdn*) for question 4, 7, and 8 (summarized and one by one) in the attitude questionnaire, addressing attitude towards the digital tutee with regard to the SE traits of the same tutee; comparisons by Mann Whitney's U-test (*W*)

	Agent self-efficacy			
	low: *Mdn*	high: *Mdn*	*W*	*p*
Question 4 + 7 + 8	10	10	2447	0.76
Question 4	4	4	2816	0.20
Question 7	5	5	2420	0.65
Question 8	1	1	2122	0.066 .
Sample size	71	71		

Significance levels: 0.1 * 0.05 ** 0.01 *** 0.001

groups teaching a digital tutee of low ($M = 0.76$, $SD = 3.55$) vs. high ($M = 0.93$, $SD = 3.61$) SE ($t(140) = -0.282$, $p = .78$).

Students' In-Game Performance (Q1c)

To determine whether the digital tutees' SE level affected students' in-game performance, we analyzed students' in-game performance (section "In-game performance"). A two sample *t*-test found a statistically significant difference on in-game performance with a small to medium effect size ($t(140) = -2.51$, $p = .013$, Cohen's $d = 0.42$) with regard to the SE level (low: $M = 57.0$, $SD = 13.7$; high: $M = 50.6$, $SD = 16.6$) of the digital tutee. The result suggests that teaching a digital tutee with low SE enhanced in-game performance.

Match Vs. Mismatch Between Student and Digital Tutee Self-Efficacy

For the next three research issues (Q2a – Q2c), the *Media Equation Theory* (Reeves and Nass 1996) suggest a possible matching effect for SE with regard to attitude (Q2a). For the other two questions (Q2b and Q2c), we did not make any predictions.

These three research questions address a comparison between the four matched vs. mismatched agent x student SE groups; followed by matched-mismatched analyses for each of the two pairs of low and high SE student groups respectively.

In order to avoid ambiguity in the comparisons of low- and high SE student groups while securing a sufficient power for the statistical analyses, 29 mid-scoring SE questionnaire students were excluded (corresponding to one fifth of the students). The resulting data set consisted of 113 participants (section "Participants").

Self-Efficacy (Mis)Match and Students' Attitude Towards Their Digital Tutee (Q2a)

Would a match or mismatch between students' and their digital tutee's SE affect the students' attitude, as measured by the total score (Range = [3, 15]) of the three attitude questions in the questionnaire (Question 4, 7, & 8), also used in the analysis of research question Q1a (above). The scores on the separate questions (Question 4, 7, & 8) were typically skewed and the aggregate score did not comply to a normal distribution, advocating non-parametric statistical methods for the analysis.

A comparison of the two matched vs. mismatched agent x student SE conditions (Table 3) revealed no significant effect using a Mann-Whitney's U test ($W = 1824$, $p = .18$).

Next, the match-mismatch analyses were repeated for low and high SE students respectively (Table 3) using Mann Whitney's U tests with *p*-values adjusted for twofold comparison by means of Holm correction. Neither of the two low and high SE student groups revealed any significant effects between matched and mismatched agent x student SE conditions (low SE students: $W = 454$, $p = .38$; high SE students: $W = 376$, $p = .49$).

Before the analyses of the total attitude scores, each of the three questions (questionnaire attitude items 4, 7, & 8) were analyzed separately. All agent x student SE combinations turned out more or less the same with regard to the three questions and there was no evidence of significant effects on attitude with regard to the different match-mismatch contrasts (Table 4).

Table 3 Test of normality (Shapiro-Wilk), descriptive statistics (n, Median, & Range), and comparison tests (Mann-Whitney's U test) for research question Q 2a evaluating attitude effects

		Shapiro-Wilk					Mann-Whitney	
		W	*P*	n	Median	Range	*W*	*p*
All students	matched	0.863	< .001	55	10	4–13	1824	.18
	mismatched	0.920	< .001	58	10	4–13		
Low self-efficacy students	matched	0.944	.140	27	10	6–13	454	.38[a]
	mismatched	0.897	.001	28	9.5	7–11		
High self-efficacy students	matched	0.891	.005	30	10	4–13	376	.49[a]
	mismatched	0.802	< .001	28	10	4–13		

[a] *p*-values adjusted by means of Holm correction

Taken together, our prediction with respect to research question Q2a was not supported; students who taught a digital tutee that was similar to them in terms of SE did not show a more positive attitude towards their tutee compared to students who taught a digital tutee that appeared dissimilar to them in terms of SE.

