Design of teachable agents and feedback in educational software

Focusing on low-performing students and students with low self-efficacy

BETTY TÄRNING
COGNITIVE SCIENCE | LUND UNIVERSITY

Printed by Media-Tryck, Lund 2018 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

Faculties of Humanities and Theology
Department of Philosophy
Cognitive Science

Lund University Cognitive Studies 171
ISBN 978-91-88473-67-7
ISSN 1101-8453

Design of teachable agents and feedback
in educational software

Design of teachable agents and feedback in educational software

Focusing on low-performing students and students with low self-efficacy

Betty Tärning

LUND
UNIVERSITY

DOCTORAL DISSERTATION
by due permission of
the Faculty of Humanities and Theology, Lund University, Sweden.
To be defended at LUX, room C126, 9 March 2018, at 10:00.

Faculty opponent
Kristen Pilner Blair
AAALab, Stanford University

Organization	Document name
LUND UNIVERSITY Cognitive Science, Department of Philosophy	Doctoral dissertation
	Date of issue: 16 February 2018
Author: Betty Tärning	Sponsoring organization

Title and subtitle: Design of teachable agents and feedback in educational software: Focusing on low-performing students and students with low self-efficacy.

Abstract

The use of educational software is rapidly expanding. The six papers comprising this thesis address the design of teachable agents (digital tutees) and feedback in educational software, with a focus on low-performing students and students with low self-efficacy.

Paper I examines differences between high- and low-performing students interacting with a teachable agent capable of off-task conversation. The two groups are shown to differ in their engagement in off-task conversation. Papers II-IV examine a characteristic of teachable agents that, to my knowledge, has not been studied before: self-efficacy. The teachable agent's self-efficacy (high vs. low) was revealed via a conversational chat. The agent provided the student recursive feedback by expressing her thoughts about each just-completed game session. The results show how a teachable agent with seeming low self-efficacy can positively impact low-performing and low self-efficacy students with respect to performance and self-efficacy.

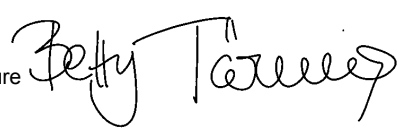
Papers V-VI focus on feedback in the broader and more common sense: namely, feedback from someone/something regarding one's performance. Paper V contributes to a more detailed understanding of where in the feedback chain students fall off: from (i) noticing feedback (ii) processing, for example read, it, (iii) making sense of it, (iv) acting on it, and finally to (v) progressing on the basis of the feedback. The results suggest that agents can be used to help students' pay attention – at least to textual feedback. Paper VI surveys the types of feedback provided by digital apps currently in use in Swedish schools. It addresses the mismatch between the feedback provided by the majority of apps and what researchers and educators alike understand to be appropriate. The majority of apps being sold as educational software are nothing more than glorified digital tests.

This thesis contributes to the educational software domain by pinpointing two characteristics of teachable agents that can be exploited to improve performance and self-efficacy for low-performing and low self-efficacy students. It fills in missing details on the processes by which students make use of – or fail to make use of – software-delivered feedback. Finally it shows how designers of educational software need to do a better job designing feedback.

Keywords. Educational technology, Pedagogical agents, Feedback, Self-efficacy, Low-performing students, Students with low self-efficacy.

Classification system and/or index terms (if any)

Supplementary bibliographical information		Language: English
ISSN 1101-8453 Lund University Cognitive Studies 171		ISBN 978-91-88473-67-7
Recipient's notes	Number of pages 270	Price
	Security classification	

Signature  Date 29 January 2018

Design of teachable agents and feedback in educational software

Focusing on low-performing students and students with low self-efficacy

Betty Tärning

LUND
UNIVERSITY

Omslagsfoto: Magnus Haake

Cognitive Science
Department of Philosophy
Faculties of Humanities and Theology

ISBN 978-91-88473-67-7
ISSN 1101-8453

Printed in Sweden by Media-Tryck, Lund University, Lund 2018

MADE IN SWEDEN

To mum and dad

Table of contents

Acknowledgement

It is with great pleasure I finally get to write the acknowledgement, not only because this means the thesis is actually finished but also because there are so many people whom I would like to thank.

First of all I would like to thank the Educational Technology Group (Agneta Gulz, Magnus Haake, Annika Silvervarg, Björn Sjödén, Jens Nirme, Camilla Kirkegaard, Erik Anderberg, and Kristian Månsson). Not only have learned so much from you, but you have also given me some of my most memorable moments at work. I will never forget all laughter when we recorded the voices to Magical Garden (remember Erik?) or how much fun one could have in the in the 16th century (or what do you say Kristian?).

I was lucky to have Agneta and Magnus as colleagues, but I was also fortunate enough to have them as supervisors. Working with you during these years has been great fun. Not only are you two of the most knowledgeable people I know but also two of the funniest. Agneta, even though you are always busy with a thousand different projects, you always have time for my many questions and fixing my spelling mistakes. I greatly admire your superhero powers like reading at the speed of light) – but, just as much, I admire your inability to find the way as well as your (in)ability to pack a bag. Magnus, with whom I had the pleasure of traveling on several occasions. Travelling with you, is an adventure in itself. Not only do I want to thank you for delayed flights, trains, and shopping time – but also for all help with statistics and layout work. On a more serious note, I cannot thank you enough for all help, I would not have made it without you two.

Annika Silvervarg and Lena Pareto, my second supervisors, thank you a lot. Lena, first of all thanks for developing "The Squares Family". Most papers would not exist without your work. Annika, thanks for always being positive and thinking that we will make it. You are a great writing partner.

And Björn, my ex colleague, my friend, and my B2. You know how smart I think you are, so no need to repeat that. I have greatly enjoyed all our experimental sessions, trips, and conversations. I both appreciate your company and your role as the devil's advocate. And I hope that the croissants will keep on coming.

Writing this thesis would not have been half as fun if it weren't for all colleagues. Thanks to all my PhD-colleagues both present and past; Andrey Anikin, Paulina Lindström, Rasmus Arnling Bååth, Katarzyna Bobrowicz, Zahra Gharaee, Ivo Jacobs, Can Kabadayi, Andreas Lind, Jens Nirme, Manuel Oliva, Åsa Harvard, Kristin Osk Ingvarsdottir, Thomas Strandberg, Richard Andersson, Joel Parthemore, Philip Pärnamets, and Trond Arild Tjøstheim. Thomas, thanks for

being my first roomie and for fun times both doing experiments and watching handball. Zahra and Manuel, thanks for keeping me company during the Christmas holidays. Andrey, you have saved me more than once; thanks for always helping. And Joel, thanks for all proofreading.

A special thanks to my current roomies Jens and Trond. Jens for always working so hard (pressing me to do the same) and for making a virtual me. Trond, a big thank you for your patience with all my questions, for always encouraging me, taking me out on café tours, and for inventing the submission dance.

To my other colleagues at LUCS, Peter Gärdenfors, Christian Balkenius, Annika Wallin, Petter Johansson, Jana Holsanova, Mathias Osvath, Tomas Persson, Kerstin Gidlöf, Magnus Johnson, Lars Hall, Megan Lamberts, Helena Osvath, Birger Johansson, Eva Sjöstrand, and Ingela Byström. Thanks for all engagement and input at seminars and for all great lunch conversations. Lars and Petter, this all started with you (I am not yet sure if I should thank you for that or not ;-) – but I have greatly appreciated all our conversations. And Lars, I hope you have stretched enough, because "the time has come…" Birger, not only are you my colleague, but you have also become a dear friend. Thanks for being my private IT-guy shrink, dance and bingo partner – there is never a dull time with you.

And thanks to "the philosophers", Anna Cagnan Enhörning, and Astrid Byrman for being great lunch and "fika" partners. Frits, I greatly enjoy our conversations but I still don't think I get anything. Anna and Astrid, thanks for being so much fun but also for always helping with all those small things.

None of this would have been possible without the help from all schools, teachers, and students participating in our experiments. A huge thanks you to all of you. In particular – a big thanks to Håkan Andersson who more than once have welcomed us to his classes, always with a smile and an interest in what we do.

I also want to thank my girlfriends, Lizette, Erika, Susanne, Louise, Josefin, Joanna, Hanna, and Emma, for never ever talking about research or work. And cannot thank you enough for all the caring and fun moments we shared; also, thanks to your spouses, they are quite cool too!

I would also like to thank my family. Mum and dad, this is for you for always believing in me no matter what. I love you! David, thanks for always asking me how it goes and for thinking that what I do is actually cool. Thanks also to my extended family, Stina, Helene, Janne, Emma, and Linnea for all fun moments outside academia.

Lastly I would like to thank my "new" family. Abbe and Elli, thank you for always reminding me of what is important in life and for being you. And finally, the biggest thank you goes to Johan, for always being there, for all encouragement, for wiping my tears away and for making me feel loved. Love you!

List of original papers

Paper I

Tärning, B., M. Haake, & A. Gulz (2011). Off-task engagement in a teachable agent based math game. In *Proceedings of the 19th International Conference on Computers in Education* (60-64). Chiang Mai, Thailand.

Paper II

Tärning, B., A. Silvervarg, A. Gulz, & M. Haake (submitted). *Instructing a teachable agent with low or high self-efficacy: Does similarity attract?*

Paper III

Tärning, B. & A. Silvervarg (submitted). *"I didn't understand, I'm really not very smart": How design of a conversational teachable agent with low self-efficacy can contribute to student performance and self-efficacy*.

Paper IV

Tärning, B., A. Gulz, & M. Haake (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). London, UK: Cognitive Science Society.

Paper V

Tärning, B., Y.J. Lee, R. Andersson, K. Månsson, A. Gulz, & M. Haake (submitted). *Looking into the black box of students' (not) handling feedback on mistakes*.

Paper VI

Tärning, B. (submitted). *Review of feedback in digital applications – does the feedback they provide support learning?*

Other, related work by the author

Tärning, B., B. Sjödén, A. Gulz, & M. Haake (submitted). *Young children's experience and preference of feedback: Sense and sensibility*.

Kirkegaard, C., B. Tärning, M. Haake, A. Gulz, & A. Silvervarg (2014). Ascribed gender and characteristics of a visually androgynous teachable agent. In *Proc. of the Int. Conf. on Intelligent Virtual Agents* (pp. 232-235). Heidelberg, Germany: Springer.

Sjödén, B. & B. Tärning (2014). Appsolute pedagogy: Reviewing digital functionality of Swedish iPad apps for early math and literacy skills. Poster presented at the *Swedish Cognitive Science Society National Conference (SweCog 2014)*, September 2014, Skövde, Sweden.

Sjödén, B., B. Tärning, L. Pareto, & A. Gulz (2011). Transferring teaching to testing: An unexplored aspect of teachable agents. In *Proc. of the Int. Conf. on Artificial Intelligence in Education* (pp. 337-344). Heidelberg, Germany: Springer-Verlag.

Silvervarg, A., M. Haake, L. Pareto, B. Tärning, & A. Gulz (2011). Pedagogical agents: Pedagogical interventions via integration of task-oriented and socially oriented conversation. Oral presentation at the *AERA 2011 Symposium Pedagogical Agent Presence, Appearance, and Agent-learner Interactions: Current Research and Future Directions*, New Orleans, LA.

Tärning, B. (2011). Effects of feedback detail and frequency in a teachable agent based learning game. Poster session presented at the *Swedish Cognitive Science Society National Conference 2011 (SweCog 2011)*, May 2011, Skövde, Sweden.

Haake, M., A. Silvervarg, B. Tärning, & A. Gulz (2011). Teaching her, him… or hir? Challenges for a cross-cultural study. In *Proc. of the 11th Int. Conf. on Intelligent Virtual Agents*, *LNCS 6895* (pp. 447-448). Heidelberg, Germany: Springer-Verlag.

Silvervarg, A., A. Gulz, B. Sjödén, M. Haake, & B. Tärning (2010). Designing a teachable agent with social intelligence. Poster presented at the *10th Int. Conf. on Intelligent Virtual Agents*, Philadelphia, PA.

Sjödén, B. & B. Tärning (2010). Transferring teaching to testing: How can a teachable agent aid performance when present on a math test. Poster presented at the *Swedish Cognitive Science Society National Conference (SweCog 2010)*, Örenäs, Sweden.

Gulz, A., M. Haake, & B. Tärning (2007). Visual gender and its motivational and cognitive effects: A user study. *Lund University Cognitive Studies, 137*. www.lucs.lu.se/wp-content/uploads/2013/09/gulz_haake_tarning_report_2007.pdf

Gulz, A., M. Haake, & B. Tärning (2007). Digitala lärmiljöer och androgynitetens potential. In *Proc. of Utvecklingskonferens LU: Att tänka om – ett kvalificerat akademiskt lärarskap*. Lund, Sweden: Lund University.

Gulz, A., M. Haake, & B. Tärning (2007). Challenging gender stereotypes using virtual pedagogical characters. Oral presentation at the *GLIT Symposium on Gender, Learning and IT*, Helsingborg, Sweden.

Introduction and scope of the thesis

When I was seven or eight years old I received The Little Professor as a Christmas gift from my parents, on recommendation from my teacher (Figure 1). It was a small device in the form of a calculator; my task was to answer random equations. My teacher thought I needed to practice my math skills, and I did since I usually do as I am told. Anyway, I did not have much faith in my ability to succeed in math.

In the beginning it was kind of fun when the professor would wiggle his moustache each time I provided the correct answer; but when I entered an incorrect answer, his lack of response did not help me much. At the time, I did not reflect that the device did not provide me with satisfactory feedback. I would just try another number and, with a little luck, the professor would soon wiggle his moustache again. Today though I have come to realize that this Christmas gift largely encapsulates what this thesis is about: educational technology, pedagogical agents, low-performing students, students with low self-efficacy, and feedback.

Figure 1. The little professor.

Since I received the professor some thirty years ago, the number of digital teaching devices has increased dramatically. Many are much better than my early encounter, but some are not that different in pedagogical quality. In any case, there is a long way to go. Research into educational software is still in its infancy. With this thesis, I hope to contribute a small piece to the larger puzzle.

Some of the devices used in school and everyday life make use of *digital pedagogical agents*: computer-controlled characters able to take on different pedagogical roles. They can be Hispanic or Asian, female or male, competent or less competent, provide more feedback or less, and so on. A large part of the

potential of pedagogical agent is that they can be designed with respect to so many characteristics. At the same time, this poses a major challenge. We know from previous research that each characteristic the designer decides upon can affect student learning. Not only do characteristics have different effects, the effects vary depending on student group. What characteristics have what effects on what students is, in many cases, unknown.

The types of pedagogical agents I focus on are *digital tutees*, often referred to as *teachable agents*: a subgroup of pedagogical agents designed to take the role of tutee. They are built around the concept of *learning by teaching*: while students instruct their tutee, they simultaneously learn for themselves. Working with a digital tutee can have many benefits. They have been shown to improve learning, ability to self-assess, and belief in one's own ability. That said, some groups benefit more than others. The two groups I chose to study more closely are low-performing students and students with low self-efficacy: two groups that stand to gain from well-designed instruction.

The characteristics of digital tutees I focused on are (i) ability to carry out both a task-oriented and an *off-task conversation* (the tutee should be able to talk not only about the learning task but also other topics such as family background, sports, and movies) and (ii) self-efficacy: the tutee's apparent belief in its ability to succeed with the task at hand.

Paper I examines differences between high- and low-performing students interacting with a digital tutee capable of off-task conversation. Papers II, III, and IV examine the effects on student performance, attitude towards the tutee, and self-efficacy of interacting with a high- or low self-efficacy digital tutee. Students and tutees were deliberately matched or mismatched with respect to self-efficacy.

Digital tutees offer a unique way of providing feedback to students. By observing how well their tutee progresses, students indirectly get information about their own knowledge and learning: something referred to as *recursive feedback* (Okita & Schwartz, 2013). In the papers that follow, recursive feedback is both used in its original sense and a slightly modified form in which the tutee contributes additional, subjective feedback: namely, her 'thoughts' on what happens in the learning activities, including how she succeeds or fails and why. These are expressed in the form of a conversation with the student, and it is through this conversation that the digital tutee displays its self-efficacy, manipulated to be high or low.

For papers I to IV, I used the educational game *Rutiga familjen* (English: "*The Squares Family*"; Pareto, 2014) developed by Lena Pareto at University West, Sweden. The game aims to teach children the base ten system, known to be a bottleneck in mathematics (Sherman, Richardson, & Yard, 2015), by encouraging

students to learn using visual representations instead of numbers (for more information, see Appendix).

Paper V uses the educational game *Historiens väktare* (English: "*Guardian of History*") (see pages 15-20 in Paper V) – which also uses a digital tutee – to gain a wider understanding of feedback and its role by focusing on the constructive but critical feedback provided to the student in text form by a so-called *correcting machine* (part of the game narrative). In one of three conditions in the experiment reported, the tutee visually highlights the feedback text by pointing and looking towards the text box. In a second condition, an arrow signals the feedback text. The third condition was a control condition with no visual signalling of the text. The study's primary research question concerns what we call student *feedback neglect* and the points at which such neglect can occur: in noticing feedback, processing (e.g. read) feedback, making sense of it, acting on it, and using it to progress. The results confirm previous studies indicating that students do neglect feedback to a large extent, while filling in detail on where in the process this happens.

Partly on the basis of paper V, Paper VI studies the ways in which system developers fail to consider feedback when designing educational software – either by failing to consider the rich variety of feedback forms available or failing to implement feedback in a thoughtful manner.

This thesis contributes valuable information to the educational software community: (a) by showing how off-task conversation affects low- vs. high-performing students; (b) by showing how digital tutees' self-efficacy affects students' performance, self-efficacy and interactions with the digital tutees; (c) by exploring the possibility of using agents to increase students' inclination to pay attention to textual feedback; (d) and, on a broader note, by filling in details on the process by which students make use of – or fail to make use of – software-delivered feedback. The final paper is a call to designers of educational software: they must put more focus on feedback if their applications are to be anything more than glorified digital tests.

Introduction to the papers

All papers save one are co-authored. My contribution to each is described below. The papers are presented in chronological order with respect to when the studies took place.

Paper I: Off-task engagement in a teachable agent based math game

This first paper is a conference paper written with my supervisors Magnus Haake and Agneta Gulz. I analysed the chat dialogues, took the lead on writing the paper, and presented it at the *19th International Conference on Computers in Education* (*ICCE 2011*), Chiang Mai, Thailand.

The paper is based on analysis of dialogues between students and their digital tutees in an educational math game targeting basic arithmetic skills related to the place-value system. Tutees' knowledge develops on the basis of what they are taught, what they observe students doing (e.g., when selecting a card), and what answers they receive to their (multiple-choice) questions. We call this interaction *on-task conversation*, since it is focused completely on the task at hand. Additional interaction takes place via free-text chat (*off-task conversation*) where any topic whatsoever can come up, and the tutees themselves raise topics unrelated to the game.

A previous study (Gulz, Haake, & Silvervarg, 2011) found that students with access to the chat function had a more positive experience and achieved better results (judging by how well they taught their digital tutees) than those who played the game without the chat. On closer analysis of high- vs. mid- and low-achieving students, it was found that, in fact, only the high- and mid-achieving students reported a more positive game experience; at the same time, they were likelier to refrain from starting a new chat and quit the chat more often. The present paper is a follow up exploring this seeming paradox through detailed analysis of the chat behaviour, with particular focus on students' engagement. The analysis was driven by two research questions:

Q1. To what extent did students seem engaged in the off-task conversations with their digital tutee?

Q2. What did students do when the chat logs indicate that they were not engaged? Did they quit the chat? Did they start a new chat at their next opportunity?

We looked for differences between high- and low-achieving students concerning the first three chats, which started automatically, and the subsequent chats, which started only at the students' initiative. No significant differences were found between high- and low-performing students with respect to the initial chats. When it comes to the optional chats, this changed: high-performing students showed considerably more engagement even as the low-performing students chatted to a considerably greater extent; the high-performing students would simply quit when they became less engaged and refrain from starting a new chat. In the discussion section, we propose that what is seen here is that high-performing students to a larger extent than low-performing students take control over their own learning.

Paper II: Instructing a teachable agent with low or high self-efficacy: Does similarity attract?

The second paper was written with my supervisors Annika Silvervarg, Agneta Gulz, and Magnus Haake. I took the lead on designing the study, collecting the data, and did a large part of the writing. I wish to give a special acknowledgement to Björn Sjödén, who helped design the study and assisted with the game sessions; along with Ludvig Londos, Axel Duvebäck, and Frida Nelhans, who assisted throughout.

The study took place over eight weeks using nine classes at four schools in Sweden. In all, 166 fourth graders took part. Due to missing data or poor attendance, 24 were removed from analysis, leaving 142, of whom 113 were assigned to either the high (58) or low (55) self-efficacy group based on a pre-questionnaire; the middle 20% (29) were excluded. The game used was the same used in Paper I, but the chat function was redesigned and extended with an eye toward manipulating the digital tutee's expression of self-efficacy.

We were interested in whether and how a digital tutee displaying high vs. low self-efficacy would affect students' in-game performance, self-efficacy, and attitude towards their tutee– both in general and with respect to matching/mismatching self-efficacy between student and digital tutee.

To determine in-game performance, we analysed how well students answered their tutee's multiple-choice questions and how well they chose their cards.

To determine change in self-efficacy and attitude towards the digital tutee, we analysed two different pre- and post-questionnaires.

The results show that students teaching a digital tutee with low self-efficacy performed significantly better overall than those teaching one with high self-

efficacy. We found no effects on students' self-efficacy or attitude towards the digital tutee.

Students with low self-efficacy increased their self-efficacy regardless of whether they had a matching (low self-efficacy) digital tutee, but the increase only became significant when teaching a matching tutee. Low self-efficacy students with matching digital tutees raised their performance level to that of the high self-efficacy students. No corresponding effects were found for students with high self-efficacy, possibly due to a ceiling effect.

This study points towards potential benefits in designing digital tutees with low self-efficacy. More studies need to be carried out, with diverse student populations and improved software design, such as allowing the digital tutee's self-efficacy to change on the fly in light of repeated experiences of success or failure.

Paper III: "I didn't understand, I'm really not very smart": How design of a conversational teachable agent with low self-efficacy can contribute to student performance and self-efficacy

Paper II led to a follow-up study further analysing the dialogues to look for possible mechanisms behind the results in paper II. Paper III was written with my supervisor Annika Silvervarg. A special thanks to Agneta Gulz and Nils Dahlbäck who offered comments. I took the lead with the design and contributed to both the data analysis and the writing.

For purposes of the new analysis, the students were divided into low (45), medium (53) or high (44) self-efficacy groups according to the results of the pre-study questionnaire. The middle group was removed from the analysis. This was done to increase the contrast between the low and high self-efficacy students and to conserve resources, given that coding and analysis are time intensive and costly. Nine additional students were removed due to missing data or scarce attendance, so that analysis was done on data from 89 participants: 47 girls and 42 boys. For comparison of the division into groups in papers II and III, see section "*Clarification note for paper II, III, and IV*".

We were interested in how these students responded to what their digital tutee said and how they perceived her. We were particularly interested in differences between high and low self-efficacy digital tutees, both in general and with respect to (mis-) matching self-efficacy. We were guided by four research questions:

Q1. To what extent, and how, did students respond to digital tutees' feedback on the just-completed round – how well or poorly it went and why – stated expectations for the coming round, and thoughts about learning to play the game better?

Q2. To what extent, and how, did students comment on their tutee's intelligence and abilities?

Q3. To what extent, and how, did students comment on their tutee's attitude?

Q4. Was there any relation between students' chat behaviour and performance?

The results show that, overall, a digital tutee with low self-efficacy received more response on its utterances, in particular more positive responses, and received less criticism of its intelligence and competence. This was particularly true for students with themselves low self-efficacy. Their performance in answering their tutees' multiple-choice questions correlated nicely to the frequency of their positive comments on the digital tutee's intelligence and competence. No such correlation was found for students with high self-efficacy.

The paper discusses possible explanations why students with low self-efficacy benefit more from a digital tutee with low (matching) rather than high (mismatching) self-efficacy. It focuses on role modelling, the importance of social presence and social relations, and a phenomenon called the *protégé effect*.

The term 'protégé effect', introduced by Chase, Chin, Oppezzo, and Schwartz (2009), refers to the fact that students make more effort when asked to learn something in order to teach someone else, compared to learning the same thing for themselves. We propose that instructing a digital tutee with low self-efficacy makes the student put in even more effort since a digital tutee with low self-efficacy comes across as being in more need of help than a digital tutee with high self-efficacy.

The paper suggests that the relationship students form with their digital tutee has an effect on their performance – based on the correlation between how well low self-efficacy students answered the digital tutees' multiple-choice questions, and the extent to which they commented positively on their tutee's competence and intelligence.

The paper attempts to address how a pedagogical agent can be designed with respect to self-efficacy to support learning for the widest possible range of students. Our tentative conclusion, as in Paper II, is that digital tutees should probably be designed to express low self-efficacy.

Paper IV: Supporting low-performing students by manipulating self-efficacy in digital tutees

The fourth paper is a conference paper written with my supervisors Magnus Haake and Agneta Gulz. I did the study design, collected the data, took the lead on writing the paper, and presented it at the 2017 conference of the *Cognitive Science Society* (*CogSci*). The paper draws on the same data as the previous two except that here the focus is strictly on the low-performing students, and so only the 62 students who performed below the median on the place-value-system pre-test are considered. Note that slightly less than half of these students fell into the low self-efficacy group identified in Paper II.

Previous studies have shown that low-performing students often benefit more than high-performing students from instructing a digital tutee. Our aim was to explore which of two explanatory mechanisms – modelling theory (Bandura, 1997) and protégé effect (Chase et al., 2009) – would best predict what happens when low-performing students instruct a digital tutee expressing low self-efficacy versus one expressing high self-efficacy.

A digital tutee with high self-efficacy models a learner with belief in her own ability to learn and succeed. This should make a good role model for low-performing students. Modelling theory predicts that low-performing students should make more progress instructing a digital tutee with high as opposed to low self-efficacy.

A digital tutee with low self-efficacy expresses uncertainty over its ability. It comes across as someone less able and more in need of help. The protégé effect predicts that students should be more engaged with and put more effort into the task, and make more progress, when instructing a digital tutee with low as opposed to high self-efficacy.

The results show that low-performing students interacting with a low self-efficacy digital tutee performed significantly better than the same group interacting with a high self-efficacy tutee. The results shed light on why low-performing students tend to benefit from educational games that include digital tutees. They confirm that at least some of the pedagogical force in a digital-tutee-based game derives from students' ability to attribute social agency to the digital tutees: in this case, concerning tutees' belief in their own capabilities. Clearly, self-efficacy is one trait that digital-tutee designers should keep in mind.