Self-Efficacy (Mis)Match and Students' Self-Efficacy Change (Q2b)

Next, we explored whether the match or mismatch of digital tutee's and student's SE affected students' subsequent SE, we compared students' pre- and post-testing SE scores (section "Self-efficacy") The dataset of 113 was divided into matching vs. mismatching subgroups (Fig. 4, left). SE change scores for the mismatch group showed a non-normal distribution (Shapiro-Wilk: $W = 0.947$, $p = .013$) advocating use of non-parametric statistical methods. A Mann-Whitney's U test was then used to evaluate the matched group ($n = 55$, *Median* = 1, *Range* = [−7, 11]) against the mismatched group ($n = 58$, *Median* = 0, *Range* = [−6, 12]). This revealed a less-than-significant trending effect ($W = 1848$, $p = .14$). At the same time, two one sample Mann-Whitney's U tests (Holm corrected for multiple measurements) revealed a significant positive effect of SE increase for the matched agent-student SE group ($V = 806$, $p = .020$), but not so for the mismatched group ($V = 781$, $p = .27$).

Table 4 Medians (*Mdn*) and Range (*Rng*) for the three questions (questionnaire item 4, 7, and 8), addressing attitude towards the agent (digital tutee) with regard to the four agent x student SE combinations

	Medians & Range (agent x student self-efficacy)							
	low x low		high x low		low x high		high x high	
	Mdn	*Rng*	*Mdn*	*Rng*	*Mdn*	*Rng*	*Mdn*	*Rng*
Question 4	4	2–5	4	1–5	4	1–5	4	1–5
Question 7	5	2–5	5	2–5	4.5	1–5	5	1–5
Question 8	2	1–3	2	1–3	1	1–3	1	1–3
Sample size	27		28		30		28	

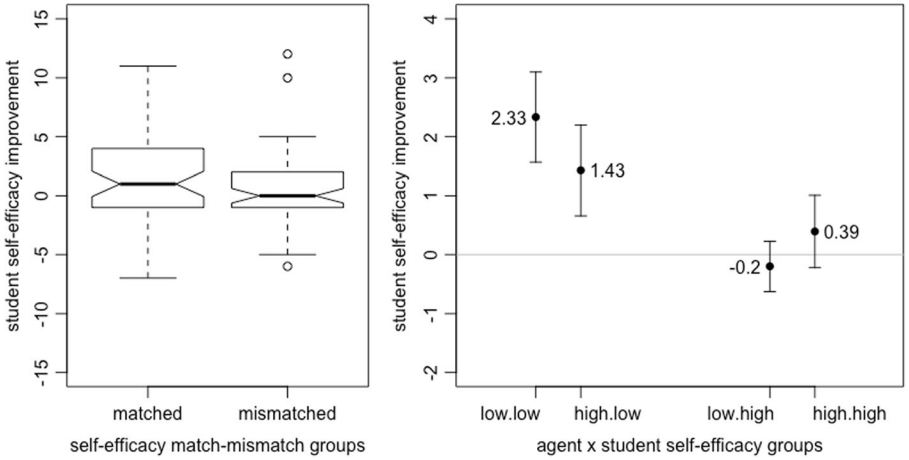

Fig. 4 Left: boxplot of SE improvement for matched vs. mismatched tutee x student SE pairings; right: SE improvement means and standard errors for matched vs. mismatched tutee x student SE pairings separated on low and high SE student groups

Diving into the low- and high-ST students groups separately (Fig. 4, right) reveals the patterns behind the overall less-than-significant trending effect between the matched and mismatched groups and allows certain observations for each of these two student groups.

Notably, separated out, the two student groups show normal distribution and homogeneity of variance, allowing use parametric statistics. The following observations were made for these two student groups.

(1) Both the low and high SE student groups show higher average SE improvements where the agent's SE matches (Fig. 4, right), though the differences are, again, less than significant as evaluated by two *t*-tests (low-SE students: $t(53) = 0.832$, $p = .82$; high-SE students: $t(56) = 0.802$, $p = .82$; *p*-values Holm corrected to adjust for two-fold measurements).

(2) The low-SE student group (Fig. 4, right), with matching agent ($n = 27$, $M = 2.3$, $SD = 3.98$) showed a statistically significant improvement on a one-sample *t*-test ($t(26) = 3.05$, $p = .011$) while the low-SE student group with mismatched agent ($n = 28$, $M = 1.4$, $SD = 4.08$) showed only a marginally significant effect on a one-sample *t*-test ($t(27) = 1.85$, $p = .075$, *p*-values Holm corrected to adjust for two-fold measurements).

(3) The high-SE students showed no significant change either for the matching or mismatched agent (Fig. 4, right); both conditions having standard error bars crossing the 'zero' line. Considering that the pre-test SE scores for the high SE students (matched: *Median* = 31, *Range* = [29, 35]; mismatched: *Median* = 31, *Range* = [29, 35]) were already close to the maximal of 35, there was little room for any increase. This points toward a likely ceiling effect.

Overall, the effects of matching vs. mismatching were not significant; i.e. no unambiguous 'similarity effect' with regard to agent x student SE match-mismatch was found. However, considering the difference between the two conditions for the low SE students and the possibility of ceiling effects for high SE students, we cannot conclusively disregard an effect of 'similarity attraction'.