Furthermore, we based our predictions on two different theoretical models: role modelling according to which a teachable agent with high self-efficacy (high SE-TA) (TA standing for teachable agent), should have the most positive influence on the performance of low-performers, and the protégée effect according to which a

teachable agent with low self-efficacy (lowSE-TA), should have the most positive influence on the low-performers performance. The latter theory was supported, and might be further elaborated on, by means of the results of our study. According to the protégée-effect students tend to make more effort and take more responsibility for the task of teaching a digital tutee than for the task of learning for themselves (Chase et al., 2009). In our study the outcome was better when low-performers taught a lowSE-TA compared to a highSE-TA. It is near at hand that they made an even larger effort and took even more responsibility for a digital tutee with low self-efficacy, since this tutee expresses a low trust in her own ability to learn, and likely comes across as someone who is more in need of help than a tutee with high self-efficacy. A highSE-TA, on the other hand, indicates that she is capable to learn and perform, and is in less need of help.

Clarification note for paper II, III, and IV

In paper II and III the students were divided into a high, mid, and low self-efficacy group based on the scores from a self-efficacy questionnaire. In paper II the mid self-efficacy group (corresponding to one fifth of the students) were excluded from analyses, leaving us with 113 students (55 belonging to the low self-efficacy group and 58 to the high self-efficacy group). In paper III the students were, likewise, divided into three groups of low, mid, and high self-efficacy. The the mid self-efficacy group (here corresponding to one third of the students) were excluded from the analyses, leaving us with 89 students (45 in the low self-efficacy group and 44 in the low high-efficacy group).

For paper IV, the target group was low-performing students identified as the students who performed below the median on a pre-test in math, leaving us with 62 students. Out of these 62 low-performing students, only 28 (45%) belonged in (overlapped with) the low self-efficacy group defined/used in paper II. With regard to paper III the overlap of low-performing students (paper IV) with low self-efficacy students was even less, as the low self-efficacy group in paper III was more strictly defined.

Paper V: Looking into the black box of students' (not) handling feedback on mistakes

Paper V was written with Yeon Joo Lee, Richard Andersson, Kristian Månsson, Agneta Gulz, and Magnus Haake. I took the lead on study design, data collection, analysis, and writing the paper.

Many studies on feedback focus on the type of feedback provided and the learning outcomes for the students that receive the feedback. The steps in between – from presentation of feedback to performance in relation to the feedback (making progress or not) – are usually left out (left in the 'black box'). These include at minimum 'noticing the feedback', 'reading or otherwise processing it', 'making sense of it',[1] and 'acting on it' (see figure 2). We were interested in where along the way students fall off: i.e., where *feedback neglect* takes place. We also explored the effects of two ways of signalling the feedback: either a digital tutee pointed and looked towards the feedback or an arrow pointed to the feedback. A control condition included no signalling.

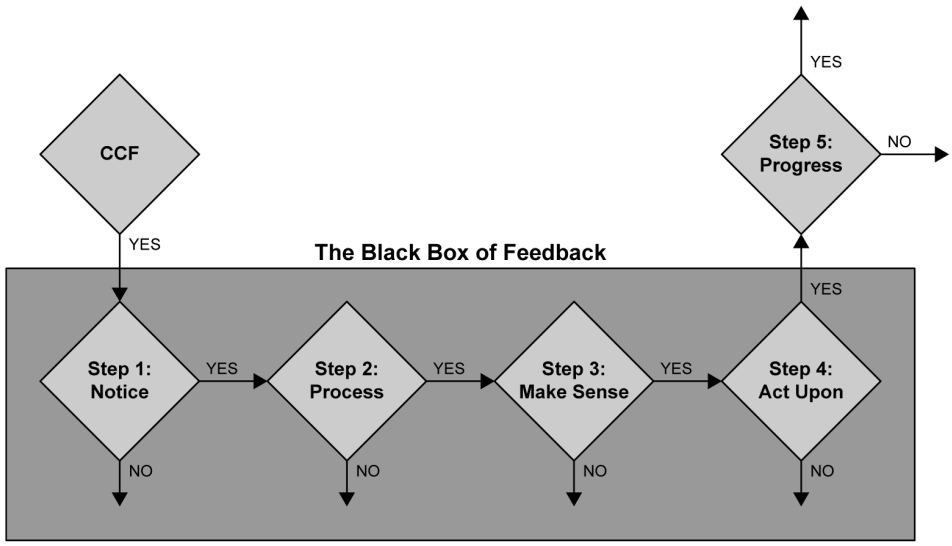


Figure 2. The CCF-processing model developed to study feedback used in this study: 1 (notice), 2 (process), 3 (make sense), 4 (act upon), 5 (make progress).

Forty-six middle-school students used the educational software Guardian of History over three lessons, the first two in their own classrooms and the third at a location at Lund University. Data were collected using behavioural data logs, eye-tracking, and a questionnaire.

We found neglect at each step. In 33% of the cases, students did not even notice the feedback. A further 39% were noticed but not read. Of those that were, 77% were not acted upon. Of those that were acted upon, 52% showed no indication of

[1] This step was not a part of our analysis since it could not be measured with the methods we were using.

increased performance. Of the three feedback conditions, the one with the digital tutor pointing and looking toward the feedback fared best. Students were likelier to notice and then to read the feedback; however, they did not in the end perform any better.

Even though we found no differences at this last step, the results are important. We showed that the methodology enables one to follow what happens at each step. We also showed that signalling matters and that using a digital tutee to do the signalling can make students likelier to notice and read feedback – including those not initially inclined to do so.

Paper VI: Review of feedback in digital applications – does the feedback they provide support learning?

For the final paper, I am sole author. Agneta Gulz assisted with supervision and comments.

This paper stands apart from the others in not being based on a classroom study. It is a review paper examining the forms of feedback available in educational apps used in Swedish schools. My interest in doing so grew out of discussions with teachers at two workshops that I held. Its purpose is not to distinguish good from bad apps but to point out how rarely educational apps are evaluated properly. It discusses what types of feedback to look for and notes the pitfalls that come with some of them.

To determine which apps are frequently in use, I sent an email to approximately forty schools around Sweden asking them to list the apps used at their school. Removing those apps that I did not consider to be educational, I ended up with 103 apps. Counting all their sub-games, I looked at 242 games in all.

The results show that 78% of the apps provide nothing but verification feedback, meaning that they only let the learner know whether their answer was correct or not. Further, 10% of all apps show the correct answer when the learner has provided an incorrect one (verification feedback) and no more than 12% provide the student with elaborated feedback, that is feedback that in some way guides the learner who fails towards the correct answer. In addition, the result show that more than half of all apps provide the learner with some type of encouragement (most often in form of written utterances like "*well done*" or visual features like stars or balloons. Almost all encouragement appears after a successfully completed task. No app encouraged the learner for good effort or partial progress. Rather, the encouragement focused on the learners' abilities/intelligence and not on the task at

hand, something that is completely contrary to what is recommended from a learning science perspective.

I concluded that educational software developers do not take full advantage of the possibilities with technology. Most of the feedback provided in the apps is not well suited for the purpose of supporting learning.

Digital pedagogical agents

Schools increasingly provide students with various educational software that they access most often via laptops or tablets. Students enter virtual environments that can sometimes be quite complex. Some educational software is built around or accompanied by a so-called pedagogical agent: a computer character in a pedagogical role, often based on some form of artificial intelligence. A pedagogical agent can, in principle, be designed to take on any role and act in any way the designer wants.

In addition to pedagogical agents, one finds another type of computer character lurking in this software: the *avatar*. Avatars are characters that represent, and are directly controlled by, people. We often encounter this type of character when we play a video game; the avatar is controlled by ourselves or another player. Pedagogical agents, on other hand, are programmed to display sufficient understanding of the learning context and subject matter to perform a useful role and act on their own (Chase et al., 2009).

Both avatars and pedagogical agents hold benefits for learning. Students can learn to take on attributes of their avatars. Avatars can motivate students to take risks from which lessons can be learned. Pedagogical agents can serve as role models for how to think or act and can themselves be a motivating force, as will be discussed in subsequent sections and in the papers. Most pedagogical agents come with some form of embodiment. They are visually represented and often have a voice. Presentation can vary widely; see Figure 3.

Figure 3. Examples of different visual forms that pedagogical agents can take on.

A key feature of pedagogical agents is that they bring a social component to the learning environment, and researchers agree that social context is critical for learning (Palinscar & Brown, 1984; Lave & Wenger, 2001; Vygotsky, 1980). A pedagogical agent simulates a social presence, mimicking aspects of human interaction and taking on various roles and personas. Pedagogical agents have been studied in a number of roles: tutor (Johnson, Rickel, & Lester, 2000; Veletsianos, 2009), co-learner (Hietala & Niemirepo, 1998; Lee, Nass, Brave, Morishima, Nakajima, & Yamada, 2006), mentor (Baylor & Kim, 2005), and tutee: the focus of this thesis.

Lester et al. (1997) conducted a much cited experiment in which they let 100 middle-school students interact with an animated pedagogical agent, Herman the Bug, in an interactive learning environment in which they learned about botany. There were five versions of Herman, which differed in explanatory behaviour but not in physical appearance. What the researchers found was that, depending on the version, students performed better or worse; but all versions had a strong positive effect on students' perception of their learning experience – what the authors call the persona effect. The authors go on to argue that pedagogical agents have a powerful motivational role to play, and it does not matter whether the agent is expressive or not; its mere presence is what counts.

Although the persona effect has been the frequent target of debate (Heidig & Clarebout, 2011), subsequent studies have confirmed pedagogical agents as powerful educational tools (Moreno, Mayer, Spires, & Lester, 2001; Plant & Baylor, 2005; Kim, Wei, Xu, Ko, & Ilieva, 2007; Pareto, Haake, Lindström, Sjödén, & Gulz, 2012; Johnson, Ozogul, & Reisslein, 2015). What was not apparent from Lester et al.'s (1997) study was the importance of a whole set of characteristics: pedagogical agents' visual appearance and vocal presentation as well as such non-embodied characteristics as competence and feedback. What later studies have found is that the mere presence of a pedagogical agent does not automatically translate to increased student performance or improved learning experience. Various characteristics play a role, with varying effects depending on group.

Visual appearance and vocal presentation

Mayer and DaPra (2012) showed that students who interacted with an embodied agent using social cues such as gesture, facial expression, and eye gaze had superior learning outcomes compared to students who interacted with the same agent lacking such cues. Johnson et al. (2015) found that, when a pedagogical agent used visual signalling (pointing), students in general learned more while low-performing students did not find the tasks as difficult when the agent used pointing.

Baylor, Rosenberg-Kima, and Plant (2006) found that female students working with embodied pedagogical agents reported a more positive attitude towards engineering after working with a female compared to a male agent. They also found that a younger compared to an older agent had a more positive effect on students' self-efficacy towards engineering – but only if the agent was "*cool*" (as judged by hairstyle and dress). The effect was indeed reversed for older agents, who had a more positive effect on student self-efficacy when they were "*uncool*". Veletsianos (2010) showed that an agent's visible characteristics influence both students' perception of the agent and learning. In their study, 94 participants interacted with a pedagogical agent while working through a tutorial on either nanotechnology or punk rock. The agent was designed to look either like a scientist or artist. Overall, participants recalled more information when interacting with the artist compared to the scientist. The results also show that people stereotype pedagogical agents just as they do human beings. Participants expressed the opinion that the agent with a punk-rock-type hairstyle seemed to know more about punk rock just because of its hairstyle (not because it actually knew more).

Two experiments by Atkinson (2002) found that students who learned to solve mathematical word problems in the company of an animated pedagogical agent

who provided oral explanations outperformed students whose agent used only text. Kim, Baylor, and Reed (2003) found beneficial effects from pedagogical agents who had either a strong (authoritarian, assertive, enthusiastic) or calm (soft, nice, kind) voice compared to agents with a 'computer' voice. Students rated the strong and calmly voiced agents more affective, affable, and credible; and the results show that they learned more. Students rated the strong voice most motivating.

Non-embodied characteristics

I have described how pedagogical agents can be designed to be female, male, or androgynous; black or Hispanic; strict or relaxed dressers; and so on. Visual and auditory presentation aside, pedagogical agents can also be designed to be more or less competent, more or less self-confident, more or less inclined to give feedback, varying in style of interaction and conversation in countless ways.

Interaction style

Baylor and Kim (2005) compared three types of pedagogical agents with respect to feedback, in a teaching environment for instructional planning. They found that students interacting with a more encouraging agent were likelier to increase their self-efficacy. Wang, Johnson, Mayer, Rizzo, Shaw, and Collins (2008) showed that a more polite agent (offering hints and phrasing feedback politely) yielded better learning outcomes, especially for students who expressed a preference for indirect help and those of lower ability to start with. Lee, Nass, Brave, Morishima, Nakajima, and Yamada (2006) studied the effects of both emphatic displays and encouraging feedback. Their results show that pedagogical agents expressing emphatic emotions and providing encouraging feedback correlated with better learning (higher recall) compared to more neutral companions that did not provide encouraging feedback. A possible explanation is that if the student feels that she is interacting with a social partner this can make her try harder.

Pedagogical agents can sometimes be perceived as flat and impersonal. Frasson and Aimeur (1996) and Kirkegaard (2016) explored the effects of intentionally more colourful pedagogical agents. Frasson and Aimeur compared a compliant digital peer with a more disturbing and challenging one. They found that the more unpredictable behaviour from the challenging agent affected student learning depending on how confident students were about the task from the beginning. Students with high confidence showed a good learning progression interacting with the challenging agent; students with low confidence did not. Kirkegaard worked with middle-school students instructing a digital tutee on history. The

tutee had one of two communicative styles. Either it was a compliant tutee who accepted everything students proposed, or it would now and again challenge students' answers or explanations. Students were balanced with respect to level of self-efficacy in history. Kirkegaard again found that students with high self-efficacy performed better with the challenging agent, students with low self-efficacy with the traditional agent.

Not only *how* things are said but *what* is said is important. The type of agent used in this thesis is a conversational digital tutee with the ability to talk to students on a diverse range of topics: i.e., engage in *off-task conversation*. Conversational agents in educational software are often able to converse only about the learning domain and the task at hand; but what goes on in most classrooms is a mix of on- and off-task conversation. Small talk cannot just make a situation more relaxed but has been shown to promote trust, build rapport (Bickmore & Cassell, 1999; Cassell & Bickmore, 2003), and provide students with a more positive learning experience (Silvervarg, Haake, Pareto, Tärning, & Gulz, 2011). However, not all students experience off-task conversation as something positive; some find it time consuming and meaningless (Veletsianos, 2012).

Paper I considers the effects of off-task conversation on high vs. low performers' engagement. Paper III considers further effects of off-task conversation in relation to students with low self-efficacy.

Competence

Kim and colleagues conducted a series of studies looking at the effects of pedagogical agent competence and intelligence on student self-efficacy and learning. Baylor and Kim (2004) studied twelve pedagogical agents differing in gender, ethnicity, and competence. They found that agents perceived as less intelligent led to increased student self-efficacy. Kim, Baylor, and PALS Group (2006) studied pedagogical agents of high vs. low competence and of responsive vs. proactive interaction style, measuring the effects on students' self-efficacy, learning, and attitude towards the agent. They found that a more proactive and competent agent had a positive impact on student learning, while a less competent agent had a positive effect on self-efficacy. More competent agents produced more positive attitudes from the student toward the agent. Kim (2007) again examined more vs. less competent agents in relation to high- vs. low-performing students. Academically stronger students showed higher self-efficacy in relation to the learning task and recalled more after working with the more competent agent. Academically weaker students showed higher self-efficacy and recalled more after working with the less competent agent. Hietala and Niemirepo (1998) had earlier reported similar results comparing a more with a less competent companion, who also differed in confidence: the more competent agent was more certain when

suggesting a solution to a problem, the less competent agent more hesitant. Participants were grouped according to extroversion vs. introversion and IQ. Extroverted students with lower IQ preferred to work with a less competent agent, introverted students with higher IQ preferred a more competent agent.

Similarity attraction

Human beings tend to like people they perceive as similar to themselves: a phenomenon known as similarity attraction (Newcomb, 1956; Byrne & Nelson, 1965; Byrne, Griffitt, & Stefaniak, 1967; Nass & Lee, 2001). This is not unique to human-human interaction but can be seen in human-computer interaction as well. In various ways, people are affected by how similar they perceive the agents to be to themselves (Hietala & Niemirepo, 1998; Moreno & Flowerday, 2006; Plant, Baylor, Doerr, & Rosenberg-Kima, 2009; Rosenberg-Kima, Plant, Doerr, & Baylor, 2010; Ozogul, Johnson, Atkinson, & Reisslein, 2013).

Moreno and Flowerday (2006) allowed a group of American students to choose which animated pedagogical agent to interact with from a choice of female vs. male and five ethnicities. Students of colour chose significantly more often than white students to work with an agent of the same ethnicity. Rosenberg-Kima et al. (2010) studied African-American females learning engineering with a female or male agent who was either African-American or white. When students interacted with an agent of the same ethnicity, they were less likely to endorse gender stereotypes regarding engineering; they also rated the need for engineering and their interest in the subject higher. When the agent was also of the same gender (i.e., female), participants judged the agent more likeable and reported a higher self-efficacy.[2] Pratt, Hauser, Ugray, and Patterson (2007) found that students changed their opinion to align with the agent to a greater degree when the agent had matching ethnicity. On the other hand, Behrend and Thompson (2011) found no effects for ethnicity in their study where a digital trainer supported students in an Excel™ exercise.

In their study, Isbister, and Nass (2000) found that participants tended to prefer a character whose personality was judged complementary rather than similar to their own. So similarity attraction is, at least, less than straightforward.

Nevertheless, similarity attraction has great potential for use in educational contexts. Bandura (1997) writes about what he calls vicarious experience: observing someone, judged similar to oneself in important respects, perform a task

[2] If the agent was white, participants reported higher self-efficacy if the agent was male.

supports one's own learning of the task (see Section "*Self-efficacy and performance*").

Papers I, II, and III all address whether the agents' self-efficacy should be taken into account when designing digital tutees. To my knowledge, this has not been studied before. Hietala and Niemirepo (1998) come closest: their agents displayed greater or lesser certainty in providing suggestions to their students, in keeping with their level of competence. However, the agents' expressed certainty was not treated independently, and so the study could not distinguish the effects of certainty from competence.

Feedback and pedagogical agents

A pedagogical agent typically seeks to guide students through the learning process and can offer feedback when the student is heading in the wrong direction. What type of – and how much – feedback an agent provides depend to large extent on what role the agent has, but also on the designer´s choices.

Azevedo et al. (2012) placed 69 students in an adaptive hypermedia learning environment targeting the human circulatory system. Pedagogical agents provided varying levels of prompting and feedback. The students were divided into three groups. Agents prompted the first group to use self-regulatory learning and offered immediate feedback. The second group received the same prompts but no feedback. The third group – the control group – received neither prompts nor feedback. Students receiving both prompts and feedback achieved significantly higher scores than those in the other two groups.

Letting students know where they are in their learning process – making them attend to their difficulties and helping them solve tasks – can reduce frustration and cognitive load. Moreno (2004) investigated the consequences for students of receiving explanatory vs. corrective feedback. In the one experimental condition, the agent said if the answer was correct and attempted to explain the student's choice; in the other, the agent only communicated whether the choice was correct. Students in the first group received higher test scores – stemming, Moreno argues, from the decreased cognitive load. Providing students with more information than just 'correct' or 'incorrect' can help them focus on the right things. Rather than wondering about what they did wrong, they can focus on correcting and learning from their mistake.

Leelawong and Biswas (2008) studied the effects of providing self-regulatory compared to corrective feedback. Under two experimental conditions, students taught a digital tutee. In the first condition ('self-regulatory'), the tutee remarked upon her learning with the purpose of teaching the student to be more self-aware.

In the second, the tutee offered only corrective feedback (when the student made a mistake). A third group acting as control group and were instead taught *by* an agent and received corrective feedback as in the second group. Results show that students in the two groups teaching a tutee performed better than the students in the control group. In addition, students who received self-regulatory feedback appeared to be better prepared to learn in new domains.

The feedback provided in the two teachable-agent-based games discussed in this thesis is different from the prototypical notion of 'feedback'. It is not feedback provided directly to the student with respect to her performance, but feedback provided indirectly to them as they observe how well their digital tutee performs: the aforementioned recursive feedback (Okita and Schwartz, 2013; see also Section "*Learning by teaching*").

Effects of varying characteristics in pedagogical agents on different student groups

Decisions about agent characteristics can influence the effects it has on the student interacting with it. However, as already been shown, the effects may differ between student groups. Silvervarg, Haake, and Gulz (2013) let 108 students interact with two digital tutees (from a choice of three) over two 45-minute sessions. One was designed to look like a boy, one like a girl, and one androgynous. The female students had the most positive attitude towards the androgynous agent, whereas the male students were equally preferential towards the androgynous and male agents. Some other studies that show that boys and girls tend to be affected in different ways by certain design choices are (Gulz & Haake, 2010; Arroyo, Woolf, Cooper, Burleson, & Muldner, 2011).

Further, consider high- versus low-performing students and those who have greater or lesser knowledge of a subject. McLaren, DeLeeuw, and Mayer (2011a) found that students with little or no prior knowledge of chemistry learned more interacting with a polite compared to more direct tutor: what the authors call the politeness effect. No effect was found for students who had prior knowledge. The authors argue that students with little or no prior knowledge are more likely to respond to social engagement by engaging more deeply with the material. Students with prior knowledge simply may not need social engagement the same way, and already have a clear strategy of how to integrate and organize novel information. A second study of 132 students interacting with the same (web-based) tutors (McLaren, DeLeeuw, & Mayer, 2011b) found that the students who benefitted the most from the polite tutor were those that during the task did most errors.

Woolf et al. (2010) examined whether low-achieving students benefited from interacting with a math learning companion who provided affective feedback. High- and low-performing students interacted with either a male or female agent. A control group received the same feedback but with no agent present. Students in all three groups showed improvement. Interacting with the agent influenced all students positively, and low-achieving students in particular improved their confidence. Low-achieving students compared to high-achieving students also reported more positive feelings when the learning companion was present.

In this thesis, I have chosen to focus on learning effects for two student groups: low-performing students and those with low self-efficacy. These two groups have the most to gain from a well-designed pedagogical agent (see sections "*Low performing students*" and "*Low self-efficacy students*").

Digital tutees

As mentioned, digital tutees are computer agents that students teach, in the process learning for themselves (Biswas, Leelawong, Schwartz, Vye, & TAG-V group, 2005). A digital tutee can be considered something of a hybrid between pedagogical agent and avatar since it is controlled and influenced by both computer and human user (Chase et al., 2009). Artificial intelligence of some form allows the digital tutee to reason, answer questions, and otherwise show independent behaviour. At the same time, the student instructing it also exercises control, and its knowledge is a direct reflection of what the student has taught it. The tutee reasons logically but may nevertheless reach the right or wrong answer.

The benefits of digital tutees are many. For example, their reasoning can be visualized so that the student can follow the reasoning process; an example of such a digital tutee is 'Betty' (Biswas et al., 2005); see Figure 4. The student teaches Betty by creating concepts maps of nodes and relationships, which the student chooses from a predetermined set. After having instructed Betty, the student has the opportunity to query her to see how much she has understood or let her take a quiz to see how well she does on questions the student might not have considered (Leelawong & Biswas, 2008). Betty then explains her answers by using stylized natural language or highlighting the relevant portion of the concept map (Biswas et al., 2005).

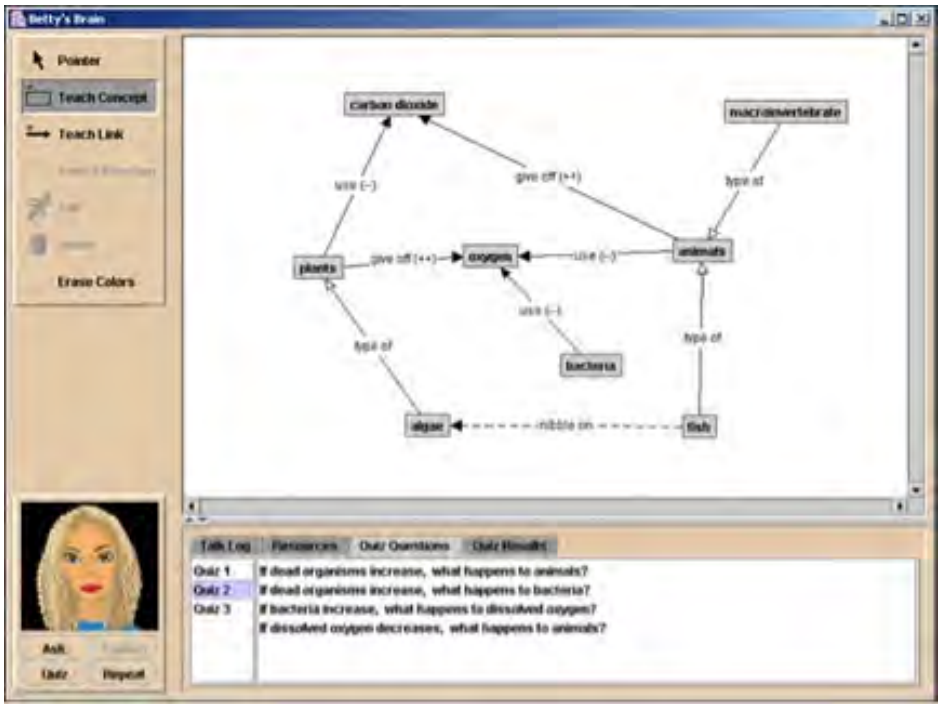


Figure 4. An example of a student instructing Betty by constructing a conceptual map.

Having a visual representation of how someone else reasons about a topic enables students to reflect on their own thinking and make comparisons. Supposing that the agent is an expert in reasoning, this brings the possibility of the student adopting a more effective way of thinking.

Learning by teaching

Already before the Christian era, the Roman philosopher Seneca the Younger coined the expression *Docendo discimus* (Latin: "*by teaching we learn*"). This is probably something everyone has experienced: to teach is to learn. Despite the Roman's wisdom, it was not until Bargh and Schul's (1980) seminal paper that the idea was put to the empirical test. Bargh and Schul demonstrated that participants who prepared to teach a text – so someone else could take a quiz – learned the material better than those preparing to take the quiz themselves.