Self-Efficacy (Mis)Match and Students' In-Game Performance (Q2c)

To determine whether the match or mismatch of agent's and student's SE affected student in-game performance, we looked at in-game performance (section "In-game performance"). The dataset was again divided into matching ($M = 58.0$, $SD = 14.0$) and mismatching ($M = 51.9$, $SD = 14.7$) groups.

Figure 5 (left) indicates an overall match-mismatch effect between matched ($M = 58.0$, $SD = 14.0$) and mismatched groups ($M = 51.9$, $SD = 14.7$) agent SE x student SE groups.

Diving into the low- and high-SE student groups separately (Fig. 5, right) shows that this effect can be uniquely attributed to the low-SE student group. A two-sample t-test between the 'low x low' ($M = 57.5$, $SD = 10.4$) and 'high x low' ($M = 48.1$, $SD = 14.6$) agent SE x student SE groups displayed a medium to large statistically significant effect ($t(53) = 2.75$, $p = .0081$, Cohen's $d = 0.74$). That is to say that the matched subgroup performed markedly better than mismatched subgroup. No such effect was found for the high-se group (matched: $M = 58.5$, $SD = 17.0$, mismatched: $M = 55.5$, $SD = 14.0$; two sample t-test: $t(56) = 0.719$, $p = .48$).

Thus, there seems to exist a similarity effect in that students in the low-SE group teaching a digital tutee low in SE (matched SE) performed significantly better compared to students in the low-SE group teaching a digital tutee high in SE (mismatched SE).

An additional interesting observation is that students in the low-SE group, teaching a digital tutee low in SE (matched SE) performed at the same level as students in the high SE group (Fig. 6).

Discussion

We had two primary research aims with matching research questions. One aim was to explore if a digital tutee's expression of high versus low SE would have any effect on students with respect to their attitude towards the tutee, own SE, or in-game

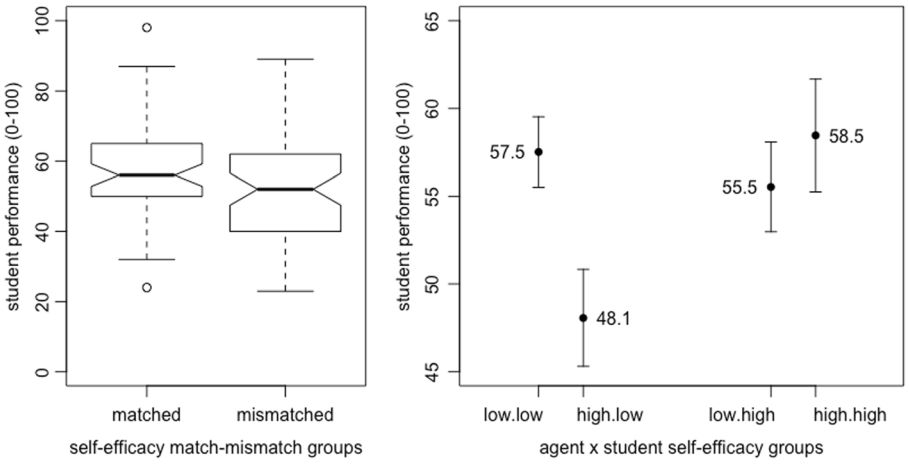

Fig. 5 Boxplot of in-game performance for matched vs. mismatched agent x student SE pairings (left); means and standard errors for all possible agent and student pairings (right)

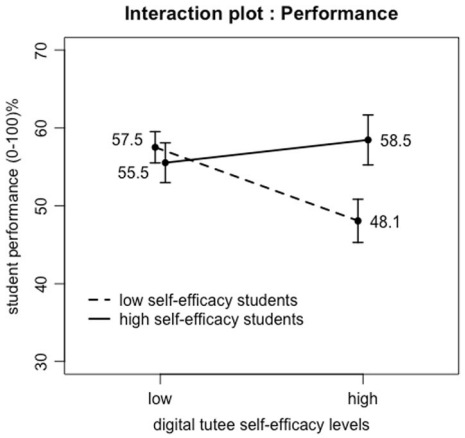

Fig. 6 Student in-game performance with regard to digital tutee vs. student SE

performance. The other was to explore whether deliberately matching or mismatching student and tutee SE would have any impact on these same outcomes.