Even though the participants in that study did not actually teach someone else – they only anticipated doing so – they clearly benefited. Probably the anticipation

made them focus better on those parts they believed to be most important and generally try harder (Bargh & Schul, 1980; Benware & Deci, 1984; Renkl, 1995[3]). Selecting and organizing information means having to rehearse what one knows – something that is always good for learning (Okita & Schwartz, 2013).

The act of preparing to teach is a metacognitive activity: one should anticipate learners' needs and questions (Okita & Schwartz, 2013). Schwartz et al. (2009) claim that this kind of metacognition – reflecting over someone *else's* thinking – is less demanding than reflecting over one's own, even as it may improve one's self-reflection.

Annis (1983) compared students who were only expected to teach but did not actually do so with those who actually taught. A control group learned only for themselves. The students who expected to teach but did not do so showed enhanced learning compared to the control group. The students who actually taught showed enhanced learning compared to the first experimental group. What they gained was needing to explain and answer questions, known to be beneficial for learning (e.g. Webb, 1989). Doing so is an exercise in reflective knowledge building as one uncovers misconceptions, explains things in different ways (Chi, Siler, Jeong, Yamauchi, & Hausmann 2001; Palinscar & Brown, 1984; Uresti, 2000), and – in the process – discovers gaps in one's existing knowledge (Graesser, Person, & Magliano, 1995; Uresti, 2000; Roscoe & Chi, 2007). Annis (1983) showed that teaching was more beneficial than merely preparing to teach; still, it remained to be shown which parts of the process are most beneficial.

In his doctoral thesis, Fiorella (2013) presents four experiments using two experimental groups – one expecting to teach, the other expecting to teach then actually teaching – and a control group who learned only for themselves. The students in the two experimental conditions equally outperformed those in the control group when tested after a short delay. Tested after a one-week delay, students in the actual teaching group outperformed the other two groups. Fiorella argues that preparing to teach helped students better manage their basic processing of the material, with solid short-term benefits; while actual teaching fostered deeper generative processing: diving into the material, something that is critical to long-term learning.

Although all the reasons behind the value of learning by teaching are not mapped out, uniform consensus is that it is powerful. Being asked to teach someone else assigns one a responsibility, and this responsibility is important to most people, if not everyone. Chen, Shohamy, Ross, Reeves, and Wagner (2008) showed that

[3] Renkl (1995) notes some drawbacks with this anticipation though, such as decreased intrinsic motivation and increased anxiety.

believing an experience is social activates the brain's reward system, and that helps cement the learning of new associations.

Chase et al. (2009) found that students asked to teach someone else – in their case, a digital tutee – put more effort into the task. This is the aforementioned protégé effect. They propose three key factors: i) the task is social, ii) students have a clear sense of responsibility, and iii) students can share failures with their tutee, which shields them from forming negative thoughts about themselves: what the authors call the *ego-protective buffer*. Other contributing factors that have been identified include:

- Needing to actively rehearse what one knows, so as to convey it to one's student(s) (Okita & Schwartz, 2013).
- Being forced to externalize one's thoughts; questions from students can put things in new perspectives (Webb, 1989; Chi et al., 2001; Palinscar & Brown, 1984; Uresti, 2000).
- Discovering that one is, indeed, capable of teaching someone else, which affects one's own self-efficacy concerning the subject (Bandura, 1997; Riggio, Fantuzzo, Connelly, & Dimeff, 1991).

These factors, of course, concern both traditional and computer-based learning by teaching. The latter does pose certain advantages though. In the computer game, every student can take the role of teacher, including those who would not get the opportunity to teach someone else: either because they are not knowledgeable enough (low performing) or because they do not believe in their own abilities (low self-efficacy). Also, teaching a digital tutee means that no harm is done if the tutee is taught the wrong things. The digital tutee need never grow tired or stop listening; she can be taught the same things over and over again. Digital tutees can be hand-tailored to the student in terms of e.g. level of knowledge and responsiveness.

Recursive feedback

When a teacher is part of a full cycle of teaching, she not only prepares and then teaches but also gets to observe their student(s) use what they have been taught. Okita and Schwartz (2013) consider this last step an important but neglected part of a teacher's learning experience. It is possible that their students have misunderstood something; but it might also be the teacher who misunderstands or has a knowledge gap.

As noted earlier, the most familiar way to think about feedback in a learning situation is in terms of direct feedback in response to a learner's performance,

ideas or thoughts: a student writes a report and hands it in to the teacher, who hands it back with comments. Feedback can likewise be delivered by a computer program or by direct consequence of one's interactions with one's environment: e.g., a high jumper who tries a new technique only to break the bar.

With learning by teaching, indirect feedback takes pride of place as the student serves as role model for her tutee – in the case of this thesis, a digital tutee – to imitate. The student is simultaneously teacher and learner. Consider a student Susanne who instructs a digital tutee on history: specifically, what factors led to the start of World War I. How well Susanne performs her teaching duties reflects her own learning, and both can be directly evaluated; but it is also possible to evaluate the tutee: if the tutee performs well, Susanne's teaching can be assumed to have paid off; if poorly, then she has room for improvement[4]. That extends to Susanne: she can, by observing her digital tutee, infer how well she is doing as a teacher: i.e., she receives *recursive feedback*.

Okita and Schwartz (2013) present two experiments showing the value of such recursive feedback. In the first, four groups of students were all told to study a text about the mechanisms by which humans maintain a fever, for teaching later to another student. One group (experimental condition) prepared, taught the other student, then got opportunity to observe how their student answered a short quiz. Of the three control conditions, one prepared and taught but did not have opportunity to observe; one prepared and then observed but never taught; and one prepared but neither taught nor observed. Compared to the control groups, students in the experimental group fared better on the post-test: i.e., the recursive feedback brought *added value*.

In the second experiment, the researchers showed that the positive effects of observing one's tutee extended to the digital realm and also how recursive compared to direct feedback was beneficial for learning. Students played a computer-based game in which they must induce a rule from available evidence then express the rule in a competition of increasing complexity. To pass a level, one had to out-perform the Evil Moby. In one condition, students played the game themselves and received direct feedback. In another, they watched the digital tutee they had taught play the game using the rule they induced earlier (recursive feedback). Students who received the recursive feedback showed greater ability to solve novel logic problems. Observing the success of one's tutee encourages

[4] It is possible, of course, that the tutee performs poorly but Susanne learns well. It is even possible that Susanne has, for some reason (perhaps as a game with herself), intentionally taught her tutee faulty things – but that requires knowing the *right* answers.

reflection, which in turn leads to deeper cognitive processing than simply noting if one's answer was correct (Chin, Dohmen, & Schwartz, 2013).[5]

Recursive versus direct feedback

Receiving recursive instead of direct feedback reduces impact from two major threats to feedback success. The first involves the cognitive complexity in interpreting the implications of feedback. The second involves feedback as an affective threat (Okita & Schwartz, 2013). How well one interprets feedback depends to great extent on how well one knows the subject. If one is well-acquainted with it rather than being a novice, it is easier and less threatening to take critical feedback (Kluger & DeNisi, 1998; Kluger & DeNisi, 1996). A professional singer, who is told to sing one octave higher, knows just what to do, whereas a novice receiving the same feedback, may not understand what to do with it. Learning by teaching helps students familiarize themselves with the task or subject which can makes it easier for them to relate to feedback.

Many tend to perceive critical (or negative) feedback on their performance or ideas as a threat, rather than as a possibility of improving (Kluger & DeNisi, 1996). If critical feedback is taken as directed towards oneself as a person, it risks leading to so called learned helplessness or social unwarranted comparisons (Hattie & Timperley, 2007). The point of the feedback – as a constructive guide to help the learner to improve - is lost and focus shifts to the wrong thing, namely the person or the ego. In effect, the learner can even start to avoid the situation that generates the feedback (Hattie and Timperley, 2007). But making use of negative (or critical) feedback is an important metacognitive strategy as a way to learn (Chin et al., 2013). Working with a digital tutee in an educational game mitigates affective threat by providing students with the aforementioned ego-protective buffer (Chase et al., 2009): the digital tutee, *not* the student, is tested for knowledge. If the tutee fails, the failure does not come back as hard on the student; the responsibility is shared.

One subject where students are often particularly sensitive to feedback (and in addition, often adopt a fixed view of intelligence) is mathematics. The subject poses a number of familiar stumbling blocks; one is the base ten system: the focus of papers I-IV. Already by the age of three or four, some children start falling behind their peers in the area of early math, due to a weak interest in numerosity (Hannula, Mattinen, & Lehtinen, 2005) and an underdeveloped number sense (Griffin & Case, 1997; Griffin, 2004a; Griffin, 2004b; Jordan, Kaplan, Ramineni,

[5] Recursive feedback is not limited to teaching situations; it also arises where e.g. someone has built their ideas into a design and then observes someone else use the final product.

& Locuniak, 2009). Falling behind early on raises the risk of also not succeeding at mathematics in the long run. Interventions to support students – of all ages – that are at risk of falling behind are therefore important. Learning by teaching, is one available intervention tool.

As noted before, the way I am using 'recursive feedback' in papers II and III is not quite the same as Okita and Schwartz; as recursive feedback, I include subjective thoughts from the digital tutee on the previous game: e.g., *Awesome! We won! I have a good grip now of tens and hundreds and all that you teach me or I´m learning the rules in the math game slowly; I´m not a very brilliant student*.

Agent effects on differing student groups

Let me return to the topic of adapting to student needs. As said, a school class is far from a homogenous group of people who respond the same ways and share the same knowledge, abilities, and preferences. Pedagogical agents offer a way to cater for some of the diversity. At this point, I would like to say more about the two groups that have been the focus of my attention: low-performing students and those with low self-efficacy.

Low-performing students

Again, research shows that low-performing students benefit more than high performers from teachable-agent-based educational software. Chase et al. (2009) found that students using a biology game that included a digital tutee spent more time and learned more, but the effect was most pronounced for low performers. In a study by Sjödén, Tärning, Pareto, and Gulz (2011), a group of nine-to-ten-year-olds used a teachable-agent-based math game, in school, over the period of seven weeks. Afterwards, students were divided into a low- and a high-performing group on the basis of a math pre-test and randomly assigned one of two post-test conditions: with a digital tutee present or without one present. The digital tutee observed what the student did and tried additional tests 'on its own' where it performed just as well as the student. With the digital tutee present – and only with it present – low performers improved significantly more than high performers from pre- to post-test. In similar fashion, Pareto, Schwartz, and Svensson (2009) found greater improvement (though below statistical significance) for low-ability students using a math game with a digital tutee present.

Several explanations as to why low performing students benefit more than high performing students from using a teachable agent-based learning game have been proposed, where some have already been mentioned in the section "*Learning by teaching*".

First, a teachable-agent-based game, the student is positioned as the most capable, teaching someone less knowledgeable – something known to influence students' view of their own competence positively (Riggio et al., 1991). Since low-performers are less likely than high-performers to have taken a teacher position before – and having had the experience of being the more knowledgeable one – the benefit is likely to be higher for low performing students.

Second, a digital tutee is typically designed to model productive learning behaviour: being curious, asking questions, revealing reasoning visually or

verbally (Blair, Schwartz, Biswas, & Leelawong, 2007). A high-performing student is likelier already to have such behaviour in her repertoire. The low performer benefits from observing and, hopefully, imprinting the behaviour into her own.

Third, a digital tutee typically knows nothing (or very little) about the subject from the beginning, needing to learn step by step – encouraging in students what Dweck (2000) calls a growth mindset in contrast to the commonly observed fixed mindset (knowledge and intelligence as inflexible properties). Particularly when it comes to mathematics, students tend to think of themselves and their peers as either gifted or not. Too many students have little confidence in their ability to progress. To make matters worse, mathematics teachers are likelier than other teachers to talk about students being 'talented' or not, reinforcing these feelings (Rattan, Good, & Dweck, 2012). Of course, both high- and low-performing students can have a fixed mindset and so benefit from observing knowledge and intelligence as properties that can and do change; but the mindset is more common (and more ingrained) among low performers. They become trapped in a vicious circle where they think they are not talented and see no point in making an effort, therefore make little effort, do not achieve much, and find the confirmation that they are not talented: a self-fulfilling prophecy.

Fourth, there is the ego-protective buffer, proposed by Chase et al., (2009), also discussed in section "*Learning by teaching*". In short, some of the responsibility for a possible failure can be shared between student and tutee, that then acts as an ego-protective buffer. Both high and low performing could possibly benefit from an ego-protective buffer but low performing students probably to a higher degree since they are more likely to fail in school than high performing students and to have negative feelings associated with such failing.

In paper IV we explored if one out of two explanatory mechanisms could predict what would happen when low-performing students taught a digital tutee with low- or high self-efficacy. The two mechanisms were (i) the modelling theory (Bandura, 1997) and (ii) the protégé effect (Chase et al., 2009). Based on the modelling theory we predicted that the low performing students would benefit more from interactions with a digital tutee with high self-efficacy since this tutee models someone who is able and whom the student could copy. Based on the protégé effect we predicted the opposite: that the students would benefit most from interactions with a low self-efficacy digital tutee. This since this tutee comes across as someone in more need of help, so that the student has to put more effort into teaching.

Self-efficacy and performance

Low performance is, on a *group level*, related to what is called *low self-efficacy*. Self-efficacy refers to peoples' belief that they can succeed in a particular task (Bandura, 1997). Importantly self-efficacy refers to task specific confidence. It refers to your belief as to whether you will succeed when performing a specific task, for example performing in a jazz dance class, giving a particular speech in a social science class, or solving math tasks that involve the place value system. You can have high self-efficacy with respect to one domain but not another. Self-efficacy is not to be confused with self-esteem: one's overall self-evaluation, which tends to be more stable over time. Self-efficacy is easier to influence and, indeed, change. Bandura identifies four primary sources for self-efficacy: performance, vicarious experience, persuasion, and physiological response.

Performance is a relatively reliable indicator of self-efficacy. In general, succeeding with a task raises self-efficacy; and failure lowers it. However, an occasional success/failure is unlikely to have much impact after a number of failure/successes.

High self-efficacy often helps a student enter into a healthy circle of reinforcement, low self-efficacy into a something of a downward spiral. Students with a high self-efficacy in a domain tend to participate more readily, work harder, achieve more, and to persist longer when encountering difficulties and achieve at higher levels (Bandura, 1997). This, in turn, tends to lead to success and progress. Students with low self-efficacy feed on negative thoughts, think of the task as threatening instead of challenging, and set low objectives (Md. Yunus & Wan Ali, 2008; Bandura, 1997).

However, it needs to be pointed out that the relationships between self-efficacy, on the one hand, and academic motivation, learning and achievement, on the other hand is one of *group level* correlation (Stajkovic & Luthans, 1988; Pajares, 2003; Schunk, 1995; Bandura, 1997). In particular, high self-efficacy in a domain does not equal high performance in the domain, nor does low self-efficacy equal low performance. Some students claim confidence in a task but have little knowledge how to go about it. It also happens that students who do not expect to succeed are in fact well-prepared for succeeding (apart from their low belief in their own capability).

In other words, self-efficacy can to *some* extent predict performance; but performance also depends on skills, abilities, and effort: simply believing (or assuming) that one will succeed is never enough. Reversely it is *possible* to succeed in a task even if one has very little confidence that one will.

Collins (1982: in Schunk & Meece, 2006) identified both high and low self-efficacy students among high, middle, and low performers in mathematics. The common dataset reported in papers II, III, and IV reveals a similar pattern: the base ten pre-test we used identified high vs. low performers, while the self-efficacy questionnaire found students of low, middle, and high self-efficacy within each of these groups. Only 29 of 62 students in the low-performing group were also classed as low self-efficacy.

I mentioned persuasion as a factor influencing self-efficacy. Hearing "*You can do it!*" certainly can help; but hearing it repeatedly at the same time as trying and failing is not likely to result in increased self-efficacy. Result in increased self-efficacy. In the paper "*Beyond 'You Can Do It!' – Developing Mathematical Perseverance in Elementary School*", Bass and Ball (2015) argue that it is not students' attitude as such that should be targeted – at least, not solely. Trying to change people's self-efficacy or mindset will not be fruitful unless one also gives them the support they need for having new experiences: experiences that help them manage the task at hand. They require encouragement to persevere instead of giving up whenever the going gets hard. At the same time, the support and encouragement need to be appropriately measured: offering too much support can be interpreted by the student as "*I am not capable of succeeding on my own.*"; and that could lower self-efficacy.

I also mentioned vicarious experience playing a role. Observing someone else succeed can – in the right circumstances – generate an expectation that I, too, can succeed (Bandura, 1977). This is likelier to work if the observer perceives sufficient similarity with the one observed: similarity of gender, ethnicity, and competence being of particular importance. The inverse is true as well: observing others that one perceives to be similar to oneself *fail* can deter one from attempting the task oneself (Schunk, 1987). Of course, observing someone else succeed (or fail) at a task has a weaker effect on self-efficacy than succeeding (or failing) oneself.

Low self-efficacy students

As I mentioned, there is a correlation between low-performing and low self-efficacy students. I wish to turn my attention now to students who have low self-efficacy. We know that students sometimes do not believe that they are capable and have the ability to learn. There is a distinction between 'belief in ability to learn' and 'belief in ability to perform' (Panton, Carr, & Wiggers, 2014), even though there is a grey zone here in that performance depends on practicing and learning. Panton et al. (2014) argue that self-efficacy for learning and self-efficacy for performing are equally important toward motivating persistence in the face of

repeated failures. A student may have a well-founded belief in her inability to perform a task – in which case, one does not want to boost her self-efficacy for performing (not straightaway) but rather her belief in the point of trying.

It also can happen that someone has high self-efficacy that is mismatched to their actual competence in the domain. Consider Markus who has quite high self-efficacy in biology but lacks the abilities to justify it. What he needs is not even higher self-efficacy. The problems with overestimating one's abilities are obvious, not least that students may not reach out for the support they need. Several studies on self-efficacy have found that young children often have high self-efficacy concerning difficult tasks, and sometimes persist in overestimating their capabilities even when provided with feedback indicating low performance (Schunk, 1995). (Conversely, although less frequently, young children may underestimate their capabilities and even believe they cannot acquire basic skills. The problem with children overestimating their abilities often has to do with the fact that they don't fully understand what is required to execute a task successfully.) 'Providing support for students to increase their self-efficacy' is therefore not an overall goal. As Bandura (1997) puts it "The objective of education is not the production of self-confident fools." (p. 65).

Mindset in relation to self-efficacy

Believing oneself capable of learning fosters a growth mindset: abilities are open to change for the better. Believing one's fate set by innate abilities or capacities that are often assumed to be fixed, like talent or intelligence, fosters a fixed mindset instead (Dweck, 2000). Fixed mindset and low self-efficacy with respect to one or another domain often stand in a viciously circular relation. If you perform below average in a domain, say mathematics, plus believe that there is nothing you can do about it (since performance in mathematics depends on talent or intelligence which is innate and fixed) it is likely that you will also foster a low self-efficacy in mathematics. You don't see it as meaningful to make an effort and refrains from doing so. Hence it becomes less likely that you will succeed (since you do not really try). In turn, the weak beliefs you had in your ability to perform will be reinforced.

It should be observed that mindset is often discussed in general terms, as relating to the person´s view on how intellectual abilities in general come about. Someone´s self-efficacy beliefs, on the other hand, relate to the person´s view on how intellectual abilities in a certain domain develop in the person herself. It is, thus, *possible* to maintain a growth mindset while – simultaneously – hold that for a particular task or area, it will not matter how much work and effort one puts into the task or area.

A digital tutee per se can foster a growth mindset since it models someone who learns incrementally, step by step. A more advanced design of the digital tutee than the one used in the studies behind papers II-IV could – as noted earlier – have made the tutee's self-efficacy responsive to its experiences of failure and success. Such a digital tutee might be an even more powerful tool for fostering a mindset of growth.

Summary

Pedagogical agents are increasingly used in educational software. Not only can they increase student learning and motivation; they are highly adaptable, practically chameleons. They can be designed to look and behave exactly as the designer wants. This is good news. However, more research needs to be done into which characteristics have what effects on which groups of students. To my knowledge, no previous research has addressed self-efficacy in digital tutees: in particular, the consequences of giving a digital tutee high or low self-efficacy (papers II-IV). My focus has been on the two student groups I believe are most in need of support: low-performing students and students with low self-efficacy. The results I report in paper II to IV suggest that a digital tutee with low self-efficacy has the greatest potential to increase student performance and self-efficacy. The results presented in paper I indicate that off-task conversation in a teachable-agent-based game affects high- and low-performing students differently.

Feedback

The previous section discussed the *recursive feedback* (Okita and Schwartz, 2013) that a digital tutee provides. The student instructs the digital tutee; the tutee's behaviour solving various tasks and reasoning about them provides information on the quality of the instruction received. The three teachers in the examples below all offer *feedback* in the sense that one more commonly thinks about it. A student offers her solution to a problem, and the teacher 'feeds back' information about the solution.

Student: *5 + 2 = 8*.

Teacher 1: That was almost correct; five plus two equals seven.

Teacher 2: That is incorrect.

Teacher 3: What happens if you have five fingers (showing hand) and then add two. How many do you have?

In the examples above, the form differs, but the purpose is the same: to give the student information about the quality of her answer and/or guide her towards the 'right' answer. The three examples above represent so called *critical* or *negative feedback*, telling the student there is a discrepancy between her solution and the one that the teacher expects (the goal state). In contrast, *positive feedback* indicates that the student has reached the goal state: on this point, at least, no further learning is needed (Schwartz, Tsang, & Blair, 2016). My focus will be on negative feedback.

That feedback is beneficial for learning has been demonstrated in many studies over the years (Black & Wiliam, 1998; Hattie & Timperley, 2007; Shute, 2008). If one never receives information how one is doing, how would one know? Unless one is a very well-disciplined, self-confident learner, one would probably just get lost.

Feedback is important as a *scaffolding device*, guiding learners towards a learning or performance goal. But even though previous research has shown that feedback often is beneficial this is not to say that all feedback is positive for learning. The wrong kind of feedback can have negative consequences for performance, self-esteem and motivation (Harlen & Deakin Crick, 2003; Black & Wiliam, 1998; Kluger & DeNisi, 1996).

The informational content of feedback can encompass a whole range (Schwartz et al., 2016) from effectively empty feedback through simple feedback that indicates

'right' or 'wrong'[6], feedback that names a specific discrepancy, and elaborative feedback that *elaborates* on just where the error resides (why a proposed solution is inadequate) – to feedback overkill. Just learning that an answer was incorrect may in some situations be enough for high-performing students, who have the background knowledge and tools to deduce the 'right' answer; but that will often not be sufficient for low performers[7]. Both low- and high-performing students will likely benefit from elaborative feedback, but the appropriate level of elaboration will probably not be the same.

Figure 5. A continuum of the information content in feedback (Schwartz et al., 2016). As is often the case in learning, there is a sweet spot that resides somewhere between 'too little' and 'too much'. *Figure courtesy of Swartz et al. (2016)*.

Unfortunately, there is no perfect guide to which feedback is most appropriate when. Feedback given to ten people will likely be perceived in ten ways, depending on e.g. personality and pre-knowledge. As a starting point though, consensus is that feedback should generally be more informative rather than less (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; McKendree, 1990; Moreno, 2004).

Inadequate or improper feedback can lead to rote learning of the 'right' answers. Especially for students with limited prior learning, rote learning can be problematic later on. At one stage in the learning process it is okay or even advantageous to be able to recite the multiplication table by heart; but if one does not understand what 'three times five' actually means (5 + 5 + 5, but just as well 3 + 3 + 3 + 3 + 3), then more complex multiplication problems will pose a challenge.

[6] Feedback that says nothing more than "*right*" is not terribly informative, yet it does tell the learner they are on the right track; feedback that says only "*wrong*" at least informs the student of a knowledge or performance gap.

[7] It depends somewhat on the task. If a student spells "*chimney*" with two m's, then simple feedback is probably enough. If the student proposes that photosynthesis starts with oxygen, a more elaborate explanation is required.

Feedback as scaffolding

Feedback in an educational context means that there is information – from a teacher, a digital system or the environment as such – on the quality of an action, an answer, a proposal, a text, etc. provided by a student. Information is 'fed' back to the student in response to what she has done. A teacher (or a digital system) can also be more proactive and provide information beforehand to a student in order to guide and support her. A useful term that includes feedback but also contains more than feedback is scaffolding.

Vygotsky (1980) described how learning can be enhanced through interaction between someone more experienced and the person learning: e.g., a parent and her child or a teacher and her student. With that assistance, a learner can move from her current performance level to her potential level. Wood, Bruner, and Ross (1976, p. 90) labelled this process scaffolding, which they defined as:

> "*[A] process that enables a child or novice to solve a problem, carry out a task or achieve a goal which would be beyond his unassisted efforts. This scaffolding consists essentially of the adult 'controlling' those elements of the task that are initially beyond the learner's capacity, thus permitting him to concentrate upon and complete only those elements that are within his range of competence.*"

Students are explorative and curious – something that should be not just allowed but encouraged. Exploring an unfamiliar environment can lead students to discover new things and let them experience the implications of their correct and incorrect answers; but leaving them on their own will generally not be enough (Mayer, 2004). Knowing that one is on the wrong track but without the knowledge of how to get where one wants to be going can lead to frustration and cognitive overload. Even though we should encourage students to explore and learn from their mistakes, there should be some scaffolding there to help the when needed (Mayer, 2004).

Feedback neglect from students

As said, feedback should be specific. Feedback should also be timely, delivered in reasonable time from when a task is completed so that learners correctly associate the feedback to the task and remember what they themselves provided as an answer (Schwartz et al., 2016). Feedback needs to be understandable (Schwartz et al., 2016; Lea & Street, 1998; Higgins, Hartley, & Skelton, 2001; Orsmond, Merry, & Reiling, 2005), non-threatening (Schwartz et al., 2016), and of reasonable amount so as not to be overwhelming (Brockbank & McGill, 1998). The student must be able to see how to use the feedback to improve (Orsmond et

al., 2005; Wiliam, 2007; Segedy, Kinnebrew, & Biswas, 2013). Even if one keeps all that in mind, however, the student still may not pay attention. Indeed, research shows that feedback neglect is common (Hounsell, 1987; Wotjas, 1998; Perrenoud, 1998; Clarebout & Elen, 2008; Conati, Jaques, & Muir, 2013).

Segedy, Kinnebrew, and Biswas (2012) discuss one of their earlier studies, where (by their calculation) 77% of the feedback in a digital learning game was ignored. They present a follow-up study where they refined the feedback to align better with students' goals in the game as well as provide more explanations and examples. A large proportion of the feedback remained ignored. The authors argue that, even though they designed the feedback to be as useful and understandable as possible, students were still unwilling to deal with it.

As noted earlier, the focus of Paper V is feedback neglect within the game Guardians of History. We looked for evidence of neglect at each step from noticing to processing, making sense of, acting upon, and finally progressing on the basis of feedback. We found neglect at each step, though the greatest number of students were lost at Step Three: acting upon the feedback. The results suggest that a digital tutee can be used to increase the likelihood that students at least notice and process feedback. Further research is needed to determine how better to avoid losing students later on, at Step Three and beyond.

Unwillingness to deal with feedback

Consider two average-performance students who receive feedback regarding an incorrect answer. They receive the same feedback at the same time but only one of them pays attention. What has happened?

Students – particularly those with low self-efficacy – have 'good' reasons to shy away from feedback they take to be negative. They may, as said, find it threatening and understand it as evaluative punishment or evidence that they are not smart enough. They may have a fixed mindset. Their response is rational, if not desirable.

Whether students are *mastery* or *performance oriented* (Dweck, 2000) affects how they perceive feedback. A student who is performance oriented seeks to prove her competence; one who is mastery oriented seeks to improve it. Performance-oriented students may take feedback badly, as evidence that they lack competence, instead of as information that can help them learn and improve.

Whether students are *task-involved* or *ego-involved* (Nicholls, Cobb, Wood, Yackel, & Patashnick, 1990) likewise plays a role. Task-involved students are less threatened by failure because their ego is less tied to success in the task. Ego-involved students can become anxious and discouraged in the face of failure. For

them, avoiding feedback means escaping a feeling of inadequacy (Hattie & Timperley, 2007).

Small things can help feedback be experienced as less threatening. Feedback should address the task or the effort put into it and avoid referencing the person or her intelligence: saying things like "*Well done! I see you have put a lot of effort into correcting your grammatical errors.*" as opposed to "*Well done! You are a smart girl!*" The latter kind of feedback can be perceived as meaningless, foster a fixed mindset by putting abilities and intelligence in focus (Hattie & Timperley, 2007), and do little for the student (Wilkinson, 1981: in Hattie & Timperley, 2007; Kluger & DeNisi, 1998); the task-related information that could assist self-efficacy or understanding is glaringly absent (Hattie & Timperley, 2007).

There is one final, prosaic reason for students disregarding feedback. Some students do not aspire to learn as much as possible but are content just to 'get by' to get through the period, the day, or the year without any major disaster – having made time for activities other than school work and studies (Perrenoud, 1991).

Feedback 'neglect' from system designers

As noted, most research on feedback to date concerns human-human interaction. Computers and software have the benefit that they can provide feedback in a versatile and highly customizable way. A textbook offers only the option of looking up the correct answer in the key at the back (if one is even provided) in the form of text, possibly with figures. With digital feedback, one can easily mix text, pictures (including interactive ones), movies, and sounds, repeating them as many times as a student wants or needs. In principle, a teacher could engage in such repetition as well, but the teacher usually does not have the time – nor, probably, the patience. That said, no one would wish for a school where each student spends all her time sitting alone in front of her computer. A competent teacher is still often the best option for helping a student. The teacher can sense where the student is in her learning process and adjust explanations accordingly. Technological solutions are needed because resources will never permit one teacher per student.

Being told that one is simply 'right' or 'wrong' – what Paper VI calls *verification feedback* – can be the best solution even on digital learning platforms, particularly when it comes to drill-and-practice. As noted earlier though, the problem with verification feedback is that it can encourage rote learning, which often means problems later on. In too many contemporary educational games, students who answer incorrectly are allowed to try again… and again… and again in a process of trial and error. There is no need to pay attention to what one is doing; sooner or later, the correct answer pops up. That these tests often have a time limit does not

necessarily help, because trying random answers may still be quicker than thinking through the problem. It might look as though the student knows what he is doing, but he does not. If a game truly lays claim to being educational, it must bring something more to the table.

Summary

The potential benefits of feedback in educational software remain underappreciated – partly because more knowledge is needed on what types of feedback work in which situations for which group of students, but also because educational software is being churned out at an extraordinary rate with little concern for proper evaluation. General design guidelines exist. Even when these are followed though, many students still ignore the feedback. The reasons for neglect can be many. One student may feel (subconsciously) that the feedback threatens her self-image; another does not understand it; yet another finds it too cumbersome to process. More knowledge is needed about what happens and why at each step along the feedback chain.

In closing, I would like to return to my 'Little Professor'. The reason I received it was to practice math since I was struggling with the subject. I always wanted to know *why* a certain rule or formula should be used, but the only answer I ever received was "*[...] because this is the way to do it.*". This did not help. I concocted my own, not very useful way of calculating: one that produced answers my teacher said were wrong. Looking back now, I can see how I would have benefitted from other kinds of feedback and better explanations.

Years later, math was still not my favourite subject. Then, in the middle of one semester at *secondary school*, I had the opportunity to change math teachers. This teacher was much more observant. He took time to give me the explanations I needed and had been missing. My grade rose from merely acceptable to the highest possible. I received first-hand knowledge that the scaffolding one receives is of huge importance. Educational software invites the possibility to take things one step further, customizing instruction to each student in a way that was not practical before.

References

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

Arroyo, I., B.P. Woolf, D.G. Cooper, W. Burleson, & K. Muldner (2011). The impact of animated pedagogical agents on girls' and boys' emotions, attitudes, behaviors and learning. In *Proceedings of the 11th International Conference on Advanced Learning Technologies, ICALT 2011* (pp. 506-510), July 6-8 2011, Athens, GA. Washington, D.C.: IEEE Computer Society.

Atkinson, R.K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, *94*(2), 416-427.

Azevedo, R., R. Landis, R. Feyzi-Behnagh, M. Duffy, G. Trevors, J. Harley, ... , & G. Hossain (2012). The effectiveness of pedagogical agents' prompting and feedback in facilitating co-adapted learning with MetaTutor. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K.-K. Panourgia (Eds.), *LNCS*, *vol. 7315: Proceedings of the 11th International Conference on Intelligent Tutoring Systems*, *ITS 2012* (pp. 212-221). Berlin, Germany: Springer.

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.

Bangert-Drowns, R. L., C.L.C. Kulik, J.A. Kulik, & M. Morgan (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213-238.

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of educational research, 61*(2), 213-238.

Bargh, J. A., & Y. Schul (1980). On the cognitive benefits of teaching. *Journal of Educational Psychology*, *72*(5), 593-604.

Bass, H., & D.L. Ball (2015). *Beyond "You can do it!" Developing mathematical perseverance in elementary school* (White paper). Chicago, IL: Spencer Foundation.

Baylor, A. L., & Y. Kim (2004). Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. In J. C. Lester, R.M. Vicari, & F. Paraguaçu (Eds.), *LNCS*, *vol. 3220: Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, *ITS 2004* (pp. 592-603). Berlin, Germany: Springer.

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

Baylor, A.L., R.B. Rosenberg-Kima, & E.A. Plant (2006). Interface agents as social models: The impact of appearance on females' attitude toward engineering. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems* (pp. 526-531). New York, NY: ACM.

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

Benware, C. A., & E.L. Deci (1984). Quality of learning with an active versus passive motivational set. *American Educational Research Journal*, *21*(4), 755-765.

Bickmore, T., & J. Cassell (1999). Small talk and conversational storytelling in embodied conversational interface agents. In *Proceedings of the AAAI Fall Symposium on Narrative Intelligence* (pp. 87-92), Cape Cod, MA, November 5-7.

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

Black, P., & D. Wiliam (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74.

Blair, K., D. Schwartz, G. Biswas, & K. Leelawong (2007). Pedagogical agents for learning by teaching: Teachable agents. *Educational Technology*, *47*(1), 56-61.

Brockbank, A. & I. McGill (1998). *Facilitating reflective practice in higher education*. Buckingham, UK: Society for Research into Higher Education and Open University Press.

Byrne, D. & D. Nelson (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.

Cassell, J., & T. Bickmore (2003). Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, *13*(1), 89-132.

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

Chen J, D. Shohamy, V. Ross, B. Reeves, & A.D. Wagner (2008) The impact of social belief on the neurophysiology of learning and memory. Abstract presented at the *Annual Meeting of the Society for Neuroscience*, Washington, DC, Nov 15-19, 2008. Washington, DC: Society for Neuroscience.

Chi, M.T., S.A. Siler, H. Jeong, T. Yamauchi, & R.G. Hausmann (2001). Learning from human tutoring. *Cognitive Science*, *25*(4), 471-533.

Chin, D.B., I.M. Dohmen, & D.L. Schwartz (2013). Young children can learn scientific reasoning with teachable agents. *IEEE Transactions on Learning Technologies*, *6*(3), 248-257.

Clarebout, G. & J. Elen (2008). Advice on tool use in open learning environments. *Journal of Educational Multimedia and Hypermedia*, *17*(1), 81-97.

Conati, C., N. Jaques, & M. Muir (2013). Understanding attention to adaptive hints in educational games: An eye-tracking study. *International Journal of Artificial Intelligence in Education*, *23*(1), 136-161.

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

Fiorella, C. L. (2013). *The cognitive benefits of learning by teaching and teaching expectancy* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 1545806)

Frasson, C. & E. Aimeur (1996). A comparison of three learning strategies in intelligent tutoring systems. *Journal of Educational Computing Research*, *14*(4), 371-383.

Graesser, A.C., N.K. Person, & J.P. Magliano (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, *9*(6), 495-522.

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

Griffin, S. (2004b). Number Worlds: A research-based mathematics program for young children. In D.H. Clements, J. Sararna, & A.-M. DiBiase (Eds.), *Engaging young children in mathematics: Standards for early childhood mathematics education* (pp. 325-342). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Gulz, A. & M. Haake (2010). Challenging gender stereotypes using virtual pedagogical characters. In S. Booth, S. Goodman, & G. Kirkup (Eds.), *Gender issues in learning and working with IT: Social constructs and cultural contexts* (pp. 113-132). Hershey, PA: Information Science Reference & IGI Global.

Gulz, A., M. Haake, & A. Silvervarg (2011). Extending a teachable agent with a social conversation module: Effects on student experiences and learning. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th International Conference Artificial Intelligence in Education*, *AIED 2011*, *LNAI*, *vol. 6738* (pp. 106-114). Berlin, Germany: Springer.

Hannula, M.M., A. Mattinen, & E. Lehtinen (2005). Does social interaction influence 3-year-old children's tendency to focus on numerosity? A quasi-experimental study inday care. In L. Verschaffel, E. De Corte, G. Kanselaar, & M. Valcke (Eds.), *Studia Paedagogica 41: Powerful environments for promoting deep conceptual and strategic learning* (pp. 63-80), Leuven, Belgium: Leuven University Press.

Harlen, W., & R. Deakin Crick (2003). Testing and motivation for learning. *Assessment in Education: Principles, Policy & Practice*, *10*(2), 169-207.

Hattie, J. & H. Timperley (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Heidig, S. & G. Clarebout (2011). Do pedagogical agents make a difference to student motivation and learning? *Educational Research Review*, *6*(1), 27-54.

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

Higgins, R., P. Hartley, & A. Skelton (2001). Getting the message across: The problem of communicating assessment feedback. *Teaching in Higher Education*, *6*(2), 269-274.

Hounsell, D. (1987). Essay writing and the quality of feedback. In J.T.E. Richardson, M.W. Eysenck, & D. Warren-Piper (Eds.) *Student learning: Research in education and cognitive psychology* (pp. 109-119). Guildford, UK: SRHE & Open University Press.

Isbister, K. & C. Nass (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., G. Ozogul, & M. Reisslein (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.

Johnson, W.L., J.W. Rickel, & J.C. Lester (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, *11*(1), 47-78.

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

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

Kim, Y., A.L. Baylor, & PALS Group (2006). Pedagogical agents as learning companions: The role of agent competency and type of interaction. *Educational Technology Research and Development*, *54*(3), 223-243.

Kim, Y., A.L. Baylor, & G. Reed (2003). The impact of image and voice with pedagogical agents. In A. Rossett (Ed.), *Proceedings of E-Learn 2003: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (pp. 2237-2240), Phoenix, AZ, November 7-11 2003. Phoenix, AZ: AACE.

Kim, Y., Q. Wei, B. Xu, Y. Ko, & V. Ilieva (2007). MathGirls: Toward developing girls' positive attitude and self-efficacy through pedagogical agents. In R. Luckin, K.R. Koedinger, & J. Greer (Eds.), *Frontiers in Artificial Intelligence and Applications*, *vol. 158: Proceedings of AIED 2007* (pp. 119-126). Amsterdam, The Netherlands: IOS Press.

Kirkegaard, C. (2016). *Adding challenge to a teachable agent in a virtual learning environment*. PhD thesis, Dept. of Computer Science, Linköping University. Linköping, Sweden: Linköping University Electronic Press.

Kluger, A. N. & A. DeNisi (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254-284.

Kluger, A. N. & DeNisi, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword. *Current Directions in Psychological Science*, *7*(3), 67-72.

Lave, J. & E. Wenger (2001). *Situated learning: Legitimate peripheral participation*. New York, NY: Cambridge University Press.

Lea, M. R. & B.V. Street (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education*, *23*(2), 157-172.

Lee, J.E., C. Nass, S.B. Brave, Y. Morishima, H. Nakajima, & R. Yamada (2006). The case for caring colearners: The effects of a computer-mediated colearner agent on trust and learning. *Journal of Communication*, *57*(2), 183-204.

Leelawong, K. & G. Biswas (2008). Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education*, *18*(3), 181-208.

Lester, J. C., S.A. Converse, S.E. Kahler, S.T. Barlow, B.A. Stone, & R.S. Bhogal (1997). The persona effect: affective impact of animated pedagogical agents. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (pp. 359-366). New York, NY: ACM.

Mayer, R.E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, *59*(1), 14.

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

McKendree, J. (1990). Effective feedback content for tutoring complex skills. *Human-computer Interaction*, *5*(4), 381-413.

McLaren, B.M., K.E. DeLeeuw, & R.E. Mayer (2011a). A politeness effect in learning with web-based intelligent tutors. *International Journal of Human-Computer Studies*, *69*(1), 70-79.

McLaren, B.M., K.E. DeLeeuw, & R.E. Mayer (2011b). Polite web-based intelligent tutors: Can they improve learning in classrooms? *Computers & Education*, *56*(3), 574-584.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science*, *32*(1), 99-113.

Moreno, R. & T. Flowerday (2006). Students' choice of animated pedagogical agents in science learning: A test of the similarity-attraction hypothesis on gender and ethnicity. *Contemporary Educational Psychology*, *31*(2), 186-207.

Moreno, R., R.E. Mayer, H.A. Spires, & J.C. Lester (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, *19*(2), 177-213.

Nass, C. & K.M. Lee (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.

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

Nicholls, J.G., P. Cobb, T. Wood, E. Yackel, & M. Patashnick (1990). Assessing students' theories of success in mathematics: Individual and classroom differences. *Journal for Research in Mathematics Education*, *21*(2), 109-122.

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

Orsmond, P., S. Merry, & K. Reiling (2005). Biology students' utilization of tutors' formative feedback: A qualitative interview study. *Assessment & Evaluation in Higher Education*, *30*(4), 369-386.

Ozogul, G., A.M. Johnson, R.K. Atkinson, & M. Reisslein (2013). Investigating the impact of pedagogical agent gender matching and learner choice on learning outcomes and perceptions. *Computers & Education*, *67*, 36-50.

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading and Writing Quarterly*, *19*(2), 139-158.

Palinscar, A.S. & A.L. Brown (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, *1*(2), 117-175.

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.

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

Pareto, L., D.L. Schwartz, & L. Svensson (2009) Learning by guiding a teachable agent to play an educational game. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A.C. Graesser (Eds.), *Frontiers in Artificial Intelligence and Applications*, *vol. 200: Proceedings of AIED 2009* (pp. 662-664). Amsterdam, The Netherlands: IOS Press.

Perrenoud, P. (1991). Towards a pragmatic approach to formative evaluation. In P. Weston (Ed.) *Assessment of pupil achievement* (pp. 79-101). Amsterdam, The Netherlands: Swets & Zeitlinger.

Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning processes: Towards a wider conceptual field. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 85-102.

Plant, A.L. & E.A. Baylor (2005). Pedagogical agents as social models for engineering: The influence of agent appearance on female choice. In C.-K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Frontiers in Artificial Intelligence and Applications*, *vol. 125: Proceedings of AIED 2005* (pp. 65-72). Amsterdam, The Netherlands: IOS Press.

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

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

Rattan, A., C. Good, & C.S. Dweck (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.

Renkl, A. (1995). Learning for later teaching: An exploration of mediational links between teaching expectancy and learning results. *Learning and Instruction*, *5*(1), 21-36.

Riggio, R.E., J.W. Fantuzzo, S. Connelly, & L.A. Dimeff (1991). Reciprocal peer tutoring: A classroom strategy for promoting academic and social integration in undergraduate students. *Journal of Social Behavior and Personality*, *6*(2), 387.

Roscoe, R.D. & M.T. Chi (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research*, *77*(4), 534-574.

Rosenberg-Kima, R.B., E.A. Plant, C.E. Doerr, & A.L. Baylor (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.

Schunk, D.H. (1995). Self-efficacy and education and instruction. In J. E. Maddux (Ed.), *Self-efficacy, adaptation, and adjustment: Theory, research, and application* (pp. 281-303). New York, NY: Plenum Press.

Schunk, D.H. & J.L. Meece (2006). Self-efficacy development in adolescence. In F. Pajares, & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 71-96). Greenwich, CT: Information Age Publishing.

Schwartz, D.L., C. Chase, D.B. Chin, M. Oppezzo, H. Kwong, S. Okita, … , J. Wagster (2009). Interactive metacognition: Monitoring and regulating a teachable agent. In D.J. Hacker, J. Dunlosky, & A.C. Graesser (Eds.), *The educational psychology series. Handbook of metacognition in education* (pp. 340-358). New York, NY: Routledge (Taylor & Francis Group).

Schwartz, D.L., J.M. Tsang, & K.P. Blair (2016). *The ABCs of how we learn: 26 scientifically proven approaches, how they work, and when to use them*. New York, NY: W.W. Norton & Company.

Segedy, J.R., J.S. Kinnebrew, & G. Biswas (2012). Supporting student learning using conversational agents in a teachable agent environment. In *10th International Conference of the Learning Sciences: The Future of Learning, ICLS 2012 - Proceedings* (Vol. 2, pp. 251-255), July 2-6 2012, Sydney, Australia.

Segedy, J.R., J.S. Kinnebrew, & G. Biswas (2013). The effect of contextualized conversational feedback in a complex open-ended learning environment. *Educational Technology Research and Development*, *61*(1), 71-89.

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

Shute, V.J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153-189.

Silvervarg, A., M. Haake, & A. Gulz (2013). Educational potentials in visually androgynous pedagogical agents. In H.C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *LNCS*, *vol. 7926: Proc. of AIED 2013* (pp. 599-602). Berlin, Germany: Springer-Verlag.

Silvervarg, A., M. Haake, L. Pareto, B. Tärning, & A. Gulz (2011). Pedagogical agents: Pedagogical interventions via integration of task-oriented and socially oriented conversation. Part of the *AERA 2011 Annual Meeting Symposium: Pedagogical Agent Presence, Appearance, and Agent-learner Interaction – Current Research and Future Directions*, April 8-12 2011, New Orleans, LA.

Sjödén, B., B. Tärning, L. Pareto, & A. Gulz (2011). Transferring teaching to testing – an unexplored aspect of teachable agents. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *LNCS*, *vol. 6738: Proc. of AIED 2011* (pp. 337-344). Berlin, Germany: Springer.

Stajkovic, A.D. & F. Luthans (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin*, *124*(2), 240-261.

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.), *LNCS*, *vol. 1839: Proceedings of the 5th International Conference on Intelligent Tutoring Systems, ITS 2000* (pp. 103-112). Berlin: Springer.

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.

Veletsianos, G. (2010). Contextually relevant pedagogical agents: Visual appearance, stereotypes, and first impressions and their impact on learning. *Computers & Education*, *55*(2), 576-585.

Veletsianos, G. (2012). How do learners respond to pedagogical agents that deliver social-oriented non-task messages? Impact on student learning, perceptions, and experiences. *Computers in Human Behavior*, *28*(1), 275-283.

Vygotsky, L.S. (1980). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

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

Webb, N.M. (1989). Peer interaction and learning in small groups. *International Journal of Educational Research*, *13*(1), 21-39.

Wiliam, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F.K. Lester, Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1053-1098). Charlotte, NC: Information Age Publishing.

Wood, D., J. S. Bruner, & G. Ross (1976). The role of tutoring in problem solving. *Journal of child psychology and psychiatry, 17*(2), 89-100.

Woolf, B.P., I. Arroyo, K. Muldner, W. Burleson, D.G. Cooper, R. Dolan, & R.M. Christopherson (2010). The effect of motivational learning companions on low achieving students and students with disabilities. In V. Aleven, J. Kay, & J. Mostow (Eds.), *LNCS*, *vol. 6094: Proceedings of the 10th International Conference on Intelligent Tutoring Systems*, *ITS 2010* (pp. 327-337). Berlin: Springer.

Wotjas, O. (1998). Feedback? No, just give us the answers. *Times Higher Education Supplement*, September 25, 1998. Retrieved from https://www.timeshighereducation.com/news/feedback-no-just-give-us-the-answers/109162.article

Md. Yunus, A.S & W.Z. Wan Ali (2008). Metacognition and motivation in problem solving. *International Journal of Learning*, *15*(3), 121-132.

Appendix

Appendix

In paper I, II, III and IV an educational game called “The squares family” has been used. Here follows a short description of the game idea and underlying pedagogy.

The squares family

The overall goal is that students should gain conceptual understanding in mathematics. For example, understand that the “3” in 30 means three sets of ten, and more generally that the position of a number plays a role for what it means. An understanding of the so-called ten-base system is critical for future learning in mathematics.

Importantly, the mathematical content of the game is interwoven with the game narrative. Another important feature of the game is that the student learns by teaching a digital tutee.

The squares family is a two player game where a player can be a student, a digital tutee or ‘the computer’. ‘The computer’ has five different skill levels; at level 1 the computer plays by choosing cards randomly and at level 5 it plays like an expert.

The game is used either in a collaborative mode or in a competitive mode.

There is a playing board in the middle and two sets of cards on each side, one for each player (see figure 1). The game is composed of different subgames to choose from, such as “the find-pair” and “within the rope up to 10, 100 or 1000”. Since the students in the studies reported in this thesis only used the latter this is what will be explained here. Besides from choosing the range – up to 10, 100 or 1000 – the student gets to choose between subtraction or addition.

The players 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.

In figure 1 Oskar plays the sub-game “within the rope: up to 100”. The set of cards to the right is his opponents cards (the computer playing at skill level 3) and set of cards to the left is Oskar’s. The game board has three framed squares; one dark blue; one light blue and one green. The dark blue squares represent ones (the two red boxes, within the rope on the dark blue, represents the number “2”), the light

blue squares represents tens (the three orange boxes, within the rope on the light blue, represents “30”) and the green squats represents hundreds (the one yellow box, within the rope on the green, represents “100”) and altogether the number 132 is represented on the board.

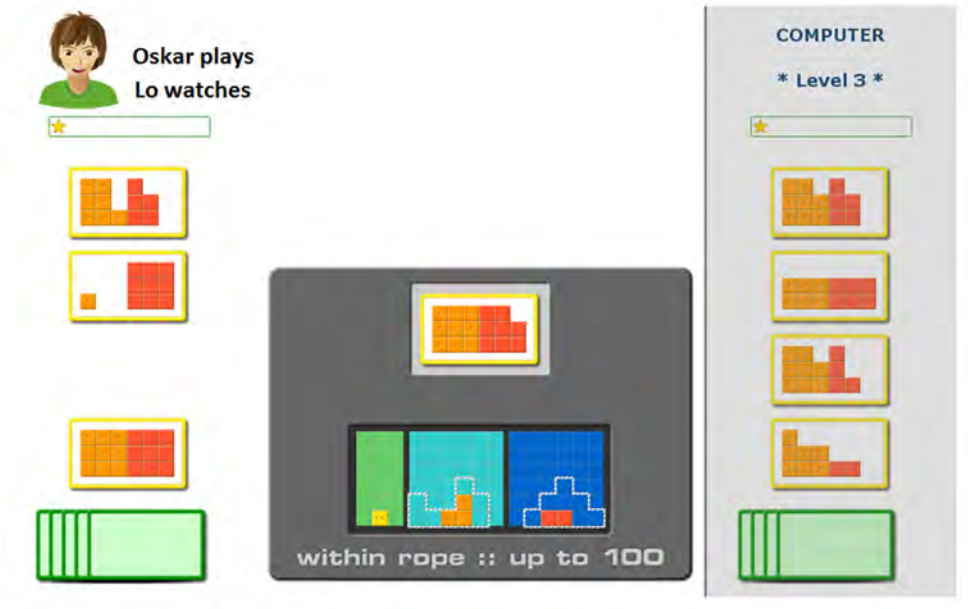


Figure 1. The game showing the board game in the middle and Oskar’s cards to the left and the computer (playing at level 3) cards to the right.

The cards which the student can chose from present tens, as orange boxes, and ones, as red boxes. In the example the student has chosen the card “98” which leads to two carry-overs and renders two stars to the student. When the first 8 red boxes are added to the dark blue area they will fill up this area within the rope (equaling ten boxes). The ten red boxes will be packed into one orange box (making the orange boxes four). Adding the nine orange boxes from the card leads to a carry-over (of ten orange boxes) into the green area; three orange squares are then left in the light blue area.

The challenge is to choose a good (preferable the best) card – that is the card that leads to most carry-overs or, if no carry over is possible, a card that stops the opponent from getting a star (i.e. a carry-over). In other words, which cards are better depends on the current game board and both players’ cards.

The student teaches her digital tutee in two different modes; “Observe” and “Try-and-be-guided”. In the observe-mode the student choses which card she thinks is best and the digital tutee observes and learns from this. In the try-and-be-guided

mode, the digital tutee suggests which card to pick, but the student always has the option of stopping the tutee and exchange the card to one that the student prefers. In the final mode "On her own" the digital tutee plays on her own based on what she has learned by observing the student play and from the student's answers to the questions she has posed. In this mode, the student cannot interfere.

The questions posed by the digital tutee are of multiple-choice format. All questions concern game play. They always relate to the current situation and most often to the choice(s) of card(s) just made. The timing of the question is so that it comes after a card is chosen but before the computation on the board is performed. This way the student has one more chance to reflect over her choice.