Results in the Overall Student Population

Q1: How do students respond to a digital tutee with high versus *low SE?*

- *Q1a: Will the digital tutee's SE affect students' attitude towards their tutee?*
- *Q1b: Will the digital tutee's SE affect potential increase or decrease in students' own SE?*
- *Q1c: Will the digital tutee's SE affect students' in-game performance?*

The results clearly show that it did not matter whether students taught a digital tutee with high or low SE when it came to what they thought of their tutee (their attitude towards their tutee) or for their own SE. It did, however, have an effect on how well they performed. Teaching a digital tutee with low SE was more beneficial to in-game performance than teaching one with high SE. A possible explanation may be found in the aforementioned *protégé effect* (section "Characteristics of digital tutees in relation to student learning") together with a general tendency to interact differently with different agent personalities. The protégé effect refers to the finding that students put more effort into teaching someone else compared to when they learn for themselves (Chase et al. 2009). Yet it is possible that students in a teacher role take *more* responsibility for a digital tutee with low SE precisely because this tutee expresses a low trust in her own ability to learn, and possibly comes across as someone who is in need of more help than a digital tutee with high SE.

Self-efficacy (Mis)match Effects

Q2: (How) do student respond to match/mismatch regarding their own SE and that of their digital tutee?

- *Q2a: Does match/mismatch between SE in student and digital tutee have effects on students' attitude to the digital tutee?*

We predicted that we would find a similarity attraction effect so that students teaching a digital tutee with matching SE would show a more positive attitude towards their tutee. Contrary to our expectation we found no such effects. The results of previous studies into (mis)matching effects in human-agent interaction are mixed (section "Previous research on matching vs. not matching characteristics between actors in an educational context"). Some (Hietala and Niemirepo 1998; Nass and Lee 2000; Kim 2007) report a similarity attraction effect, i.e. more positive attitude towards an agent when matched; others (Isbister and Nass 2000; Behrend and Thompson 2011) do not.

We speculate that the divergence in results have to do which characteristic of (dis)similarity is addressed and the way attitude is measured. Hietala and Niemirepo's (1998) measure was how much time students chose to spend with different available agents, while Isbister and Nass (2000) asked participants to indicate how well certain words (e.g. 'assertive', 'friendly', and 'bashful') corresponded to the agent they had interacted with.

It is possible that our way of measuring attitude was inappropriate. It is also possible that the characteristic of SE plays no significant role in similarity attraction. Yet another possibility behind our none-result is the particular role of a digital tutee. In most other studies the pedagogical agent is a peer or a collaborator (or both). A peer is somehow equal to the students; it can think and, in some ways, act on its own, whereas a digital tutee has a more submissive role having to learn from the student (who is its teacher).

- *Q2b: Does match/mismatch between SE in student and digital tutee have effects on students' potential increase in SE?*

One study (Pareto et al. 2012) has shown that software including digital tutees as such can influence students' SE, compared to a control-group using the same software without digital tutees. But this provides no bases for predicting whether a (mis)match between student and tutee SE will influence students' SE. As it turned out, we found no effect when all the matching and all the non-matching students were considered together. Diving down into the high-vs low-SE student subgroups, however, revealed certain interesting patterns.

Low-SE students increased their SE both in the matched and the mismatched condition, though the increase was only statistically significant in the matched condition, i.e. the digital tutee also had low SE. Thus, one can speculate that a tutee with low SE may indeed have a larger potential to boost SE in students who themselves have initially low SE. A tutee with low SE expresses a feeling of not knowing, which may boost the student's confidence for knowing more than the tutee.

High-SE students, from the start have a great deal of confidence in their ability to deal with the math tasks in the game. The room for any increase in SE is small, producing a ceiling-effect in our data. In addition, many students with high SE have stable SE judgments over time and are not easily influenced by momentary or single experiences of non-success (Bandura 1997). Nevertheless, the results reveal a small (less than statistically significant) *decrease* in SE for the high-SE group in the mismatched condition (with a low-SE tutee), and a small (less than statistically significant) *increase* in SE in the matched condition (with a high-SE tutee). In other words, it cannot be excluded that there is a similarity attraction effect that is hidden behind a ceiling effect.

To determine whether that is true would require circumventing the ceiling effect – at the least, a significant methodological challenge. However, there is a more practical pedagogical question: Is it pedagogically meaningful to boost the SE in a domain for someone who already has a high (stable) SE in the domain?

- *Q2c: Does match/mismatch between SE in student and digital tutee have effects on students' in-game performance?*

Once more, the, lack of previous studies on (mis)matching students and digital tutees with respect to SE meant we could make no predictions whether our manipulation would affect students' in-game performance. No effect was found for students with high SE. An effect *was* found for students with low SE. Low-SE students performed significantly better when teaching a digital tutee with low SE; a clear similarity-attraction effect.

As discussed above, low-SE students are generally more likely than high-SE students to benefit from teaching someone else. It did not seem to matter to the high-SE students in our study whether they instructed a tutee with high or low SE. Overall, the low-SE students in our study seemed to benefit more than high-SE students from playing the game and instructing the tutee.

In addition, we made the following observation for low-SE students teaching a low-SE tutee: Their in-game performance was, in this case, comparable to the in-game performance of the high-SE student group (a group that in general performs at a higher level). When low-SE students taught a digital tutee with high SE their in-game performance was considerably lower and did not reach the level of the high-SE student group.

When the digital tutee expresses a sense of not being able to manage the task – which is what the low-SE tutee routinely does – the student could experience this as negative, or critical, feedback on her teaching. As mentioned before students with high SE are less susceptible than students with low SE to single instances of failure or other forms of 'negative' feedback. In contrast to low-SE students, they tend to forget quickly about it (Bandura 1997). However, in our study *also students with low SE* were *positively* influenced by the feedback from a low SE tutee, even though it was 'negative' and 'critical' in the sense explicated above. What probably matters is that the feedback is *recursive* in the sense used by Okita and Schwartz (2013): it does not *directly* target the student herself – even though most students understand that the performance of the digital tutee reflects how well they themselves instruct it. The recursiveness of the feedback functions as an ego-protective buffer and gives the student a teaching comfort zone.

One final note: it is often assumed that performing better is closely related to liking something more. Our results might appear to argue against this. We found a statistically significant effect of (mis)matching SE on in-game performance but *not* on attitude.

Limitations

One important limitation in the study is that the digital tutee's SE level was held constant – either low or high – through all sessions. As discussed above, the risk

is that high-performing students (a group that overlaps with high-SE students) who are assigned a low-SE tutee, that they teach well, may with time get frustrated.

Though the tutee makes progress and performs well – when it is taught well – it continues despite all successes to express low belief in its ability to succeed. That is, nothing changes in the tutee's SE even though it repeatedly gets to 'experience' success.

Another limitation lies with the high-SE students' ceiling effect – at least for concluding whether SE (mis)match plays any role for high-SE students' SE as it does for low-SE students. Another kind of instrument or measurement would be required to explore this further.

Yet another limitation has to do with the digital tutee's limited conversational abilities. Over the course of all the sessions, students chatted with the tutee extensively, and it became clear that they were getting frustrated with the tutee's inability to answer many of their questions. For future studies, either the agent's conversational abilities should be extended or the opportunities to chat with it curtailed.

Finally, there is the problem regarding generalizability of results. A teachable agent (digital tutee) is not like other kinds of digital pedagogical agents being more intertwined with the student it interacts with. A digital tutee depends on the student for its learning. This is not the case with pedagogical agents that function as instructors or coaches Correspondingly, a digital tutee is more compliant than other kinds of pedagogical agents. It may collaborate to some extent with its student teacher, but the relation is less *even* than in the case of digital peers or other learning companions. With this said, there is a grey scale between a learning companion agent and a digital tutee. Therefore, what is found to apply for digital tutees is sometimes relevant for companion agents too.

All participants in this study were of the same age group and socioeconomic background. In order to reach more general and conclusive results studies with other populations are needed.

Conclusion and Future Work

It remains far from clear how best to design an agent for a digital learning environment so as to support learning in a wide range of students. Some design choices appear to have more consequences than others; some work out in unexpected ways, affecting different groups of students differently. Yet, for each digital learning environment including pedagogical agents, there *are* a large number of design choices to be taken by developers and designers.

In this study and paper, we have looked at one particular design choice with respect to digital tutees: namely, what level of SE a digital tutee should express regarding the domain of instruction (and, with that, what the effect is of matching or mismatching its SE to that of the student). Through our study, we have attempted to collect knowledge on which this design choice can be based.

Our research questions were: (How) will it affect the students instructing the tutee if the tutee expresses low or high SE, respectively? (How) will it affect the

students' in-game performance, attitude towards the tutee and own subsequent SE? We approached the questions both with respect to an entire student population and while examining match versus non-match in high-low SE between digital tutee and student teaching it.

What we found was that the tutee's SE had no effect on students' attitude toward the agent; neither did it have any statistically significant effect on students' own SE. It did, however, have a significant impact on in-game performance – at least with respect to the sub-group of low-SE students. One might conclude that, in general, students gained more performance-wise from instructing a digital tutee with low rather than high SE – but the effect was by far the most pronounced for students whose own SE was low: i.e., whose SE matched that of the agent.

Separating the low- from the high-SE students, we found some tantalizing effects of tutee SE on student SE. Low-SE students increased their SE considerably regardless of condition – but with a trend towards a stronger effect when they taught a low-SE agent. The high-SE students – perhaps not surprisingly – did not change their SE much and may well have encountered a ceiling effect. The small differences we found between conditions would, however, be interesting to try to study further, despite our lack of statistically significant results: in particular the way that, when high-SE students instructed a low-SE agent, their own SE seemed to *decrease* slightly, and when they instructed a high-SE agent, it increased slightly. Thus, it cannot be excluded on the basis of our study that there is a matching-effect with respect to high-SE students, but in our case hidden behind a ceiling-effect.