Figure 2. The digital tutee asks the question: "*I thought of the card 99, why´s 98 better?*" The answering options are: "*The card 98 gives 2 points, but 99 gives nothing*", "*Both are just as good and give 2 points*", "*It turns out I was wrong, the card 99 gives more points*", and "*I don't know*".

The questions are organized into five categories, reflecting the stages of becoming a skilful player. The categories are (p. 262, Pareto, 2014).

1. *Game idea:* understanding what the game is about (see figure 3 left panel).
2. *Graphical model:* understanding the graphical model and how it relates to mathematics (see figure 3 right panel).

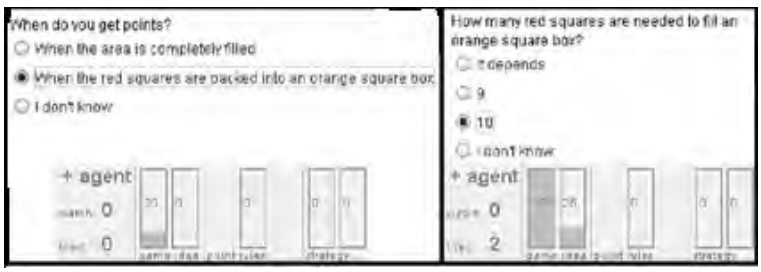


Figure 3. Left panel the agent asks a question regarding the game idea and in the right panel the agent asks a question regarding the model category.

3. *Scoring:* knowing how points are awarded in the game and predicting the outcome of a choice
4. *Basic strategic thinking*: how to choose the best card considering their own cards only which involves predicting each of the four cards' effects and choosing (see figure 4 A-D).
5. *Advanced strategic thinking*: how to choose the best card considering both players' cards and predicting two steps ahead. Normally this means considering and discriminating between 16 alternatives that are two arithmetic computations ahead (see figure 4 E).

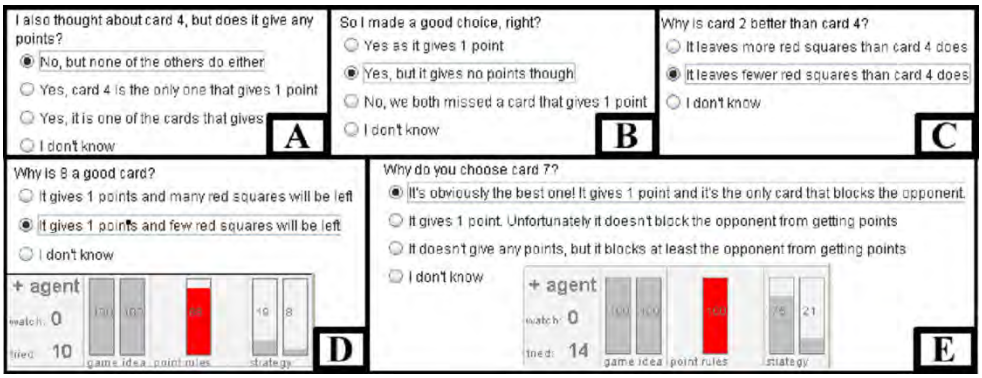


Figure 4. Different examples of strategic agent questions: in show-mode (A), try-mode (B), either model (C), advanced level (D), most challenging type of question E).

The digital tutee asks questions depending on how much it has been taught by its teacher (i.e. the real student). In the beginning when not knowing that much, it asks more basic questions like "when do you get a star" and progresses to more advanced questions as the learning progresses.

For further information regarding the game and how the digital tutee is programmed for example I refer to Pareto (2014).

Paper I

T. Hirashima et al. (Eds.) (2011). Proceedings of the 19th International Conference on Computers in Education. Chiang Mai, Thailand: Asia-Pacific Society for Computers in Education

Off-task Engagement in a Teachable Agent based Math Game

Betty TÄRNING[a*], Magnus HAAKE[b] & Agneta GULZ[a]
[a]*Lund University Cognitive Science,* [b]*Department of Design Sciences, Lund University*
*betty.tarning@lucs.lu.se

Abstract: A previous study compared two student groups that played a mathematics game based on a teachable agent. One group played with, and the other without, the inclusion of a social conversation module: a chat between the student and the teachable agent. Results were that students who used the game with the chat included had a more positive experience of the game and learned more in the sense of teaching their agent better. However, patterns differed between sub-groups of students. Low-achievers did not prefer the game with the chat included, whereas high- and mid-achievers did, but simultaneously low-achievers tended to chat more. Low-achievers tended not to use the options of not starting the chat or quitting a chat beforehand as much as high- and mid-achievers did. In this paper we pursue a more detailed analysis of the students' conversational behavior in the chat. The analytic focus is on the notion of engagement. Results point towards differences between the student groups in their engagement in the off-task conversation, that in turn can help explain the previous somewhat paradoxical result.

Keywords: teachable agent, off-task conversation, natural language dialogue, engagement, low- and high-achievers, quality of conversation, learning

Introduction

The paper approaches engagement in off-task conversation between students and a teachable agent. The starting point was an intriguing result from a previous study regarding how math low- and high-achievers respectively responded towards off-task conversation. We wanted to reach a better understanding of this result by undertaking an additional analysis. In this the notion of *engagement* became the analytic focus.

1. Background: the original result of different patterns for low- and high achievers

The underlying mathematics game trains basic arithmetic skills [1,2], and the student teaches her *Teachable Agent*, TA, to play the game. For more details we refer to [2]. The focus in this paper is on the chat-like conversation between student and TA by means of natural language text input. We refer to this as *off-task conversation* as opposed to the *on-task conversation* between student and TA, which is a multiple-choice guided conversation during the math game sessions, targeted at mathematical content. The motive behind the chat is to enhance students' experiences and increase their inclination to use the game over time. Yet another motive is to enable additional pedagogical interventions such as influencing students' math self-efficacy, cf. [3]. Three classes in a Swedish school participated in the original study, with 38 female and 42 male 12-14 year olds. 18 were by their teachers classified as math low-achievers, 39 as math mid-achievers and 23 as math high-achievers. Each class was divided into two groups with an even distribution with respect to

math achievement and gender in each group. All students got to play the game during three lessons. The *WithC*-group used the game with, and the *NoC*-group without the chat module. For switching between playing the game and chatting, the TA uses the metaphor of "breaks" between lessons in school. During the first three "breaks" the chatting with the agent was mandatory until the "break" ended after three minutes. After this students were always offered the *choice* between taking a "break" to chat with the agent or continuing to play the game. They could also choose to end the chat before the "break" was over. However students in both conditions did get breaks in the sense of cognitive rest and change of activity. Instead of training their TA, they could let it play against the computer which only requires passive viewing. It was obvious that students did make use of the latter as a kind of break.

After having completed the third lesson, students filled out a questionnaire regarding their experience of the game. In sum the results showed (for details see [1]) that students in the *WithC*-group tended to have a more positive game experience than students in the *NoC*-group (diff=0.54, p=0.07). Furthermore, students in the *WithC*-group tended to reach better result in terms of how well they taught their TA (diff=0,3, p=0.07). But separating the results into student sub-groups, revealed that it was only the medium- and high-achieving students that reported a more positive experience of the game for the *WithC* compared to the *NoC* condition. The *low-achievers* did not seem to prefer any of the two system versions before the other. A comparison of the tendency to refrain from starting a chat or quitting a chat before chat time was out also revealed different patterns in the subgroups. High-achievers, and to some extent mid-achievers refrained from starting a chat or quitting a chat beforehand – to a considerably higher extent than low-achievers.

a) ***Student:*** hola senorita
TA: What do you do on the weekends?
Student: Play handball and just chilling with my friends
TA: OK
Student: and you?
TA: I quite often listen to music
Student: OK, what do you like listening to?
TA: Mostly rock and pop
Student: What's so great about rock do you think?
TA: I did not understand, what do you mean?
Student: Why do you like rock?

b) ***Student:*** are you good at school?
TA: pretty good
Student: neeerd
TA: can't we talk about spare time interests instead
Student: ok, I like football a lot
TA: how, what do you mean?
Student: It's fun to go for the footballs and hit them
TA: Well, it's difficult for me to understand what you say, can we talk of something else?
Student: Do you have problems with the Swedish language or what?
TA: I'll go to the cinema this weekend, it's so nice just forget about everything and go into another world kind of
Student: Can I come along?
TA: What do you do usually on the weekends?

c) ***TA:*** maybe we could talk about music
Student: .
TA: I did not understand, what do you mean by that?
Student: .
TA: It's difficult for me to understand what you say, can we talk of something else?
Student: .
TA: I think Coldplay is really cool, which is your favorite artist?
Student: .
TA: well
Student: .
TA: okay
Student: .

d) ***Student:*** hi
TA: hi
Student: do you like animals?
TA: I don't understand, ask me something else
Student: dfjhasjk
TA: Maybe we can talk about hobbies?
Student: fghjkl
TA: What do you mean?
Student: hfhdfdjdjdj
TA: Yesterday I scored twice in floor ball in gym class, do you like sports?
Student: djdjdjd
TA: OK
Student: djdjd

Figure 1: Examples of chats with high (a & b) and low (c & d) engagement. (Translated from Swedish.)

Taken together there is a seeming paradox in that high-achievers and mid-achievers tended to like the game better when the chat was included (compared to low-achievers) while they also were more inclined to quit and to refrain from starting the chat when they had the option to. In other words: high- and mid-achievers *liked* the software more when there was a chat included, but they *chatted less* than the low-achievers. Low-achievers on the other hand

chatted more but did *not like* the software to any higher degree with the chat included. We therefore decided to conduct a more detailed analysis of the participating students' chat behavior. At first, two researchers did an independent screening of the chat logs in order to familiarize them selves with the material. They made the common observation that the *engagement* in the conversation, as indicated by the chat logs, varied considerably between chats of each individual student as well as between the individual students. Figure 1a & 1b show two examples of parts of chat logs that, according to the researchers coding scheme, indicate high engagement, and Figure 1c & 1d present two examples of parts of chat logs that indicate low engagement.

Engagement is increasingly discussed within the educational technology field, e.g. [4]. The research questions that we hoped to illuminate by the analyses to be presented in this paper were the following: 1) To what degree did the students seem *engaged* in the off-task conversations with their teachable agent? Did low- and high-achievers differ in this respect? 2) What did students do when the chat logs indicate that they are or have become very disengaged in the chat conversation? In particular: did they quit or not? Did they refrain from starting the next chat or not? Did they continue chatting in a similar way indicating low engagement, or not? Did low- and high-achievers differ in this respect?

2. Method

6 = the student is driving the conversation and asks relevant questions, sometimes in the form of a counter question when having answered a question from the TA *("do you have siblings?"; "I feel good right now, how about you?"*

5 = the student is driving the conversation but asks more curious and exploratory questions *("do you have problems to understand the Swedish language?"; "are you gay?")*

4 = the TA urges the student to ask a question and the student gives a relevant answer *(TA: "what type of music do you like?" Student: "techno"; TA: "Ask me something else" Student: "do you like horses?")*

3 = the TA urges the student to ask question and the student gives an answer that is (semi)-relevant but questionable in terms of being engaged in driving the conversation further: *(TA "I don't get it, can we talk about something else?" Student: "You are stupid"); (TA: "it seemed fun to me" student: "ha ha, you nerd").*

2 = the student drives the conversation but asks irrelevant questions *("are you fat?")*

1 = the TA either asks a question or urges the student to ask a question but utterances provided by the student are irrelevant/not furthering the conversation *(TA:"What do you mean by that?" Student: "go and drown yourself"; TA:" I don't get it can we talk about something else?" Student: "asshole")*

0 = nonsense syllables and blank space (*Student: ". "; Student: "xasdfghhhh")*

Figure 2: The coding scheme for the chats with some example utterances. (Translated from Swedish.)

Measuring off-task conversation engagement: Chat logs were collected in the *WithC*-group for 30 students: 11 high-achievers, 13 mid-achievers and 6 low-achievers. Each student was involved in 3 to 8 chats and on the average each student exchanged 130 phrases with their TA. Half of them were uttered by the agent and half by the student. Each phrase produced by a student was categorized in terms of the degree of engagement in the conversation and was given a value between 0 and 6 (0 representing extremely low engagement and 6 very strong engagement, see Figure 2). The context of the conversation so far and the utterance just made by the TA was taken into consideration, with the main objective to estimate to what degree the student's utterance was an engaging conversation initiative, or a suitable and engaged response, given the TA:s previous utterance. Two researchers coded the dialogue (inter-rater reliability measure (Cohen's kappa): $\kappa = 0.86$) and means were calculated.

3. Results and analysis

Engagement in off-task conversation: We chose to calculate two engagement values, one for the first three mandatory chats (*chat 1-3*), and one for the following chats (*chats 4-x*), where the chatting was optional and could be quitted beforehand. What then comes forth is the following (Table 1). For chat 1-3 there is no significant difference in engagement between high- and low-achievers (*t*-test [one-tailed]: $p = 0{,}171$, but for chat 4 and onwards the engagement value is significantly higher for high achievers than for low achievers (*t*-test [one-tailed]: $p = 0{,}044$; all participants with 4 or more chat sessions included).

Table 1: Off-task engagement for mandatory chats (chat 1-3); Off-task engagement for optional chats (chat 4-x); all participants with 4 or more chat sessions included.

	off-task (chat 1-3)			*off-task (chat 4-x)*			
Groups	*n*	Mean	Var.	*n*	Mean	Var.	**one-tailed *t*-tests**
low	6	3,32	2,3141	4	2,29	1,2542	chat 1-3: $p_{high,low} = 0{,}171$
mid	9	4,04	0,6380	11	3,97	1,3725	chat 4-x: $p_{high,low} = 0{,}044$
high	8	3,93	0,5908	8	3,82	1,9366	

What did students do when disengaged in the chat conversation? This analysis started by an identification of those chat passages where a student clearly seemed to have lost engagement in the conversation with the agent. One criterion is when a student repeats a blank, a dot, a single word, or meaningless strings, and continues to do so without getting back to a productive conversation. Another criterion is when a student goes on with something that seems more like a monologue, sometimes including harassment, which does not relate to any of the utterances by the TA and, again, does not get back to a productive conversation. All chat logs for all participants were evaluated. Fourty-three instances of disengagement according to the criteria above were identified (see Table 2). Out of these, 22 were within high-achiever chats, 11 within low-achiever chats, and 10 within mid-achiever chats. Since there were 11 high-achievers and 6 low-achievers in the *WithC*-group, it was equally common that a low-achiever and a high-achiever did get strongly disengaged. What differed between the two student groups, however, was the behavioral pattern in this kind of situation – even though we cannot claim statistical significance given the limited number of students involved in the analysis.

Table 2: Actions taken in situations where the chat log indicates strong disengagement in the conversation on the part of the student. (All participants and chat logs included.)

	low-achievers ($n = 6$)	**mid-achievers ($n = 13$)**	**high-achievers ($n = 11$)**	*Σ (sum)*
Quits the chat	0	3	**5**	*8*
Does not start next chat	3	2	**13**	*18*
Gets on with unengaged chat	**8**	5	4	*17*
Σ (sum)	*11*	*10*	*22*	*43*

For the 22 instances of low engagement in conversation between high-achievers and the TA, the student quitted the chat 5 times, and refrained from starting the next chat 13 times. Only in four of the instances did the student both continue the disengaged chat and also start the next one. Reversely for the 11 instances with low-achievers, the student continued the disengaged chat and also started the next chat 8 times, with only 3 instances of refraining from starting the next chat and no case of a student quitting a chat.

In other words, low-achievers were more inclined to continue a chat even when there is a strong indication that they were unengaged, whereas high-achievers were more inclined in these situations to take control or action: they quit the chat and/or refrained from starting

next chat. This result is in concordance with the results reported in table 3: the engagement value for chat 4 and onwards, where the student could control the starting and ending of a chat, was significantly higher for high-achievers than for low-achievers.

Two aspects of the differing behavioral patterns may contribute to an explanation of why high-achievers tended to like the game better with the chat included whereas low-achievers did not. First, it is well known that the experience of having control over one's situation plays an important role for a positive learning experience. Second, with the behavioral differences described, low-achievers tended to spend more time than high-achievers with a chat they might have experienced as non-engaging (boring, meaningless).

4. Discussion and conclusion

The analysis presented in this paper did provide us with a possible explanation of the previous somewhat intriguing result that high–achievers but not low-achievers *liked* the software more when there was a chat included, but that they *chatted less* than the low-achievers. The suggested explanation is the different actions taken when getting into a clearly disengaged conversation with the TA. Low-achievers in the study did not tend to take control over a situation of disengagement in the sense of quitting the chat and/or refraining from chatting next time, whereas the high-achievers tended to do so. This can be important, in more general terms, to consider for designers of pedagogical games: how are possibilities of taking control in the game presented and to what extent will different students use these possibilities? For our case this is important since we are aiming for pedagogical interventions regarding math self-efficacy beliefs via the chat. In turn this is most important for low-achievers, and therefore we need a chat that works well for them.

We plan to make more information from the math game sessions accessible to the chat module, so that more detailed conversations about the game play can be conducted in the chat. This plan gets support from the chat-logs from the present study. More than a third of the students spontaneously initiated chat conversations about the math game with their TA, wanting to discuss whether the TA found it difficult, whether the TA thought it had learnt much, etc. Notably this applies to both high- mid- and low-achievers and thus seems a promising venue for pedagogical interventions.

Acknowledgments: We thank the Swedish Knowledge foundation for support.

References

[1] Gulz, A., Haake., M., & Silvervarg, A. (2011). Extending a Teachable Agent with a Social Conversation Module – Effects on Student Experiences and Learning. In G. Biswas et al. (Eds.): AIED 2011, LNAI 6738, pp. 106-114, 2011. Springer.

[2] Pareto, L., Arvemo, T., Dahl, Y., Haake., M., & Gulz, A. (2011). A Teachable-Agent Arithmetic Game's effects. In G. Biswas et al. (Eds.): AIED 2011, LNAI 6738, pp. 247-255, 2011. Springer.

[3] Kim, Y., Wei, Q., Xu, B., Ko, Y., & Ilieva, V. (2007). MathGirls: Increasing girls' positive attitudes and self-efficacy through pedagogical agents. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Proc. of AIED 2007* (pp. 119-126). Amsterdam, The Netherlands: IOS Press.

[4] Quinn, C. (2005). *Engaging Learning – Designing e-Learning Simulation Games*. San Francisco, CA: John Wiley & Sons, Inc.

Paper II

Instructing a teachable agent with low or high self-efficacy – does similarity attract?

Since the publication of this thesis, a revised version of this article is published as:
Tärning, B., Silvervarg, A., Gulz, A., & Haake, M. (2019). Instructing a teachable agent with low or high self-efficacy – does similarity attract? *International Journal of Artificial Intelligence in Education*, *29*(1), 89–121.
https://doi.org/10.1007/s40593-018-0167-2

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

[1] *Div. of Cognitive Science, Lund University, Lund, Sweden*
[2] *Dept. of Computer and Information Science, Linköping University, Linköping, Sweden*

Abstract. Previous studies have shown that such characteristics of pedagogical agents as communicative style, gender, ethnicity and level of competence, can affect students' performance, self-efficacy and attitude towards the agent. This study targets teachable agents, or digital tutees, and the potential effects of a characteristic that to our knowledge has not been studied in this kind of agent before, namely the agents' self-efficacy. A total of 166 students, aged 10-11, used a teachable agent based math game focusing on the base-ten number system. Through data logging and questionnaires, the study compared the effects of high vs. low agent self-efficacy on students' attitude towards the agent, their own math self-efficacy and their performance in the math game. The study further explored the effects of matching vs. mismatching 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 digital tutee with low self-efficacy. They also increased their self-efficacy when instructing a digital tutee 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 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, digital tutee, similarity attraction hypothesis, mathematics.

1. 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 learning, including performance (Lee, Nass, Brave, Morishima, Nakajima, & Yamada, 2007; Veletsianos, 2009; Johnson, Ozugul, & Reisslein, 2015) performance growth or learning (Frasson & Aimeur, 1996; Arroyo, Woolf, Royer, & Tai, 2009), motivation (Baylor, Ryu, & Shen, 2003; Ebbers, 2007; Plant, Baylor, Doerr, & Rosenberg-Kima, 2009) and self-efficacy, i.e. the belief in one's capacity to succeed with a task or in a domain (Baylor & Kim, 2005; Kim & Baylor, 2006; Ebbers, 2007).

Furthermore, research shows that these effects differ between groups of students. A certain design choice often has different impact on different groups of students (Hietala & Niemirepo, 1998; Kim & Wei, 2011; Ozogul, Johnson, Atkinson, & Reisslein, 2013; Arroyo, Woolf, Cooper, Burleson, & Muldner, 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 student, 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, Richardson, & Yard, 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 self-efficacy (Bandura, 1997). It is therefore of interest to explore possibilities to influence performance and self-efficacy in the area of mathematics, and two of the outcomes we focus on in the present study are precisely these: students' performance in math, specifically place value, and a possible change in the students' self-efficacy 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, Plant, Doerr, & Baylor, 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, Leelawong, Schwartz, Vye, & The Teachable Agent Group at Vanderbilt, 2005) in relation to whom the real student takes a teacher role. Teachable agent based software implement 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 & 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.

An important 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, Raukola, Haake, & Gulz, 2012; Silvervarg, Haake, & Gulz, 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 & 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, some things are 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 short, its self-efficacy (SE), in a task or domain. 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 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 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 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' 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 & Nelson, 1965; Byrne, Griffitt, & Stefaniak, 1967).

2. Background and related work

2.1. The phenomenon and concept of self-efficacy

Self-efficacy 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), self-efficacy (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, Carr, & Wiggers, 2014). Second, it is important to distinguish SE from more general self-attitudes, such as self-confidence or self-esteem. SE's subject-specific nature 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 writes, this "assuredly spawns failure" (Bandura, 1997, 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, Good, & Dweck, 2012).

For all that SE and performance in an area are connected – experiences of success in an area is one of the primary factors (see 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 self-efficacy in the area. As Bandura puts it: "The objective of education is not the production of self-confident fools" (Bandura, 1997, 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 self-efficacy 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.

2.2. 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, Wagster, & Biswas, 2008; Chase, Chin, Oppezzo, & Schwartz, 2009; Sjödén, Tärning, Pareto, & Gulz, 2011, Okita & Schwartz, 2013). One study (Pareto, Arvemo, Dahl, Haake, and Gulz, 2011) also shows that teaching a digital tutee can affect self-efficacy 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 and colleagues (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 self-efficacy, 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' 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 Bolay (2004) conducted a follow-up-study with similarly strong vs. weak learning companions. Under the one condition, students were regularly reminded by the system to collaborate with the learning companion and

encouraged to work for a high score; under another conditions 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 (Kirkegaard, Tärning, Haake, Gulz, & Silvervarg, 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, Haake, and Gulz (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 for two 45-minute 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' potential self-efficacy change, performance and attitude towards the agent.

2.3. Characteristics of pedagogical agents more broadly, in relation to student learning – potential SE-change, performance and attitude towards one's agent

Baylor and Kim have conducted a series of studies exploring whether certain characteristics in pedagogical agents can affect *self-efficacy (SE) change* in

students. One study (Baylor & 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 & 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. Another (Baylor & 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, Baylor, Plant, and Doerr (2008) studied possible effects of pedagogical agents' perceived gender, age and 'coolness' (equated with having a cool hairstyle and cool clothes) on female students' self-efficacy in engineering related fields. Students interacting with a young and cool agent had higher SE at the experiment's end.

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 Johnson, Mayer, Rizzo, Shaw, and Collins (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, Ozogul, and Resisslein (2015) found that an agent who pointed 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, Hamilton, Zheng, and Baylor (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 self-efficacy. Students who collaborated with a responsive, but not with a non-responsive agent, showed a significant increase in their self-efficacy 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.

2.4. Similarity attraction

Human beings often tend to like people they perceive as similar to themselves; a phenomenon known as *similarity attraction* (Newcomb, 1956; Byrne & Nelson, 1965; Byrne, Griffitt, & Stefaniak, 1967; Nass & 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 self-efficacy 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 likelier to work if the learner sees herself as sufficient similar to whom she is observing. Bandura 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.

2.5. Previous research on matching versus 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, Moon, Fogg, Reeves, & Dryer, 1995; Nass & 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, Hauser, Ugray, and Patterson (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 and colleagues (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. 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 & Niemirepo, 1998, p. 182). In other words, there is a high-performing companion with high self-efficacy versus a low-performing companion with low self-efficacy – 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 & Niemirepo, 1998, p. 191).

2.6. Matching versus mismatching SE 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 modelling whom? In effect, role modelling 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 performance, potential SE change and attitude toward the tutee – become difficult to predict.

3. 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' 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 self-efficacy – 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.

3.1. How do students respond to digital tutees with high versus low self-efficacy?

Q1. How do students respond to a digital tutee with high versus low self-efficacy?

- *Q1.a. How will the digital tutee's SE – high or low – affect students' attitude toward the tutee?*
- *Q1.b. How will the digital tutee's SE – high or low – affect a potential change in students' own SE?*
- *Q1.c. How will a digital tutee's SE – high or low – affect students' performance?*

According to previous studies we know that digital tutees as such, regardless of whether or how they express their self-efficacy, 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.

3.2. Does a match vs. mismatch between SE – high or low – between student and digital tutee have effects on 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?

- *Q2.a. 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 & 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.

Q2.b. Does match/mismatch in SE between student and digital tutee have effects on students' potential SE change?

Q2.c. Does match/mismatch in SE between student and digital tutee have effects on students' performance?

Some studies have shown, compare section 2.5 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/mismatch in SE – neither for digital companions, nor for digital tutees.

4. Method

The study comprised one pre-test session, seven game-playing sessions, and one post-test session – all sessions lasting 30-40 minutes. During the pre-test session students took a math test and filled out an SE questionnaire (Appendix B). 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 A).

4.1 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 self-efficacy or one who expressed low self-efficacy. 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).

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

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

4.2. Material

4.2.1. 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

hundreds). The student is to receive one point. The digital tutee, Lo (see figure 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, Haake, and Gulz (2013) report positive educational effects for visual androgyny.

Figure 1. The digital tutee Lo.

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 figure 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 figure 2 is "*Yes*, 1 point" – which is what the student has selected.

[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.