As a tentative conclusion, we propose that a designer facing the choice between a digital tutee expressing high or low SE, should opt for one with low SE. We base this recommendation on two principal findings: (i) for the entire student sample, in-game performance was stronger in the groups instructing a low-SE tutee; (ii) students with low SE benefited greatly from interacting with the low-SE agent, while students with high SE suffered, at most, a very minimal decrease in SE and no decrease in in-game performance.

As an additional recommendation, we propose that future studies use a design where the teachable agent's SE can develop over time.

On a more general level our study contributes – together with similar studies on how agent characteristics affect students learning outcomes – by pointing to design choices that designers of pedagogical agents need to deal with.

Design choices have effects on learning; on the one hand for an entire, broad, student population, and, on the other hand, perhaps for different groups of students in different ways. With a likely increasing role for agent-based educational software in the future, the burden lies on the academic community to conduct the necessary research for making informed choices starting today.

Educational software has a tremendous and still largely untapped potential to cater for a wide range of students. One single software can offer a pedagogical agent with several levels of expertise, several communicative styles, gender expressions, SE levels, and so on. Before this can be realized in practice, however, more research is needed. We see the study reported in this paper as just among the many needed for charting out the still-unmapped territory.

Appendix 1

Name: **Class:**

Here are some questions regarding Lo, please mark what you think/how you feel.

1. **How well do you think Lo has learned?**

Not at all well	Not well	Neither nor	Well	Really well
☐	☐	☐	☐	☐

2. **How well do you think Lo has played the game when alone?**

Not at all well	Not well	Neither nor	Well	Really well
☐	☐	☐	☐	☐

3. **Who do you think Lo´s learning depends on?**

   a. On me

   b. On Lo

   c. On both me and Lo

   d. Neither on me nor on Lo

4. **How has it been to instruct Lo?**

Really boring	Boring	Neither nor	Fun	Really fun
☐	☐	☐	☐	☐

5. **What do you think about your own ability to train Lo?**

Really bad	Bad	Neither nor	Good	Really good
☐	☐	☐	☐	☐

6. **What do you think about Lo´s confidence in math?**

Really uncertain	Uncertain	Neither nor	Confident	Really confident
☐	☐	☐	☐	☐

7. How has it been to chat with Lo?

Really boring ☐ Boring ☐ Neither nor ☐ Fun ☐ Really fun ☐

8. Would you like to continue to instruct Lo?

Yes ☐ No ☐ Maybe ☐

Why? _____

_____

9. Circle the words you think describes Lo

Not smart Weird Nice

Mean Uncertain Certain

Kind Annoying Cocky

Boring High self-confidence Normal

Sissy Smart Funny

Low self-confidence

Appendix 2

Nr: _____

I would like you to try to estimate how *well you would do* if you were asked to solve a number of tasks. You do NOT have to solve the tasks. Just mark how good you *would be* at solving them.

How good would you be at solving these tasks?

	Really bad	Bad	Neither nor	Good	Really good
1136 + 346	☐	☐	☐	☐	☐
184 − 64	☐	☐	☐	☐	☐
What number is missing? 670 − ____ = 485	☐	☐	☐	☐	☐
You have the number 274. Will the result end with 00 if you add 3826?	☐	☐	☐	☐	☐
Which of the totals is largest? 295 + 16 + 1719 or 32 + 2234 + 123	☐	☐	☐	☐	☐
What number do you get if you swap the hundred and the ten in 437?	☐	☐	☐	☐	☐
What number has the largest value in 6275?	☐	☐	☐	☐	☐

References

Annis, L. F. (1983). The processes and effects of peer tutoring. *Journal of Educational Psychology, 2*(1), 39–47.

Arroyo, I., Woolf, B. P., Royer, J. M., & Tai, M. (2009). Affective gendered learning companions. In *Proc. of* the *International Conference on Artificial Intelligence and Education,* (pp. 41–48). IOS Press.

Arroyo, I., Woolf, B. P., Cooper, D. G., Burleson, W., & Muldner, K. (2011). The impact of animated pedagogical agents on girls' and boys' emotions, attitudes, behaviors and learning. In I. Aedo, N.-S. Chen, D. G. Sampson, J. M. Spector, Kinshuk (Eds.), Proceedings of the 11th conference on advanced learning technologies, ICALT 2011 (pp. 506–510). Piscataway, NJ: IEEE.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*(2), 191–215.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W. H. Freeman.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Multifaceted impact of self-efficacy beliefs on academic functioning. *Child Development, 67*(3), 1206–1222.

Baylor, A. L., & Kim, Y. (2004). Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. In J.C. Lester, R.M. Vicari, & F. Paraguaçu (Eds.), Lecture Notes in Computer Science, vol 3220: Proceedings of Intelligent Tutoring Systems 2004 (pp. 592–603). Berlin/Heidelberg; Germany: Springer.

Baylor, A. L., & Kim, Y. (2005). Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education, 15*(2), 95–115.

Behrend, T. S., & Thompson, L. F. (2011). Similarity effects in online training: Effects with computerized trainer agents. *Computers in Human Behavior, 27*(3), 1201–1206.

Biswas, G., Leelawong, K., Schwartz, D., Vye, N., & TAG-V. (2005). Learning by teaching: A new agent paradigm for educational software. Applied Artificial Intelligence, 19(3–4), 363–392.

Byrne, D., & Nelson, D. (1965). Attraction as a linear function of proportion of positive reinforcements. *Journal of Personality and Social Psychology, 1*(6), 659–663.

Byrne, D., Griffitt, W., & Stefaniak, D. (1967). Attraction and similarity of personality characteristics. *Journal of Personality and Social Psychology, 5*(1), 82–90.

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

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

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development*. Philadelphia: Psychology Press.

Ebbers, S. J. (2007). *The impact of social model agent type (coping, mastery) and social interaction type (vicarious, direct) on learner motivation, attitudes, social comparisons, affect and learning performance*. Doctoral dissertation, Florida State University, Tallahassee, FL. http://etd.lib.fsu.edu/theses/available/etd-07092007-151016/.

Fiorella, L., & Mayer, R. E. (2013). The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology, 38*(4), 281–288.

Hietala, P., & Niemirepo, T. (1998). The competence of learning companion agents. *International Journal of Artificial Intelligence in Education, 9*, 178–192.

Isbister, K., & Nass, C. (2000). Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies, 53*(2), 251–267.

Johnson, A. M., Ozogul, G., & Reisslein, M. (2015). Supporting multimedia learning with visual signalling and animated pedagogical agent: Moderating effects of prior knowledge. *Journal of Computer Assisted Learning, 31*(2), 97–115.

Kim, Y. (2007). Desirable characteristics of learning companions. *International Journal of Artificial Intelligence in Education, 17*(4), 371–388.

Kim, Y., & Baylor, A. L. (2006). A social-cognitive framework for pedagogical agents as learning companions. *Educational Technology Research and Development*, *54*(6), 569–596.

Kim, Y., & Wei, Q. (2011). The impact of learner attributes and learner choice in an agent-based environment. *Computers & Education, 56*(2), 505–514.

Kim, Y., Hamilton, E. R., Zheng, J., & Baylor, A. L. (2006). Scaffolding learner motivation through a virtual peer. In *Proc. of the 7th International Conference on Learning Sciences, ICLS'06* (pp. 335–341). Bloomington, IN: International Society of the Learning Sciences.

Kirkegaard, C. (2016). Adding challenge to a teachable agent in a virtual learning environment. In *Doctoral dissertation, Linköping University*. Linköping, Sweden: Linköping University Electronic Press.

Kirkegaard, C., Tärning, B., Haake, M., Gulz, A., & Silvervarg, A. (2014). Ascribed gender and characteristics of a visually androgynous teachable agent. In T. Bickmore, S. Marsella, & C. Sidner (Eds.), *Lecture notes in computer science, vol 8637: Proceedings of intelligent virtual agents 2014* (pp. 232–235). Cham, Switzerland: Springer.

Lee, E. J., & Nass, C. (1998). Does the ethnicity of a computer agent matter? An experimental comparison of human-computer interaction and computer-mediated communication. In *Proceedings of the 1st workshop of embodied conversational characters, WECC'98* (pp. 123–128). Tahoe City, CA: ACM Press.

Lee, J. E., Nass, C., Brave, S. B., Morishima, Y., Nakajima, H., & Yamada, R. (2007). The case for caring colearners: The effects of a computer-mediated colearner agent on trust and learning. *Journal of Communication, 57*(2), 183–204.

Mayer, R. E., & DaPra, C. S. (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology: Applied, 18*(3), 239–252.

Nass, C., & Lee, K. M. (2000). Does computer-generated speech manifest personality? An experimental test of similarity-attraction. In *Proc. of the SIGCHI conference on Human Factors in Computing Systems* (pp. 329–336). The Hague/Amsterdam, The Netherlands: ACM Press.

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied, 7*(3), 171–181.

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, C. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies, 43*(2), 223–239.

Newcomb, T. M. (1956). The prediction of interpersonal attraction. *American Psychologist, 11*(11), 575–586.

Okita, S. Y., & Schwartz, D. L. (2013). Learning by teaching human pupils and teachable agents: The importance of recursive feedback. *Journal of the Learning Sciences*, *22*(3), 375–412.

Panton, M. K., Paul, B. C., & Wiggers, N. R. (2014). Self-efficacy to do or self-efficacy to learn to do: A study related to perseverance. *International Journal of Self-Directed Learning, 11*(1), 29–40.