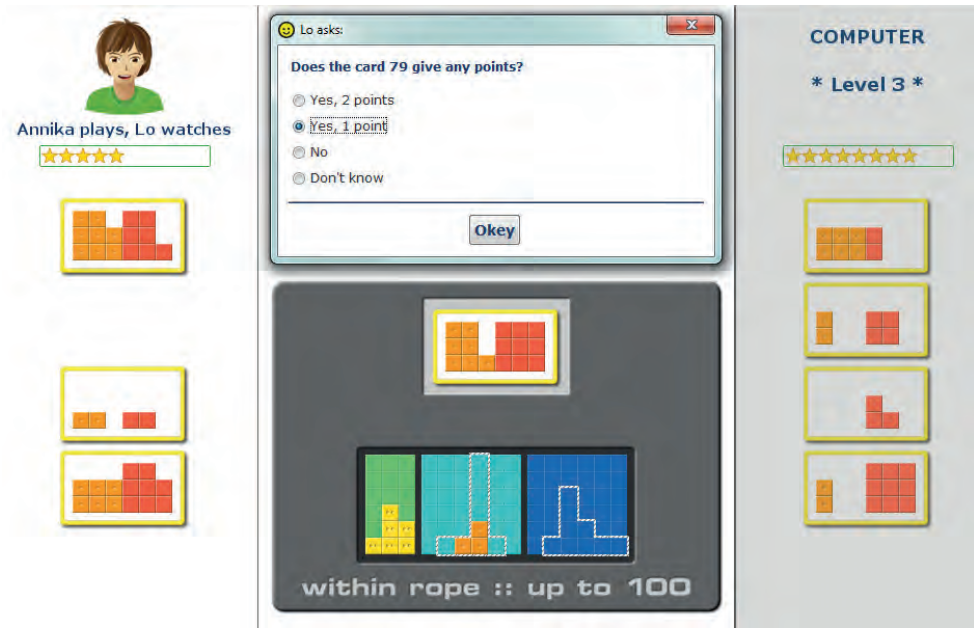


Figure 2. With the game in observe mode, the digital tutee asks the question: "*Does the card 79 give any points?*" The answering options are: "*Yes, 2 points*", "*Yes, 1 point*", "*No*", and "*I don't know*".

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).

4.4.4. The chat

In addition to the scripted 'conversation' via multiple-choice questions, the present version of the game also includes a chat (Silvervarg & 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 (figure 3).

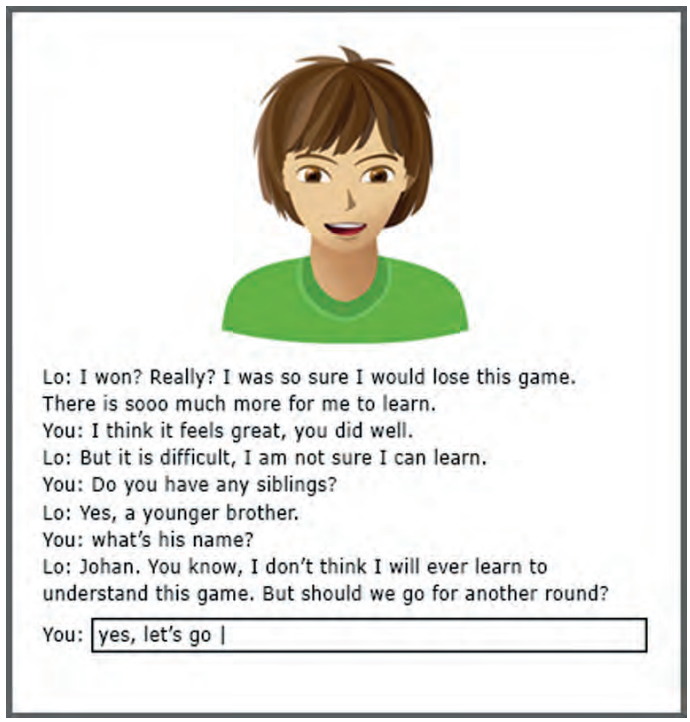


Figure 3. An example from chatting with Lo when expressing high SE.

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 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.

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 began 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 self-efficacy). 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 self-efficacy). 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 self-efficacy).

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).

4.2.3. Measurements

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

4.2.4. Attitude towards the tutee

The questionnaire on the students' experiences and opinions (Appendix A) 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 booring' 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 to 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 self-efficacy 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.

4.2.5. Self-efficacy

The pre- and post-test SE questionnaire (Appendix B) was based on Bandura, Barbaranelli, Caprara, and Pastorelli (1996); adapted for this study and translated into Swedish.

Seven items targeting the students' self-efficacy 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 B), making up a five-level Likert scale equivalent to the one in the 'opinions and experiences questionnaire'.

The self-efficacy score for each student was calculated as the sum over all items, resulting in a value ranging from 7 to 35. Self-efficacy 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.

4.2.6. Performance

Student 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?*".

A 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.

4.3. 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 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.

[2] The math test was done for another paper (Tärning, Haake & Gulz, 2017) and the results were not used in this paper.

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 B). They were also asked to fill out the opinions and experiences questionnaire Appendix A). 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.

5. 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 4.1. This second dataset was divided into four groups of agent vs. student low/high self-efficacy; self-depending on the student's and agent's SE: high/high, high/low, low/high, and low (see table 1).

5.1. 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 self-efficacy the way it was intended.

5.2. Student responses to a high vs. low SE digital tutee

Previous research relating to our first research questions (Q1.a – Q1.c), points in different directions. We therefore felt that we could not make any predictions.

5.2.1. Attitude toward digital tutee (Q1a)

To determine whether the digital tutee's SE affected students' attitude towards the tutee, we compared scores for question 4, 7 & 8 from the questionnaire regarding students' experiences and opinions (Section 4.2.4.); 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.

Table 2. Medians (*Mdn*) for question 4, 7 & 8 (summarized and one by one) in the attitude questionnaire, addressing attitude towards the digital tutee with regard to the self-efficacy 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

5.2.2. Student self-efficacy change (Q1.b)

To determine whether the digital tutee's SE affected students' own SE, we compared students' pre- and post-testing SE scores (Section 4.2.5.). A two sampled t-test revealed no statistically significant difference on SE increase between the two student groups teaching a digital tutee of low ($M = 0.76$, $SD = 3.55$) vs. high ($M = 0.93$, $SD = 3.61$) self-efficacy ($t(140) = -0.282$, $p = .78$).

5.2.3. Student performance (Q1c.)

To determine whether the digital tutees' self-efficacy level affected students' performance, we looked at students' in-game performance (section 4.2.6). 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 self-efficacy 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.

5.3. Match vs. mismatch between student´s and digital tutee´s self-efficacy

For the next three research issues (Q2.a – Q2.c), the *Media Equation Theory* (Reeves and Nass, 1996) suggest a possible matching effect for self-efficacy with regard to attitude (Q2.a). For the other two questions (Q2.b and Q2.c), 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 self-efficacy questionnaire students were excluded (corresponding to one fifth of the students). The resulting data set consisted of 113 participants (see section 4.1).

5.3.1. (Mis)match and attitude toward digital tutee (Q2a)

Would a match or mismatch between students' and their digital tutee's self-efficacy affect the students' attitude, as measured by the total score (Range = 3 to 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 self-efficacy 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 self-efficacy 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 self-efficacy student groups revealed any significant effects between matched and mismatched agent x student self-efficacy conditions (low self-efficacy students: $W = 454$, $p = .38$; high self-efficacy students: $W = 376$, $p = .49$).

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

		Shapiro-Wilk					Mann-Whitney		
		W	*p*	*n*	*Mdn*	*Rng*	*W*	*p*	
all students	matched	0.863	< 0.001	55	10	4-13	1824	0.18	
	mismatched	0.920	< 0.001	58	10	4-13			
low SE students	matched	0.94362	0.140	27	10	6-13	454	0.38	(1)
	mismatched	0.89738	0.001	28	9.5	7-11			
high SE students	matched	0.89053	0.005	30	10	4-13	376	0.49	(1)
	mismatched	0.80193	< 0.001	28	10	4-13			

(1) *p*-values adjusted by means of Holm correction

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 self-efficacy 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 4. Median (*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' self-efficacy combinations.

	Median (*Mdn*) & Range (*Rng*) for '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	

Taken together, our prediction with respect to research question Q2.a was not supported; students who taught a digital tutee that was similar to them in terms of self-efficacy 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 self-efficacy.

5.3.2. (Mis)match and potential SE change (Q2a)

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 4.2.5.) The dataset of 113 was divided into matching vs. mismatching subgroups (figure 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 to 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 self-efficacy group ($V = 806$, $p = .020$), but not so for the mismatched group ($V = 781$, $p = .27$).

Diving into the low- and high-SE students' groups separately (figure 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 self-efficacy student groups show higher average SE improvements where the agent's SE matches (figure 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: $t(56) = .802$, $p = .82$; p-values Holm corrected to adjust for two-fold measurements).

(2) The low-SE student group (figure 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 wither for the matching or mismatched agent (figure 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 self-efficacy match-

mismatch was found. However, considering the difference between the two conditions for the low self-efficacy students and the possibility of ceiling effects for high self-efficacy students, we cannot conclusively disregard an effect of 'similarity attraction'.

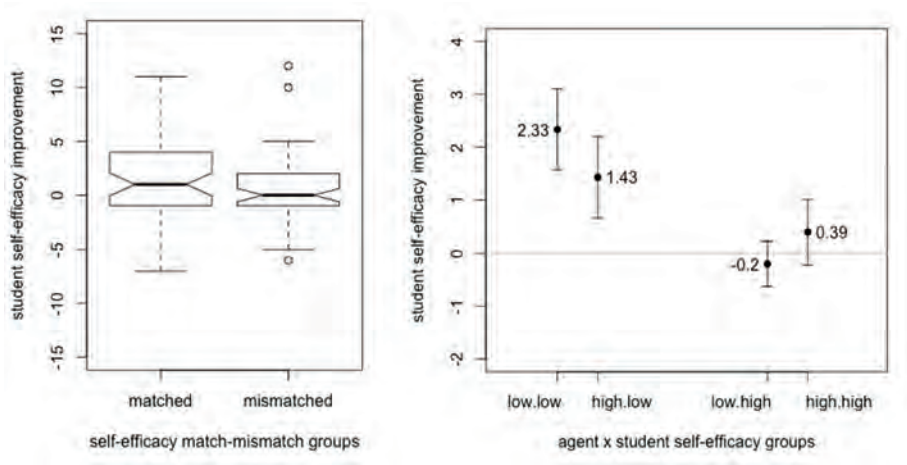


Figure 4. Left: boxplot of self-efficacy improvement for matched vs. mismatched tutee x student self-efficacy pairings. Right: self-efficacy improvement means and standard errors for matched vs. mismatched tutee x student self-efficacy pairings separated on low and high self-efficacy student groups.

5.3.3. (Mis)match and student performance (Q2.c)

To determine whether the match or mismatch of agent's and student's SE affected student performance, we looked at in-game performance (Section 4.2.6). 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 (figure 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: $t(56) = 0.719$, $p = .48$).

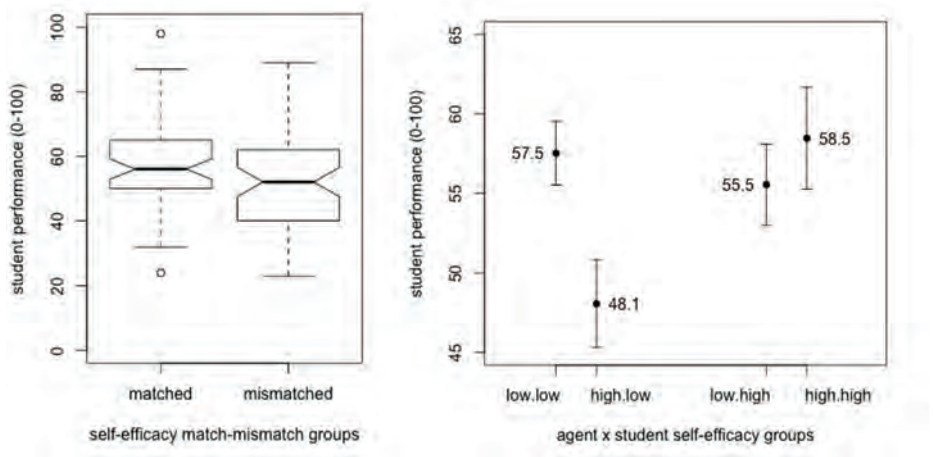


Figure 5. Boxplot of performance for matched vs. mismatched agent x student SE pairings (left); means and standard errors for all possible agent and student pairings (right).

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.

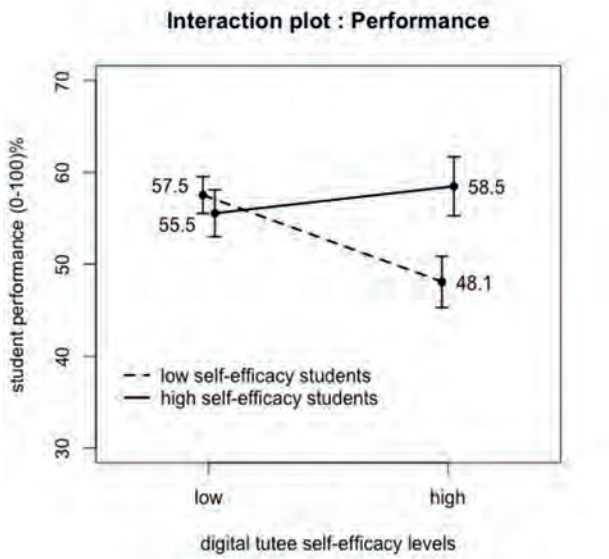


Figure 6. Student in-game performance with regard to digital tutee vs. student SE.

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

6.1. Results in the overall student population

Q1: How do students respond to a digital tutee with high versus low self-efficacy?

Q1.a. Will the digital tutee's self-efficacy affect students' attitude towards their tutee?

Q1.b. Will the digital tutee's self-efficacy affect potential increase or decrease in students' own self-efficacy?

Q1.c. Will the digital tutee's self-efficacy affect students' 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 performance than teaching one with high SE. A possible explanation may be found in the aforementioned *protégé effect* (section 2.2.) 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.

6.2. Matching/mismatching effects

Q2. (How) do student respond to match/mismatch regarding their own self-efficacy and that of their digital tutee?

Q2.a. Does match/mismatch between self-efficacy 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 2.5). Some (Hietala & Niemirepo, 1998, Nass & Lee, 2000; Kim, 2007) report a similarity attraction effect, i.e. more positive attitude towards an agent when matched; others (Isbister & Nass, 2000; Behrend & 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).

Q2.b. Does match/mismatch between self-efficacy in student and digital tutee have effects on students' potential increase in self-efficacy?

One study (Pareto, Haake, Lindström, Sjödén, & Gulz, 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 – but 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 self-efficacy in a domain for someone who already has a high (stable) self-efficacy in the domain?

Q2.c. Does match/mismatch between self-efficacy in student and digital tutee have effects on students' 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' 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 likelier 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 performance was, in this case, comparable to the 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 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 la ow 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 performance but not on attitude.

7. 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.

8. 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' 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 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, 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 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.

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 *Proc. of the 11th Conference on Advanced Learning Technologies (ICALT)*, (pp. 506-510). Piscataway, NJ: Institute of Electrical and Electronics Engineering.

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., Ryu, J., & Shen, E. (2003). The effects of pedagogical agent voice and animation on learning, motivation, and perceived persona. Paper presented at the *Annual World conference of Educational Multimedia, Hypermedia, & Telecommunication*, Honolulu, Hawaii.

Baylor, A. L., & Kim, Y. (2004). Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. Paper presented at the *Intelligent Tutoring Systems* (pp. 592-603). Maceió, Alagoas, Brazil.

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, NJ: Lawrence Earlbaum Associates.

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

Frasson, C., & Aimeur, E. (1996). A comparison of three learning strategies in intelligent tutoring systems. *Journal of Educational Computing Research*, *14*(4), 371-383.

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., 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* (pp. 335-341). Bloomington, IN: International Society of the Learning Sciences.

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.

Kirkegaard, C., Tärning, B., Haake, M., Gulz, A., & Silvervarg, A. (2014). Ascribed gender and characteristics of a visually androgynous teachable agent. In *Proc. of International Conference on Intelligent Virtual Agents* (pp. 232-235). Heidelberg, Germany: Springer International Publishing.

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

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 *Proc. of the 1st Workshop of Embodied Conversational Characters (WECC'98)* (pp.123-128). Tahoe City, CA: ACM Pres.

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 co-learner 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.

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.

Ozogul, G., Johnson, A. M., Atkinson, R. K., & Reisslein, M. (2013). Investigating the impact of pedagogical agent gender matching and learner choice on learning outcomes and perceptions. *Computers & Education*, *67*, 36-50.

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., 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.

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 *Proc. of International Conference on Artificial Intelligence in Education* (pp. 247-255). Berlin/Heidelberg, Germany: Springer-Verlag.

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* [Software]. R Foundation for Statistical Computing, Vienna, Austria.

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*. 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 *Proc. of Intelligent Virtual Agents* (pp. 153-160). Springer Berlin/Heidelberg.

Silvervarg, A., Haake, M., & Gulz, A. (2013). Educational potentials in visually androgynous pedagogical agents. In *Proc. of International Conference on Artificial Intelligence in Education* (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 *Proc. of International Conference on Artificial Intelligence in Education* (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 *Proc. of the 39th Annual Conference of the Cognitive Science Society* (pp. 1169-1174). London. UK: Cognitive Science Society.

Uresti, J. A. R. (2000). Should I teach my computer peer? Some issues in teaching a learning companion. In *Proc. of International Conference on Intelligent Tutoring Systems* (pp. 103-112). Berlin/Heidelberg, Germany: Springer-Verlag.

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.

Appendix A

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 B

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?	☐	☐	☐	☐	☐

Paper III

"*I didn't understand, I'm really not very smart*" – How design of a conversational teachable agent with low self-efficacy can contribute to student performance and self-efficacy

Since the publication of this thesis, a revised version of this article is published as:
Tärning, B. & Silvervarg, A. (2019). "I didn't understand, i'm really not very smart" – How design of a digital tutee's self-efficacy affects conversation and student behavior in a digital math game. *Education Sciences*, *9*(3), 197.
https://doi.org/10.3390/educsci9030197

Betty Tärning[1] & Annika Silvervarg[2]

[1] *Div. of Cognitive Science, Lund University, Lund, Sweden*
[2] *Dept. of Computer and Information Science, Linköping University, Linköping, Sweden*

Abstract. In this paper, we explore how self-efficacy of a teachable agent in an educational math game affects how students interact with the agent in a chat conversation. 89 students interacted with a teachable agent that expressed either high or low self-efficacy. Results showed that a teachable agent with low self-efficacy was treated in a more positive manner than a teachable agent with high self-efficacy. The students replied more often to its comments and in a more positive way. The students also gave less negative comments regarding its intelligence and competence than they did to a teachable agent with high self-efficacy. For students with low self-efficacy there was also a correlation between how well the students performed in their teaching of the agent and the frequency of positive comments they provided on the teachable agents' intelligence or competence in the chat conversation. Some students encouraged the tutee with low self-efficacy, but some were a bit frustrated or saddened by its low self-efficacy. In conclusion, our study shows that designing a teachable agent with low self-efficacy can have several benefits. However, there were some indications that student reacted differently, and therefore more studies are needed to explore this further.

Keywords. Educational game, Teachable agent, Conversational agent, Feedback, Self-efficacy

Introduction

How should a pedagogical agent in a virtual learning environment be designed to support student learning? The question is complex, since there are many types of pedagogical agents, which can take different roles in the learning process. There are, for example, tutor agents from which the student can learn (Graesser, Wiemer-Hastings, Wiemer-Hastings, Kreuz, & Tutoring Research Group, 1999), agents that work together with the student as companions or peers (Chan & Chou, 1997; Hietala & Niemirepo, 1998, Kim, Baylor, & PALS Group, 2006a), agents that take an expert role (Johnson, Rickel, & Lester, 2000; Graesser, Person, Harter, & Tutoring Research Group, 2001) and agents that act as mentors (Baylor, 2000; Baylor & Kim, 2005). Yet another pedagogical role is that of a tutee, where the agent acts as the one being taught while the real student takes the teacher role. These agents are called teachable agents (Blair, Schwartz, Biswas, & Leelawong, 2007) or, as we refer to them, *digital tutees*. A digital tutee is based on the idea that by teaching someone else you learn for yourself. Indeed, *learning by teaching* has been shown to be an efficient way to learn (Bargh & Schul, 1980; Annis, 1983; Renkl, 1995).

Results of many studies involving pedagogical agents indicate that they can have positive effects on learning (Atkinson, 2002; Baylor, 2002; Moreno, Mayer, Spires, & Lester, 2001; Kim & Baylor, 2007) and on self-efficacy (Kim, Hamilton, Zheng, & Baylor, 2006b; Kim, Wei, Xu, Ko, & Ilieva, 2007b; Pareto, Arvemo, Dahl, Haake, & Gulz, 2011; Kim & Lim, 2013). However, the mere addition of a pedagogical agent to a learning environment does not automatically improve learning or provide other beneficial effects. There are several aspects that need to be carefully considered when designing a pedagogical agent: visual appearance (Gulz & Haake, 2006; Baylor, 2009; Kim, 2016), how the pedagogical agent behaves and interacts (Baylor & Kim, 2005; Kim, Baylor, & Shen, 2007a; Wang, Johnson, Mayer, Rizzo, Shaw, & Collins, 2008; Veletsianos, 2009), and other characteristics such as whether the agent has high or low domain competence (Hietala & Niemirepo, 1998; Uresti, 2000; Baylor & Kim, 2005; Kim et al., 2006a; Kim et al., 2006b; Kim, 2007).

In order to learn more about how pedagogical agents should be designed more studies are called for, that evaluate design features of agents and the effect they have in relation to different student groups. In this paper, we have chosen to study the characteristic of self-efficacy (i.e. someone's belief in his or her own ability to succeed with a task). Self-efficacy has, to our knowledge, not been implemented and studied in digital tutees.

This paper focuses on whether high or low self-efficacy in a digital tutee affects the way a student interacts with it in a chat conversation incorporated in an

educational math game. We have analysed chat dialogues between student and their digital tutee in order to find possible differences in the extent and manner in which the students engaged in conversation with their tutee. More specifically, we looked at the following: how students responded to feedback provided by the digital tutee, how they commented on its attitude and intelligence/competence, and the relations between these chat behaviours and students' performance.

Following previous research focusing on matching/mismatching effects of characteristics in students and pedagogical agents, respectively, we also investigated possible effects of the digital tutee and student having similar or dissimilar self-efficacy. A matching pair was when the digital tutee and the student both had low *or* high self-efficacy regarding mathematics (more specifically regarding the base ten concept) and a mismatching pair when the agent hade high self-efficacy while the student had low self-efficacy and vice versa.

Background

The type of pedagogical agent used in this study is a teachable conversational agent. It is teachable in that sense that it knows nothing about the topic from the beginning but then learns from the student, who acts as teacher. By teaching the agent, the student at the same time learns for herself. This pedagogical method has been proven efficient in human – human interaction. Moreover, similar results have been shown for human – agent interaction (Okita & Schwartz, 2013; Chase Chin, Oppezzo, & Schwartz, 2009). Chase et al. (2009) found that students who were asked to learn in order to teach a digital tutee put more effort into the learning task compared to when students were asked to learn in order to take a test for themselves. This difference in effort and engagement is referred to as the *protégé effect*. Not only did the students in the study put more effort into the task when they were later supposed to teach; but they also learned more in the end. Having a protégé (such as a digital tutee) can thus increase motivation for learning. In addition, Chase et al. (2009) propose that teaching a digital tutee can offer what they term an *ego protective buffer*: that is, the digital tutee protects them from the experience of direct failure, since it is the tutee that fails at a task or a test – even though students are generally aware of the fact that the (un)success of the tutee reflects their own teaching of it. Nevertheless, the failure can be shared with the tutee, which then shields the student from forming negative thoughts about themselves and their accomplishments.

Sjödén, Tärning, Pareto, and Gulz (2011) made another observation regarding the social relation a student can build with her digital tutee. They found that low-performing students improved dramatically on a post-test when they had their

digital tutee present during testing compared to when they did not. This difference was not found for high performing students. Even though the digital tutee did not contribute with anything but its mere presence, this presence had a positive effect for low performing students.

With respect to the learning by teaching paradigm, the kind of feedback students receive when using such software needs to be highlighted. In most studies on feedback and learning, feedback is something that is provided to the student from the teacher and concerns the students' performance. In the teachable agent paradigm, the direction of the feedback is different. Here it is the teacher (i.e. the real student) that receives feedback on how well (s) he has been teaching by observing how well her tutee (i.e. the digital tutee) performs. The digital tutee provides feedback to the student regarding its (the tutee's) own ability to solve tasks without explicitly saying anything about the students teaching abilities. Implicitly, however, a student can make use of the feedback from the digital tutee, including information on how the tutee performs, to infer how much she herself knows/or how well she has taught her tutee. This type of feedback is what Okita and Schwartz (2013) call recursive feedback; namely, feedback that occurs when the tutor observes her students use what (s)he has taught them. This type of recursive feedback is present in the math game used in this study. It appears when the digital tutee attempts to play the game independently, competing against a computer agent, using its knowledge regarding the rules and strategies of the game as learned from the student. However, it also appears in the chat dialogue when the digital tutee reflects upon its own learning and performance. These reflections make up our manipulation in that they are coloured by the self-efficacy the digital tutee is assigned (high or low).

Social conversation with pedagogical agents and digital tutees

The digital tutee used in the game also belongs to the pedagogical agent subgroup called conversational pedagogical agents. Conversational agents in the area of education are primarily able to carry out conversations relating to the learning topic at hand, which is referred to as *on-task conversation*. However, some of them are also able to carry out *off-task conversation* or 'small-talk', not related to the learning topic as such. Off-task conversation can make a learning situation more relaxed and has been shown to promote trust and rapport-building (Bickmore & Cassell, 1999; Cassell & Bickmore, 2003). Off-task conversation is also something that many students are familiar with from real world learning experiences. Classroom interactions encompass a mix of on-task and off-task interactions and the teacher does not just go on about the topic to be learned, usually there is an ongoing conversation with little (apparent) relation to the topic to be learned. To be noted is that not all students experience off-task conversation

as something positive; some find it time-consuming and meaningless (Veletsianos, 2012).

Previous research with the educational game used in this study investigated the effects of the off-task conversation within the chat[1]. Those results showed that overall, the students did not experience the off-task conversation as disturbing, and students who were allowed to engage in off-task conversation had a more positive game experience compared to students who did not have the opportunity (Silvervarg, Haake, Pareto, Tärning, & Gulz, 2011). The study also explored whether high-, mid-, and low-achievers would differ with respect to their experience of the off-task conversational module (the chat module). The outcome was that high- and mid-achievers liked the software more when the off-task conversation (the chat) was included – but that they chose to chat less than the low-achievers. Conversely, low-achievers were more indifferent towards the chat – but they chatted more than high- and mid-achievers. In a follow-up analysis of the material Tärning, Haake, and Gulz (2011) found that, the engagement differed between these sets of students. High-achieving students showed greater engagement than the low-achieving students did when chatting, but in situations where they appeared unengaged in the chat, they tended to choose to quit the chat and refrain from starting a new. The low-achievers on the other hand were more inclined to continue a chat even when they appeared disengaged. The authors speculate that low-achievers do not take control over their learning situation to the same extent as high-achievers.

Self-efficacy of students and digital tutees

A student's self-efficacy has been shown to be a predictor for academic success (Bandura, 1997). Students with high self-efficacy are more willing to take on a task they know they might not succeed with and persist longer with such a task compared to students with low self-efficacy (Bandura, 1997). Especially in the domain of mathematics, some students have such low self-efficacy that it hinders them in making progress. Therefore, it is relevant to understand if and how students' self-efficacy can be influenced. According to Bandura (1977), one way to influence self-efficacy is through observation. That is, observing someone else perform a task can change the belief one has in one's own abilities. Seeing someone else succeed with a task may create a sense of "If he can do it so can I".

For example, Kim et al. (2007b) showed that girls who interacted with a pedagogical agent in an educational math game developed a more positive attitude

[1] To be noted is that the previous version of the chat did not display any differences in personality traits of the digital tutee and it responded to and asked the same questions to every student.

towards mathematics and increased their self-efficacy beliefs in the subject compared to girls who played the same game but without an agent. Kim and Lim (2016) similarly found that girls increased their self-efficacy beliefs in learning mathematics after working with an animated agent embedded in computer-based learning. Pareto and colleagues (2011) found that third graders who taught a digital tutee for nine weeks showed a significantly larger gain in self-efficacy compared to the control group who had engaged in regular math classes.

One characteristic often analysed in pedagogical agents is their competence level and how this affects students' learning and self-efficacy. Kim et al. (2006a) manipulated the competence (high vs low) in combination with different interaction styles (proactive vs responsive) in a learning companion agent. They found that students who interacted with a companion with high competence were better at applying what they had learned and showed a more positive attitude towards their companion. On the other hand, students interacting with a learning companion with low competence showed an increase in self-efficacy. An increase in self-efficacy was also found for students who worked with a more responsive companion agent. Hietala and Niemirepo (1998) similarly found in their study that students in general preferred to collaborate with a more competent digital peer compared to a weak digital peer when left with a choice. Uresti (2000) presents opposite results pointing towards a trend (although not significant) that students who interacted with a weak learning companion learned more than students who interacted with a more competent companion.

Therefore, even though the evidence is not conclusive as to whether an agent with high or low competence is most beneficial, we know that pedagogical agents with a high or low competence can have an effect on students' self-efficacy and learning. Notably, this question has not been, and cannot be, studied with regard to digital tutees. The reason is that for digital tutees, the level of competence or expertise is not a variable that an experimenter can manipulate, since the competence of the digital tutee reflects the real students teaching. Simply put, if the student teaches the digital tutee well, the tutee will learn and increase its knowledge and competence. On the other hand, if the student does not teach her digital tutee well, the tutee will not increase its knowledge and competence.

In contrast, the characteristic of *self-efficacy* is possible to design and manipulate in a digital tutee and this is exactly what we have done. In (Tärning, Silvervarg, Gulz, & Haake, 2018), we studied whether the manipulation of self-efficacy in a digital tutee – in terms of low versus high self-efficacy – would affect any of the following for the (real) students who acted as teachers for the digital tutee: i) their self-efficacy, ii) their in-game performance, iii) their attitude towards the digital tutee. The study made use of an educational game targeting mathematics and the base ten concept (Pareto, 2014). The digital tutee interacted with the student both

via a scripted multiple-choice conversation and via a natural language chat conversation. In the chat conversation in which the digital tutee commented on her performance, expectations and ability to perform and learn, the tutees' self-efficacy was manipulated to be low or high.

The analysis in (Tärning et al., 2018) showed that overall students who interacted with a digital tutee with low self-efficacy performed better than students interacting with a digital tutee 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, the digital tutee with low self-efficacy had another positive effect on this student group. Students with low self-efficacy significantly increased their self-efficacy when interacting with a digital tutee with low self-efficacy. (They also increased their self-efficacy when interacting with a digital tutee with high self-efficacy, yet not as much nor significantly.)

In the present paper, we set out to (i) further understand the reasons behind the effects in Tärning et al. (2018), as well as (ii) explore how the effects might be harvested in future designs of digital tutees. To do this we conducted an analysis of the chat dialogues between students and their digital tutees, collected during the study. Our focus was on how the students responded to the feedback provided by the tutee that expressed either high or low self-efficacy, and how they perceived and interacted with their digital tutee.

Research questions

The previous study and analysis (Tärning et al., 2018) indicated that a digital tutee that displays low self-efficacy is more beneficial when it comes to increasing low self-efficacy students' performance and self-efficacy than a digital tutee with high self-efficacy. A possible reason for this is that students put additional effort into the task of teaching when teaching someone that appears to be in more need of it. The protégé effect (Chase et al., 2009) proposes that students put more effort into the task when they learn in order to teach someone else as compared to when they learn for themselves. Assumingly this applies to the general situation of teaching someone else. However, it is possible that additional effort is made when teaching someone with low self-efficacy who is likely to come across as someone in more need of help.

By analysing the chat dialogues between students and their digital tutees, we hoped to find patterns in how students interacted with a digital tutee with high vs low self-efficacy that could support this possible explanation or indicate alternative ones. Since the study is explorative, we made no predictions but openly

explored potential differences in interaction patterns with digital tutees exhibiting high self-efficacy and with digital tutees exhibiting low self-efficacy, respectively. We also wanted to compare these two conditions for matched and mismatched cases, i.e. where students had similar or dissimilar self-efficacy (low or high) as their digital tutee. More specifically the research questions were the following:

Q1. To what extent and how do students react and respond to the digital tutees' feedback?

Q2. To what extent and how do students comment on the digital tutees' intelligence and competence?

Q3. To what extent and how do students comment on the digital tutees' attitude?

Q4. Are there any relations between students' chat behaviour and students' performance?

Method

This study was part of a larger data collection that included several instruments; pre and post self-efficacy questionnaires, a pre-test in math, a post-questionnaire targeting student experience of the agent and game, data logging of the students' game play, and chat logs. For this study, we have used the self-efficacy questionnaire, parts of the game play log and the chat logs.

Participants

In total 166 fourth graders (83 girls and 83 boys) participated in the data collection. They were recruited from four schools and nine classes in Southern Sweden in areas with relatively low socio-economic status and school performance below average. Student's self-efficacy was assessed with a questionnaire based on Bandura, Barbaranelli, Caprara, and Pastorelli (1996), adapted to fit this study's purposes. The questionnaire used the same question stem; "*How good are you at solving these types of tasks?*" translated into Swedish. All seven questions related to the base ten concept since this was the topic in the game, for example "*How good are you at solving these types of task?*": "*Which number should be in the blank 670 − ____ = 485?*" or "*You have the number 274, if you add 3826 will the result end in 00?*". All items were graded in five steps from "*not good at all*" to "*very good at*".

The students were assigned to one of two conditions: a digital tutee that expressed high self-efficacy or a digital tutee that expressed low self-efficacy. The two

groups were balanced with regard to the students' self-efficacy. Thus, the number of matched and mismatched pairs of student and tutee with respect to self-efficacy was equal in the two conditions. For the purpose of the analysis, the students were divided into one of three groups (low, mid or high) according to the results on the self-efficacy questionnaire. The groups were adjusted so that all students with the same result were categorized in the same group. Students who belonged to the mid self-efficacy group were then removed from the analysis since we wanted to focus on the students at the extreme ends, those with the highest and lowest self-efficacy. Further nine students were removed from the analysis due to missing data or scarce attendance, and so the study included data from 89 participants (47 girls and 42 boys). Thus, we ended up with 44 students belonging to the high self-efficacy group (M = 32.39, SD = 4.13), and 45 students belonging to the low self-efficacy group (M = 20.31, SD = 1.83). More details on how these were distributed for the student groups and tutee condition of high vs low self-efficacy are provided in table 1.

Table 1. Descriptive statistics for the four groups used for analysis, with tutee and student self-efficacy being high or low.

Student self-efficacy	Tutee self-efficacy	N	*M*	*SD*
Low	Low	23	20.35	4.16
Low	High	22	20.27	4.18
High	Low	23	32.30	1.69
High	High	21	32.48	2.02

The math game

The educational game targets basic arithmetic skills related to the base-ten concept, and is composed by a set of board games, see figure 1. Instead of using numbers, the game uses blocks and boxes to visualize the base-ten concept. For more details regarding the game, see Pareto (2014).

The game also incorporates a digital tutee, named Lo (see figure 2) whom the student should teach how to play the game, which requires understanding of the base ten concept. For this article, we were not interested in whether or not the digital tutee's visual gender would influence the conversation between student and digital tutee. Lo was therefore designed to look androgynous which meant that the students could form their own opinion of Lo's gender . A previous study (Silvervarg, Haake, & Gulz, 2013) has established that students in the target group indeed perceive this visual character as androgynous.

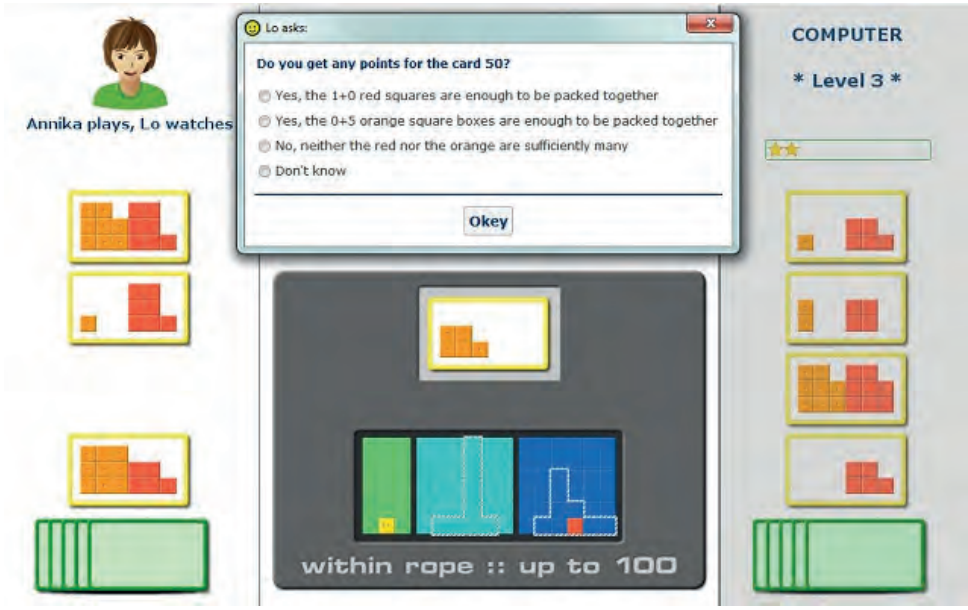


Figure 1. The game in Observe mode, Lo is asking a multiple choice questio.

The student can teach Lo in three different modes ('*Observe*', '*Try and be guided*', and '*On her own-mode*'):

– *Observe mode:* the student plays and the digital tutee learns by watching and by asking multiple choice questions to the student (see figure 1).
– *Try and be guided mode:* the digital tutee suggests which card to choose but can be corrected by the student who has the possibility to pick another possibly better card.
– *On her own mode:* the digital tutee plays on her own and the student has the opportunity to watch how well their digital tutee performs.

In 'observe' and 'try and be guided' mode the digital tutee sometimes asked multiple choice questions regarding the game and underlying math model (see section 3.4.2 Performance measures).

The chat

The chat is where the manipulation of the digital tutee's self-efficacy takes place. The chat appeared after every finished game, apart from when the student played alone without their digital tutee, which only happened in the beginning when the student was learning the game. The chat always started with a feedback- sentence that expressed the digital tutees self-efficacy (high or low) for example "*I have learnt a lot really quickly. I think I will have learned everything very soon*" (high self-efficacy) or "*It felt like I did not understand everything we went through*

during this round, I´m really not that smart" (low self-efficacy). That is, each chat started with the digital tutee saying something reviewing the just completed round and it always ended with a question asking the student how (s)he thought the next round would turn out, for example "*How do you think the next round will go?*" The chat always ended in a way that also reflected its self-efficacy as, for example "*You know, I don't think I will ever learn to understand this game. But should we go for another round*" (see figure 2 for an example of a typical chat dialogue). Every chat session lasted for one minute and was closed automatically when time was up.

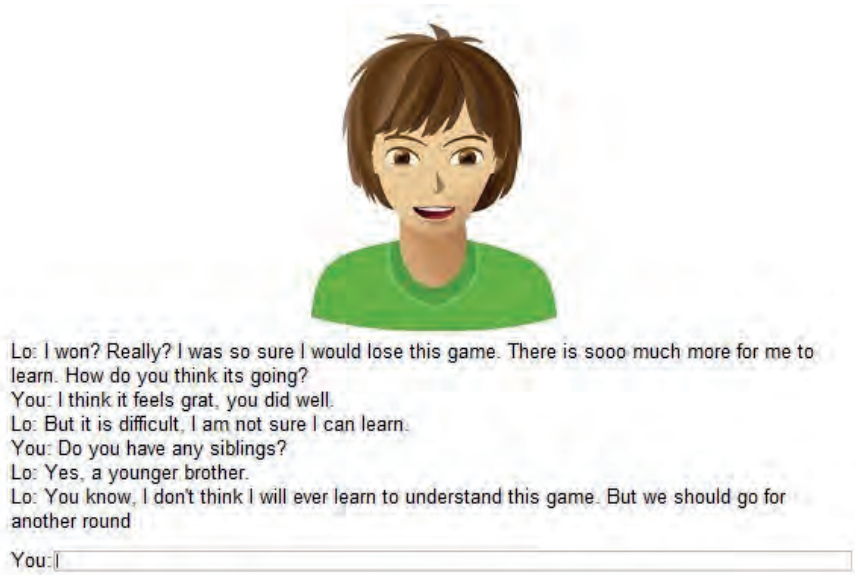


Figure 2. An example from chatting with Lo when displaying low self-efficacy.

Within the chat the students were free to talk about whatever they wanted to, for example, on-task topics such as school, math, the game or learning in general as well as off-task topics such hobbies, music and movies. In the current version, the digital tutee is able to handle greetings, ask and respond to questions and statements on various topics, ask and respond to follow up questions, and to tell mini-narratives, illustrated in figure 3.

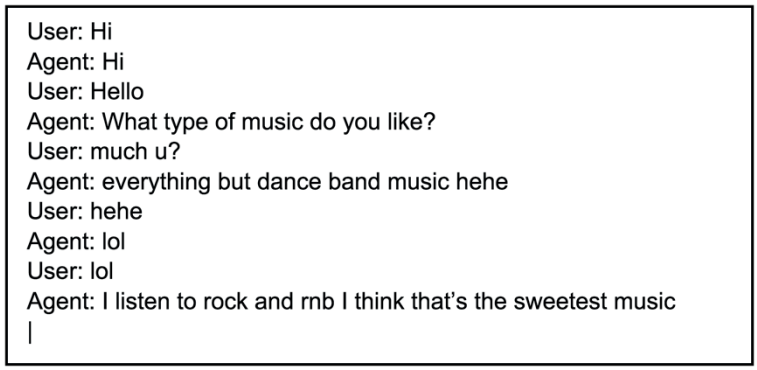


Figure 3. An example of a chat session with greetings, questions, and follow-up questions.

The chat allows for mixed-initiative, which means that both the digital tutee and the student can take the initiative and ask questions. The student can ignore a question from the tutee and instead pose a new question. When the digital tutee did not understand it had a strategy where it first asked for a clarification, then made a general request to change topic, and thirdly suggested a novel topic, see figure 4.

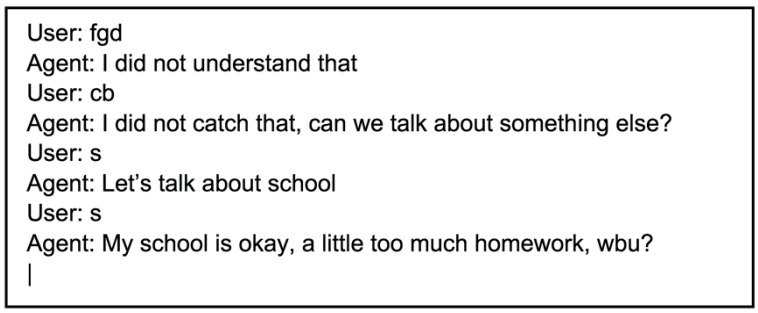


Figure 4. An example of a chat session where the tutee does not understand the student's utterances.

Feedback sentences

One of the authors constructed the feedback sentences and pilot tested them on 22 fourth graders who were not a part of the major study. The students in the pilot study were asked to read the sentences (presented in a randomized order) in order to evaluate whether they reflected high or low self-efficacy. They were asked to judge whether each sentence sounded like something being said by someone that was confident, not confident, or neither. The sentences that were not considered as either high or low in self-efficacy were then adjusted to match the approved sentences. Overall, there were 136 different sentences, 68 portraying Lo with high self-efficacy and 68 portraying Lo with low self-efficacy.

Since the game itself has three different modes (‘*observe*’, ‘*try and be guided*’ and ‘*on her own*’) the sentences also needed to correspond to these three modes. In the observe mode the student plays him/herself and the digital tutee is just learning from observing. For example, a sentence that appeared after a game in observe mode could say “*I’m learning the rules slowly, I’m not such a brilliant student*” (digital tutee with low self-efficacy). All sentences in this mode were expressed in a first-person perspective (“*I*”), since the digital tutee only observed what the student did.

In the try and be guided-mode (where the digital tutee could try for herself with the student correcting if they thought that the digital tutee proposed the wrong card), the digital tutee could express sentences in both I- and we-form, for example “*That’s great! I was sure that we were going to win, I think we played really well*” (digital tutee with high self-efficacy).

In the last mode (On her own) the student was not actively participating, instead the digital tutee played herself while the student was watching. After a game in this mode the sentences were again expressed in first person for example “*Buhu, how could I lose?! I played awesomely well, and I chose the best cards*” (digital tutee with high self-efficacy).

Each game mode was in turn divided into subcategories: ‘*game result + gameplay*’, ‘*game result + learning*’ and ‘*game result + agent knowledge*’. Each sentence started with a comment on the outcome of the previously played round of game – victory, defeat or even (i.e. ‘game result’). ‘Gameplay’ refers to how well Lo thought she had played “*That’s awesome, we won since we choose the best cards the whole time*” (high self-efficacy). ‘Learning’ reflected how much she thought she had learned during the previously round “*I really didn’t learn much this round, but maybe that was not so unexpected*” (low self-efficacy), and ‘agent knowledge’ how much she thought she knew about the game in total “*Wahoo, I won! But that was not so unexpected considering how good I am and how much I have learned by now*” (high self-efficacy). The order in which the sentences appeared in the chat was randomized according to a pre-programmed schedule and appeared for the suitable game mode.

However, in game-mode ‘observe’ there was no ‘game result + gameplay’ because the tutee was not involved in playing but only observed, and in game-mode ‘on her own’ there was no ‘game result + learning’ since the tutee did not learn anything from the student in that mode.

The chat always ended with a sentence from Lo regarding her thoughts about the upcoming round, for example “*I have a feeling that the next round will go really good, let’s go!*” (high self-efficacy) or “*I don’t feel like I understand anything but*

let's play another round" (low self-efficacy). For a summary of feedback examples, see table 2.

Table 2. Examples of sentences by the digital tutee, reflecting either high or low self-efficacy.

Game mode	Self-efficacy	
	Low	High
Observe	game result + learning	game result + learning
	"*It felt like I did not understand everything we went through during this round, I am really not smart.*"	"*It felt like I understood everything we went through during this round, I really am a genius.*"
	game result + agent knowledge	game result + agent knowledge
	"*I haven't learned so very much yet. I guess I have a lot more things to learn.*"	"*I have learnt a lot quickly. I think I will have learned everything very soon.*"
Try and be guided	game result + gameplay	game result + gameplay
	"*Did we win?! Wow, I thought we chose the wrong cards the whole time.*"	"*We got pretty bad cards, but we still won We really play brilliantly.*"
	game result + learning	game result + learning
	"*We lost… But I feel rather uncertain regarding the rules so maybe it wasn't so strange that we didn't win.*"	"*We didn't win but that was just bad luck. I feel very certain about the rules and how the game is played.*"
	game result + agent knowledge	game result + agent knowledge
	"*I still don't feel like I know anything about the game, I am glad we won!*"	"*I feel like I know everything about the game now, I don´t know how we could lose?!*"
On her own	game result + gameplay	game result + gameplay
	"*I lost, maybe I am not so very good at choosing the right cards…*"	"*It sucks that we lost! I was so sure we were going to win this round, I thought we played really well.*"
	game result + agent knowledge	game result + agent knowledge
	"*Did I win?! I was so sure I would lose, it feels like I have so much more to learn.*"	"*Brilliant! I feel very certain about the rules now, so this round felt really easy.*"
Finishing sentences		
	"*I don't feel like I understand anything but let's play another round.*"	"*I have a feeling that the next round will go really good, let's go!*"

Procedure

The seven game sessions in which the students interacted with their digital tutee were preceded by a pre-session in which the students filled out the self-efficacy questionnaire and a math pre-test (the math test was not used in this study). The self-efficacy questionnaire was the base for which the division between students in high, mid and low self-efficacy groups was made (where the mid-group was not a part of our analysis).

Within the game sessions, the students used the game individually, sitting in front of a stationary computer or a laptop (depending on schools). Each game session lasted approximately 30-40 minutes each. The first time they were instructed to play the game themselves (without the digital tutee) in order to be acquainted with the game. When they had grasped the gist of the game, they were asked to start instructing their digital tutee. Each student always instructed one and the same digital tutee and therefore always got consistent feedback in that sense that they only communicated with a digital tutee that was portrayed as having low or high self-efficacy.

After the seven game sessions, there was a post session in which the students filled out the same self-efficacy questionnaire once more. The also filled out an attitude questionnaire and a math post-test (that were not used in this study). The students were also thanked for their participation and were debriefed regarding the gist of the study.

Dependent measures

Data collected through chat logs and data logs of game play formed the basis for the dependent measures presented below.

Chat measures

Based on the research questions the authors constructed a coding schema by adding some new categories to an already existing schema (Silvervarg & Jönsson, 2013). Categories that could account for frequency and valence (positive/negative/neutral) of the students' responses to the digital tutee´s feedback were added, as well as categories for frequency and valence of students' comments on the tutees intelligence, competence and attitude. Below the measures and the corresponding coding categories are explained.

Responses to feedback

To measure the students' responses to the feedback provided by the tutee we used the code *AnswerToFeedback* (AFB) for when the student replied to the feedback.

The valence of the reply –positive, negative or neutral – was also coded. In addition, we coded what topic the reply related to: math, the game, learning or knowledge. Examples of such sentences are: "*Good, you are a brilliant student*", "*I think we got bad cards*", "*Good, but I think I chose a too difficult level for you, sorry*" and, "*The computer must have cheated*".

IgnoreFeedback (IFB) was used when the students ignored the given feedback from the digital tutee and when they started talking about something completely different not relating to the feedback from the digital tutee. Examples of this could be that they started by asking their digital tutee something not related to the game, such as "*What is your mother's name?*" or "*Do you like football?*" Some students also replied with nonsense such as random letters.

These categories were used to compute the measures of frequency of responses to feedback, as well as frequency of positive feedback. Both expressed as a value between 0 (%) and 100 (%).

Comments on the tutee's intelligence or competence

Many students remarked on their digital tutee's intelligence or competence, which was coded with *CommentOrQuestionOnIntelligenceOrCompetence* (CI). It was noted whether the comments were positive, negative or neutral. Examples of these types of remarks are: "*You are good*", "*You are very good at math*". These would be regarded as positive remarks about the digital tutee. Examples of negative remarks could be: "*I am a bit worried about you and your way of playing*".

This category was used to compute two measures, the number of comments regarding the digital tutees intelligence and/or competence made from students to their tutees, and the frequency of positive comments in relation to neutral and negative comments.

Comments on the tutee's attitude

CommentOnAttitude (CA) was coded for whenever the students remarked on the digital tutee's attitude (towards the game and learning) and whether or not this was done in a positive, negative or neutral way. For example, "*You have to believe in yourself*" and "*What do you mean, you have learned a lot, I lost*".

Due to sparse data, where most students had given none or only one positive or one negative comment, the frequencies of positive or negative comments on attitude were not calculated. Only the number of positive and negative comments from students to their agents was computed.

Performance measures

We also measured how well the students performed while teaching the digital tutee, which indirectly measures their own learning and skills. This was measured

in two ways; through the logging of their answers to the in-game multiple-choice questions about the game and the underlying model of the base ten concept, and how well they played.

Multiple-choice questions

Multiple choice questions where posed by the digital tutee three times during each game-mode (except in 'on her own' were the digital tutee played on her own.), and the student could provide a correct, incorrect or a 'Don't know' answer (see figure 1). 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?*" The answer reflects if the students understand that 10 red squares make up an orange box, and 10 orange boxes make up a yellow box.

A measure was calculated based on the percentage of correct answers in relation to incorrect answers using the formula (Correct answers – Incorrect answers + 100)/2. This resulted in a number between 0 and 100 where 100 means that the student answered all the questions correct, 0 means that all questions were answered incorrectly and 50 means that as many questions were answered correct as incorrect.

In-game performance

In-game performance was also measured in terms of the students' quality of gameplay, as represented by the average 'goodness value' (0-100) of each card the student selected during a game. In short, the goodness value is based on a comparison between the actual card, the best possible card available to the player and a theoretically best possible card, which reflects how good the choice is given the available options. Both the number of points the player can receive from the card and its strategic value in terms of preventing the opponent to receive points is taken into account. Importantly, even though goodness correlates with competitive outcome (winning correlates with high goodness), there are situations where the player cannot win (for example due to getting 'bad' cards) which can still reflect the player's knowledge and ability to choose the best alternative from a 'poor' selection. In other words, the goodness value provides a measure of performance, which, over time, reflects the player's learning progression in the game, independent of the number of wins and losses. For further details on the relationship between goodness values and game progression, we refer to Pareto (2014).