Papert, S. (1993). *The children's machine: Rethinking school in the age of the computer*. New York, NY: Basic books.

Pareto, L. (2014). A teachable agent game engaging primary school children to learn arithmetic concepts and reasoning. *International Journal of Artificial Intelligence in Education, 24*(3), 251–283.

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.), *Lecture notes in computer science, vol 6738: Proceedings of artificial intelligence in education 2011* (pp. 247–255). Berlin/Heidelberg, Germany: Springer-Verlag.

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

Plant, E. A., Baylor, A. L., Doerr, C. E., & Rosenberg-Kima, R. B. (2009). Changing middle-school students' attitudes and performance regarding engineering with computer-based social models. *Computers & Education, 53*(2), 209–215.

Pratt, J. A., Hauser, K., Ugray, Z., & Patterson, O. (2007). Looking at human–computer interface design: Effects of ethnicity in computer agents. *Interacting with Computers, 19*(4), 512–523.

R Core Team. (2016). *R: A language and environment for statistical computing [computer software]*. Vienna, Austria: R Foundation for Statistical Computing.

Rattan, A., Good, C., & Dweck, C. S. (2012). "It's ok – Not everyone can be good at math": Instructors with an entity theory comfort (and demotivate) students. *Journal of Experimental Social Psychology, 48*(3), 731–737.

Reeves, B., & Nass, C. (1996). *How people treat computers, television, and new media like real people and places*. Stanford, CA: CSLI Publications.

Roscoe, D., Wagster, J., & Biswas, G. (2008). Using teachable agent feedback to support effective learning by teaching. In *Proc. of Cognitive Science Conference* (pp. 2381–2386). Washington, DC: Cognitive Science Society.

Rosenberg-Kima, R. B., Baylor, A. L., Plant, E. A., & Doerr, C. E. (2008). Interface agents as social models for female students: The effects of agent visual presence and appearance on female students' attitudes and beliefs. *Computers in Human Behavior, 24*(6), 2741–2756.

Rosenberg-Kima, R. B., Plant, E. A., Doerr, C. E., & Baylor, A. L. (2010). The influence of computer-based model's race and gender on female students' attitudes and beliefs towards engineering. *Journal of Engineering Education, 99*(1), 35–44.

Schunk, D. H. (1987). Peer models and children's behavioral change. *Review of Educational Research, 57*(2), 149–174.

Sherman, H. J., Richardson, L. I., & Yard, G. J. (2015). *Teaching learners who struggle with mathematics: Responding with systematic intervention and remediation*. Long Grove: Waveland Press.

Silvervarg, A., & Jönsson, A. (2011). Subjective and objective evaluation of conversational agents. In *Proc. of the 7th Workshop on Knowledge and Reasoning in Practical Dialogue Systems* (pp. 65-72). Barcelona, Spain.

Silvervarg, A., Raukola, K., Haake, M., & Gulz, A. (2012). The effect of visual gender on abuse in conversation with ECAs. In Y. Nakano, M. Neff, A. Paiva, & M. Walker (Eds.), *Lecture notes in computer science, vol 7502: Proceedings of intelligent virtual agents 2012* (pp. 153–160). Berlin/Heidelberg, Germany: Springer.

Silvervarg, A., Haake, M., & Gulz, A. (2013). Educational potentials in visually androgynous pedagogical agents. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Lecture notes in computer science, vol 7926: Artificial intelligence in education 2013* (pp. 599–602). Berlin/Heidelberg, Germany: Springer-Verlag.

Sjödén, B., Tärning, B., Pareto, L., & Gulz, A. (2011). Transferring teaching to testing – An unexplored aspect of teachable agents. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Lecture notes in computer science, vol 6738: Proceedings of artificial intelligence in education 2011* (pp. 337–344). Berlin/Heidelberg, Germany: Springer-Verlag.

Tärning, B., Haake, M., & Gulz, A. (2017). Supporting low-performing students by manipulating self-efficacy in digital tutees. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society* (pp. 1169–1174). Austin, TX: Cognitive Science Society.

Uresti, J. A. R. (2000). Should I teach my computer peer? Some issues in teaching a learning companion. In G. Gauthier, C. Frasson, & K. VanLehn (Eds.), *Lecture notes in computer science, vol 1839: Proceedings of intelligent tutoring systems 2000* (pp. 103–112). Berlin/Heidelberg, Germany: Springer.

Uresti, J. A. R., & du Boulay, B. (2004). Expertise, motivation and teaching in learning companion systems. *International Journal of Artificial Intelligence in Education, 14*(2), 193–231.

Veletsianos, G. (2009). The impact and implications of virtual character expressiveness on learning and agent–learner interactions. *Journal of Computer Assisted Learning, 25*(4), 345–357.

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies, 66*(2), 98–112.