Research design and data analysis

This study employed a between subject 2 x 2 factorial design, with tutee self-efficacy and student self-efficacy as the two factors. For research questions, Q1 and Q2 two-way ANOVAS were performed to investigate the effect of the independent variables on the dependent variables regarding responses to feedback and comments on the tutees' intelligence and competence. Due to sparse data, for research question Q3 only the number of comments on the tutee's attitude was calculated. Instead, a qualitative approach was taken where all comments were collected and grouped based on their content. For research question Q4, a correlation analysis was performed to explore if the students' chat behavior related to students' game performance.

Results

The starting point for this paper was a wish to learn about the underlying mechanisms for what make students – in particular those with low self-efficacy – perform better and gain a higher self-efficacy belief when they interact with a digital tutee showing low self-efficacy than when they interact with a digital tutee showing high self-efficacy. In addition, we were interested in how the result might be exploited when designing pedagogical agents.

Responses to feedback

Since the self-efficacy of the teachable agent was expressed through feedback delivered in a chat, the first question, Q1, was: "*How do the students react and respond to the digital tutees feedback on what went on in the game?*" Our first step was to see if the students would acknowledge the feedback and questions from the tutee, such as for instance "*What do you think about the next round?*", or if they ignored this feedback from the digital tutee. Results were that, overall the students responded to 53% of the feedback.

A two-way ANOVA showed a small to medium sized significant main effect of the tutee's self-efficacy on frequency of response ($F(1,88) = 3.99$, $p < .05$, $\eta^2 = .045$), where students responded more frequently to feedback from the digital tutee with low self-efficacy ($M = 58.78$, $SD = 24.41$), than to feedback from the digital tutee with high self-efficacy ($M = 47.74$, $SD = 27.10$). There was no main effect of student self-efficacy on frequency of response ($F(1,88) = 1.39$, $p = .24$), nor an interaction effect of student and tutee self-efficacy ($F(1,88) = .245$, $p = .62$), see table 3 for means and standard deviation for these groups.

Table 3. Mean and standard deviation, M(SD), for frequency of response to the digital tutees feedback.

	Tutee with high self-efficacy	*Tutee with low self-efficacy*
Student with high self-efficacy	52.43 (25.95)	60.35 (19.12)
Student with low self-efficacy	43.27 (28.02)	56.91 (29.09)

The second step was to look at the cases where the student had actually responded to the feedback and the digital tutee´s question, formulated as for example: "*How does it feel for you?*". When responding the student could do it in either a positive way such as "*It feels very well, you did very well*", or a negative way such as "*It doesn't go very well, you need to practice more*", or in a neutral way, writing for example "*okay*". Results show that on average, 72% the responses were positive.

A two-way ANOVA showed a significant small to medium sized main effect ($F(1,87) = 4.87$, $p < .05$,$\eta^2 = .055$) between students with high self-efficacy who responded positively more often ($M = 78.55$, $SD = 24.46$), than students with low self-efficacy ($M = 65.34$, $SD = 30.85$). There were no main effect of tutee self-efficacy ($F(1,87) = 0.23$, $p = 0.63$) nor an interaction effect ($F(1,87) = 0.59$, $p = 0.30$), see table 4 for means and standard deviation for these groups.

Table 4. Mean and standard deviation, M(SD), for frequency of positive responses to the digital tutees feedback.

	Tutee with high self-efficacy	*Tutee with low self-efficacy*
Student with high self-efficacy	78.76 (28.43)	78.35 (20.85)
Student with low self-efficacy	62.27 (35.54)	68.41 (25.80)

The most frequent type of positive answers from the students was simply a "*good*" or "*well*" when answering the digital tutee how well they thought it proceeded. These replies accounted for approximate one third of all positive answers. Some were more superlative like *"great*" and *"awesome*" but these were rather few. One out of six answers commented on the tutees intelligence or competence, the most frequent answers being of the type "*You are good*", or "*You are learning*", or more seldom, "*You are clever*". The neutral answers were usually a "*don't know*", "*so-so*" or "*ok*". There were also occasions were the student instructed the digital tutee to observe more carefully or put more effort into the next round.

Frequent negative answers when the digital tutee asked, "*How do you think it's going?*" were "*badly*" or "*really badly*". Almost half of the negative answers were derogatory or even abusive comments about the tutee's intelligence or competence, like "*You suck*", "*You are dumb*" or "*You lost, idiot*".

To conclude, with respect to question Q1: *To what extent and how do the students react and respond to the digital tutees feedback on what went on in the game*, we note that students responded more frequently to feedback from the digital tutee with low self-efficacy, and that the responses were mostly positive. Thus, the trait of having low self-efficacy in a digital tutee can lead to more engagement and positive responses to comments about the tutee's self-efficacy, learning and performance.

Comments on the tutees' intelligence and competence

Next, we looked at research question Q2: *To what extent and how do the students comment on the digital tutees intelligence and competence?* These comments appeared in the free conversation following the feedback from the digital tutee. They were not prompted by the tutee, but rather came spontaneously from students, and occurred for 89% of the students overall – for 93% of the students interacting with a tutee with low self-efficacy, and 84% of the students interacting with a tutee with high self-efficacy. On average, each student gave 4.9 comments to their digital tutee, 5.7 for students interacting with a tutee with low self-efficacy and 4.0 comments for students interacting with a tutee with high self-efficacy. This means that nearly all students, on their own initiative, gave several comments to their tutee regarding its intelligence or competence.

Most of the comments regarding the digital tutees intelligence or competence, on average 60%, were negative, and 40% were positive. A two-way ANOVA showed a significant medium sized main effect of the tutee's self-efficacy ($F(1,78) = 5.71$, $p < .05$, $\eta^2 = .071$), where the digital tutee with low self-efficacy on average received more positive comments ($M = 49.91$, $SD = 37.26$), than the digital tutee with high self-efficacy ($M = 28.97$, $SD = 38.13$). There was no main effect of student self-efficacy ($F(1,78) = 0.34$, $p = 0.56$) nor an interaction effect ($F(1,78) = 0.21$, $p = 0.65$), see table 5 for mean and standard deviation for these groups.

Table 5. Mean and standard deviation, M(SD), for frequency of positive comments on the digital tutees intelligence or competence.

	Tutee with high self-efficacy	*Tutee with low self-efficacy*
Student with high self-efficacy	33.79 (36.98)	50.45 (37.30)
Student with low self-efficacy	24.67 (39.62)	49.35 (39.46)

Most of the negative comments involved saying that the digital tutee was an idiot or that (s)he sucked. However, some of the comments referred to the tutee's abilities to learn math and the game, for example, "*You are not very good at math*" or things like "*How can you be so stupid*" and "*I mean, do you actually have a brain?*". The positive comments mostly concerned the tutee's performance and ability to play, the student saying things like "*You are super good*" or "*You did very well*". However, students also expressed happiness regarding their digital tutees performance saying things like "*It feels very nice when you play as good as you do*" or "*Oh my God, you are really good, that is so fun to see!!!*"

Thus, the results for Q2: *To what extent and how do the students comment on the digital tutees intelligence and competence*, are in line with the results on Q1. The digital tutee with low-self efficacy received more positive comments about its intelligence and competence than the digital tutee with high self-efficacy from both students with low and high self-efficacy.

Comments on the tutees' attitude

Of special interest was to see if the students commented on the digital tutees attitude towards her learning and performance, since this attitude relates to the self-efficacy that the tutee expressed. Thus, we explored Q3: *To what extent and how do the students comment on the digital tutees attitude?* Our results show that only 21 out of 89 students made any comments regarding the digital tutees attitude. Out of these, 17 directed the comments to the digital tutee with low self-efficacy. In other words, for the digital tutee with low self-efficacy 17 out of 45 tutees received comments, while for the digital tutee with high self-efficacy only 4 out of 44 tutees received comments on its attitude. Equally, as many students with high self-efficacy (10) as students with low self-efficacy (11) provided these comments.

Of the four comments to the tutee with high self-efficacy one was negative "*What do you mean*", "*you have learned a lot, I lost*" and the other three were positive

"*It's good that you are confident. It will go well*", "*It will go well, just believe in yourself*" and "*Ok, but we need to fight on*".

The comments to the digital tutee with low self-efficacy are all listed in table 6 with the exception of similar comments from the same student in the same chat session. The comments are grouped together based on similarity. It was also noted whether a student with low or high self-efficacy gave the comment.

Table 6. Comments to the TA with low self-efficacy regarding its intelligence or competence from students with both low and high self-efficacy.

Student self-efficacy	Comment
High	*You could say well*
High	*Why do you think so negatively*
Low	*Yes, but you have to say something positive too*
Low	*You shouldn't be so fucking negative*
Low	*I know, but you are pretty good just think positively*
Low	*Don't be unsure you idiot*
Low	*But you should be, idiot*
Low	*Do not be unsure you will win*
High	*Good you did that well just stop being so unsure*
Low	*Lo it will be fine just relax*
Low	*Do not worry you will do it!*
Low	*It's cool, Lo*
High	*Good if you believe in yourself, I think you can do it*
High	*Tell yourself you can win*
High	*I think you can do it!*
Low	*You need to focus more, do what you should, do not think of anything else!*
Low	*I don't like it when you say so*
High	*Why do you ask when you score points all the time*
High	*I won with 21-8, what's wrong with you*
Low	*Good, but you will probably say that it was not good*
Low	*What do you talk about, it went really well*
Low	*You are super GOOD DONT YOU GET IT ????*
High	*Yes, I think it goes well for both of us, why don't you*
Low	*I got 11 stars, it's not nice to say that*

Out of the 25 comments to the tutee with low self-efficacy 14 state that the student and/or tutee is doing well. Eight of these comments express some frustration or sadness over the tutees negative attitude, for example "*Yes, I think it goes well for both of us, why don't you.*" The other six are more encouraging, trying to boost the tutee, for example "*Don't worry you will do it*" Overall many of the comments tell the tutee to be less negative and more positive (e.g. "*I know, but you are rather good, just think positive*"), to not be unsure (e.g. "*Do not be unsure you will win*") and believe in itself (e.g. "*Tell yourself you can win*"). Some students also tell the digital tutee to relax (e.g. "*It's cool, Lo*"), focus on the task (e.g. "*You need to focus more, do what you should, don't think of anything else!*"), or that it not kind to be so negative (e.g. "*I don't like it when you say so*"). These comments vary in tone, with some being rather harsh (e.g. "*You shouldn't be so fucking negative, don't be unsure you idiot*") and some very encouraging (e.g. "*Don't worry you will do it! Good if you believe in yourself, I think you can do it*").

The sparse data makes it hard to draw any definite conclusions, but it is important to note that while some students encourage the tutee when it express a low self-efficacy, some students also get a bit frustrated with it, especially if they think they or the tutee is performing well. Here there likely are differences for students with high and low self-efficacy. Since students with high self-efficacy often perform and teach their tutees better there will be a mismatch between the tutees' low self-efficacy and high performance, which can lead to frustration. For students with low self-efficacy of which many will also not teach their tutees equally well, the discrepancy between self-efficacy and performance of the tutee will not be as obvious.

Relations between chat and performance measures

Thus far, we have found differences in how students with high and low self-efficacy interact with digital tutees expressing high or low self-efficacy in the chat. Since our starting point was to explore the effect that students with low self-efficacy perform better as well as gain a higher self-efficacy belief when interacting with a digital tutee displayed as having low rather than high self-efficacy, our final analysis concerned Q4. *Are there any relations between students' chat behaviour and students' in-game performance?*

Students performance was calculated in two ways: (i) the proportion of correct and incorrect answers given by the student in relation to the multiple-choice questions posed by the digital tutee regarding the game and its underlying mathematical model and (ii) the goodness of cards chosen by the student during gameplay (see section 3.4.2 Performance measures). The first measure is more directly related to explicit teaching of the digital tutee, whereas the other is more of an indirect

measure of how well the student performs during gameplay when the tutee is learning through observation.

We computed a Pearson correlation coefficient (table 7) for the two performance measures as well as for the two measures from the chat on students' positive or negative attitudes towards the digital tutee and the feedback it provided: frequency of positive responses to feedback and frequency of positive comments on the tutees intelligence and competence. The comments on the tutees attitude had to be excluded due to the sparse data.

Table 7. The Pearson product-moment correlation coefficients for performance measures: answers to multiple choice questions, goodness, and chat measures: frequency of positive responses to feedback, and frequency of positive comments on the tutees' intelligence and competence.

	1	2	3	4
1. Answers to multiple choice questions	–			
2. Goodness	.338**	–		
3. Pos. feedback responses	.199	.031	–	
4. Pos. comments on intelligence and competence	.248*	-.016	.593**	–

*. Correlation is significant at the .05 level (2-tailed).
**. Correlation is significant at the .01 level (2-tailed).

Overall, we found a significant correlation with large effect size ($r(78) = .593$, $p < .01$) between the frequency of students' positive answers to the digital tutee´s feedback and the digital tutees comments on its own intelligence and competence. There was a significant correlation of medium effect size ($r(88) = .388$, $p < .01$) between the two performance measures; correctly answered multiple-choice questions and goodness (i.e. choosing the best card). We also found a significant correlation of small effect size between how well the student answered the multiple-choice questions and the frequency of providing positive comments on the digital tutees intelligence or competence ($r(79) = .248$, $p < .05$).

Table 8. The Pearson product-moment correlation coefficients for students with low self-efficacy.

	1	2	3	4
1. Answers to multiple choice questions	–			
2. Goodness	.204	–		
3. Pos. feedback responses	.280	-.130	–	
4. Pos. comments on intelligence and competence	.359*	.130	.698**	–

*. Correlation is significant at the .05 level (2-tailed).
**. Correlation is significant at the .01 level (2-tailed).

When looking at the different student groups we found no significant correlation between performance measures for students with low self-efficacy. But we did find a significant correlation of medium effect size ($r = .359$, $p < .05$) between their proportion of correct answers to multiple-choice questions and the frequency of positive comments they provided on the digital tutees intelligence or competence. The correlation between the frequency of positive comments they provided on the digital tutees intelligence or competence and the frequency of positive feedback responses was also significant and of large effect size ($r(39) = .698$, $p < .01$) (see table 8).

Table 9. The Pearson product-moment correlation coefficients for students with high self-efficacy.

	1	2	3	4
1. Answers to multiple choice questions	–			
2. Goodness	.368*	–		
3. Pos. feedback responses	.040	.041	–	
4. Pos. comments on intelligence and competence	.113	-.216	.464**	–

*. Correlation is significant at the .05 level (2-tailed).
**. Correlation is significant at the .01 level (2-tailed).

For the students with high self-efficacy another pattern came forth, see table 9. There was a significant correlation of medium effect size ($r = .368$, $p < .05$) between the proportion of correctly answered questions and goodness, as well as a significant correlation between the frequency of positive responses to the feedback and the frequency of positive comments they gave on the digital tutees intelligence or competence ($r(39) = .464$, $p < .01$). However, no significant correlation was found between their proportion of correct answers to the multiple-choice questions and the frequency of positive comments they provided on the digital tutees intelligence or competence.

Thus, students with low self-efficacy who express a more positive attitude towards their digital tutee in the sense of providing more positive comments on their digital tutee's intelligence and competence also perform better when they answer the digital tutee's multiple-choice question However, they do not play better, with reference to how they choose cards (i.e. the goodness value). The competence of students with high self-efficacy seems to be the driving force in how they perform. In this group, students who play the game well choose good cards (having a high goodness value) also answer the multiple-choice questions better, regardless of their attitude towards their digital tutee as expressed through their chat comments on the tutee's intelligence and competence.

Discussion

Based on the study Tärning et al. (2018) we drew the tentative conclusion that designing a digital tutee with low self-efficacy would be a good choice since the results of that study suggested that students with low self-efficacy benefitted from interacting with a digital tutee with low rather than high self-efficacy. At the same time, students with high self-efficacy were not negatively affected when interacting with a digital tutee with low self-efficacy; rather they performed equally well when interacting with both types of tutees.

With the follow-up analysis carried out in the present study, we hoped to get a deeper understanding of the results from the previous study and of how they may be exploited for the design of pedagogical agents in educational software. The analysis was based on the chat dialogues between student and digital tutee, dialogues in which the digital tutee expressed its self-efficacy (either high or low) when giving feedback to the student.

Our findings can be summarized as follows:

(i) Students responded more frequently to feedback from a digital tutee with low self-efficacy, and these responses were mostly positive.

(ii) Students gave a digital tutee with low-self efficacy more positive comments about its intelligence and competence than they did to a digital tutee with high self-efficacy.

(iii) Students' comments about the tutees attitude were almost exclusively given to the tutee with low self-efficacy. Most comments were positive, expressing that the tutee and/or student was doing well or were of an encouraging type. There were, however, also some comments that expressed frustration regarding the tutees low opinion of itself.

(iv) Students with low self-efficacy who expressed a more positive attitude towards their digital tutee, in the sense of providing more positive comments on the digital tutees intelligence and competence, also performed better when they answered the digital tutee's multiple-choice questions. However, they did not play better in the sense of choosing more appropriate cards (i.e. goodness value). For students with high self-efficacy we found another pattern, namely a relation between how well they played and how well they answered the questions asked by the digital tutee.

Below we discuss how these findings can be understood in the light of the following three constructs: the protégé-effect, role modelling, and the importance of social presence and relations. We know from previous studies that the protégé-effect is one of the underlying factors that make students who teach someone else (for example a digital tutee) learn more and be more motivated compared to

students who learn for themselves (Chase et al., 2009). That is, having someone who is dependent on you to learn and that you have responsibility for seems to lead to an increased effort. From our analysis, we see that students responded more frequently and more positively to a tutee with low self-efficacy than to one with a high self-efficacy, and that many students tried to encourage a tutee with low self-efficacy when they commented on its attitude, for example saying things like "*Tell yourself that you can win*" and "*Don't worry, you can do it!*". The tutee with low self-efficacy also received less negative comments on its intelligence and competence than the tutee with high self-efficacy.

Possibly students treat a digital tutee with low self-efficacy in a more positive manner since such a tutee comes across as someone more in need of help and who is more subordinate compared to a digital tutee with high self-efficacy. The experience of having a protégé to care for and to support might be especially relevant for students with low self-efficacy, in this case, notably low self-efficacy in mathematic. These students will, more often than students with high self-efficacy in math, lack the experience of being someone who teaches someone else Students with high self-efficacy are more likely to already in regular classes have taken a teacher role and assisted or supported a less knowledgeable and/or less confident peer.

Based on Banduras (1977) findings that a person's self-efficacy may be influenced by observing someone else performing a special task, one could have suspected that a digital tutee with high self-efficacy would function as a role model and thus boost the students' self-efficacy. Seeing someone else doing something may boost the thought: "*If (s)he can do it so can I.*" But in our analyses, we only found three instances of comments where the students agreed with the digital tutee when it expressed high self-efficacy, saying for example "*It's good that you are confident. It will go well*" and two of these comments came from students who themselves had high self-efficacy. Instead a kind of reversed role modelling may be going on in which the student can be a role model for the tutee with low self-efficacy in feeling that they are capable of completing a task and also teach it to someone else (i.e. the digital tutee). In our analysis, we found that when the digital tutee expressed a very negative attitude some of the students were positive and encouraged it with wordings such as "*You shouldn't be worried, you will make it*" or "*I know… but you are pretty good, you just have to think positive*" We did not find any comments where the student agreed with the tutee with low self-efficacy and expressed his or her own low self-efficacy.

Finally, we turn to the importance of the tutee's social presence. Sjödén et al. (2011) have previously shown that the social presence of a digital tutee can have a positive impact on low-performing students. Students with low self-efficacy do not equal low-performing students, but there is often a correlation between the groups

(Pajares, 2003; Schunk, 1995; Bandura, 1997). Looking at our analysis, we found that for students with high self-efficacy, their performance (i.e. goodness value) correlated with how well they answered the digital tutees multiple-choice questions. This correlation is not surprising since someone who answers the questions correctly is likely to be good at choosing good cards. What was interesting, however, was that we did not find this correlation for students with low self-efficacy. Instead, we found a correlation between how well they answered the digital tutees multiple-choice questions and to what extent they gave positive comments regarding their tutees' competence and intelligence. One can speculate that the social relationship the students had formed with their digital tutees had an effect on the students' performance.

With respect to the protégé effect, we found yet another interesting difference between students with high and low self-efficacy, respectively. Students with low self-efficacy seemed to make more effort than students with high self-efficacy when it came to the more social 'parts' in the game, such as interacting with the digital tutee in the chat and answering its questions within the game, parts that can be said to be more social than for example choosing cards when the digital tutee is more of a passive bystander. Possibly the social bond to the tutee was more important for students with low self-efficacy than for students with high self-efficacy. Students with high self-efficacy are already confident in themselves and what they know and have things under control, whereas students with low self-efficacy perhaps are more in a need of a social bond, a friend to support and maybe also get support from (someone that can strengthen their sense of knowing – something we also see in the result from Tärning et al. (2018)).

Limitations

Even though the digital tutee had the ability to talk about a wide array of topics its abilities were limited. You could sense that some of the students were a bit frustrated at points when the digital tutee could not answer the questions asked by the student. Maybe this led to more negative comments and frustration than otherwise would be the case.

The research questions focused on looking at agents and students on the extreme ends of the self-efficacy scale. The digital tutee was designed to have either clearly high or low self-efficacy and the analysis was restricted to the students with the highest and lowest self-efficacy score, with mid students excluded. Another way to do the study would be to look at self-efficacy as a continuous metric and investigate if there are linear relationships between the students' self-efficacy and other variables. The choice not to do so was partly based on limitations in resources to code the chat dialogues.

Conclusions and future work

How should a pedagogical agent in a virtual learning environment be designed to support learning? A previous paper by Tärning et al. (2018) explored the aspect: How should a digital tutee express its self-efficacy? A tentative conclusion was that it is more beneficial to design a digital tutee with low self-efficacy than one with high self-efficacy. In this paper we support this claim with our results that a digital tutee with low self-efficacy seems to boost the protégé-effect more and also promote a reversed role modelling where the student can boost herself through boosting the digital tutee.

Nevertheless, follow-up studies are required since we also found comments indicating a frustration with the feedback from the digital tutee with low self-efficacy. These have especially occurred in cases where the students or their digital tutee performed well but the tutee expressed a negative attitude. It is likely that the tutee's self-efficacy needs to be more adaptive and better reflect the rate at which it actually learns, which in turn reflects the proficiency of the student that is teaching it.

In our study, we only compared students with low and high self-efficacy, but we know from observations in classrooms that these are not homogenous groups. There are other factors that influence how students interact with pedagogical agents in virtual learning environments. For example, as pointed out, students with low self-efficacy to some extent overlap with the group of low-performing students, and in the latter group one can easily distinguish students that do not perform well due to not caring or trying, while others do care and try but fail nevertheless. This was also observed in another study using the same conversational chat, but without the digital tutee giving feedback (Silvervarg & Jönsson, 2011).

Another interesting area to explore is if more discussion regarding attitude and self-efficacy could lead to larger effects. Now the student can choose to ignore the tutee's feedback and refrain from responding to it, and it is not possible to respond to the tutee's concluding comment in each chat regarding how the next round will turn out. The chat could instead be designed so that the digital tutee took more initiative to discuss on-task topics, as well as its own intelligence and competence.

References

Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, *94*(2), 416-427.

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.

Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. *Journal of Educational Psychology*, *72*(5), 593-604.

Baylor, A. (2000). Beyond butlers: Intelligent agents as mentors. *Journal of Educational Computing Research*, *22*(4), 373-382.

Baylor, A. L. (2002). Expanding preservice teachers' metacognitive awareness of instructional planning through pedagogical agents. *Educational Technology Research and Development*, *50*(2), 5-22.

Baylor, A. L. (2009). Promoting motivation with virtual agents and avatars: role of visual presence and appearance. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1535), 3559-3565.

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

Bickmore, T., & Cassell, J. (1999). Small talk and conversational storytelling in embodied conversational interface agents. In *Proc. of AAAI Fall Symposium on Narrative Intelligence* (pp. 87-92). Cape Cod, MA.

Blair, K., Schwartz, D., Biswas, G., & Leelawong, K. (2007). Pedagogical agents for learning by teaching: Teachable agents. *Educational Technology & Science*, *47*(1), 56-61.

Cassell, J., & Bickmore, T. (2003). Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, *13*(1), 89-132.

Chan, T. W., & Chou, C. Y. (1997). Exploring the design of computer supports for reciprocal tutoring. *International Journal of Artificial Intelligence in Education*, *8*, 1-29.

Chase, C. 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.

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & Tutoring Research Group. (1999). AutoTutor: A simulation of a human tutor. *Cognitive Systems Research*, *1*(1), 35-51.

Graesser, A. C., Person, N., Harter, D., & Tutoring Research Group. (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*, *12*(3), 257-279.

Gulz, A., & Haake, M. (2006). Design of animated pedagogical agents – A look at their look. *International Journal of Human-Computer Studies*, *64*(4), 322-339.

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

Johnson, W. L., Rickel, J. W., & Lester, J. C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial intelligence in Education*, *11*(1), 47-78.

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

Kim, Y. (2016). The role of agent age and gender for middle-grade girls. *Computers in the Schools*, *33*(2), 59-70.

Kim, Y., Baylor, A. L., & PALS Group. (2006a). Pedagogical agents as learning companions: The role of agent competency and type of interaction. *Educational Technology Research and Development*, *54*(3), 223-243.

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

Kim, Y., & Baylor, A. (2007). Pedagogical agents as social models to influence learner attitudes. *Educational Technology*, *47*(1), 23-28.

Kim, Y., Baylor, A. L., & Shen, E. (2007a). Pedagogical agents as learning companions: the impact of agent emotion and gender. *Journal of Computer Assisted Learning*, *23*(3), 220-234.

Kim, Y. & Lim, J. H. (2013). Gendered socialization with an embodied agent: Creating a social and affable mathematics learning environment for middle-grade females. *Journal of Educational Psychology*, *105*(4), 1164-1174.

Kim, Y., Wei, Q., Xu, B., Ko, Y. & Ilieva, V. (2007b). MathGirls: Toward developing girls' positive attitude and self-efficacy through pedagogical agents. In *Proc. of the 13th International Conference on Artificial Intelligence in Education (AIED)* (pp.119-126). Los Angeles, CA: IOS Press.

Moreno, R., Mayer, R. E., Spires, H. A., & Lester, J. C. (2001). The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cognition and Instruction*, *19*(2), 177-213.

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.

Pajares, F. (2003). Self-efficacy beliefs, motivation, and achievement in writing: A review of the literature. *Reading &Writing Quarterly*, *19*(2), 139-158.

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 *Proc. of International Conference on Artificial Intelligence in Education* (pp. 247-255). Berlin/Heidelberg, Germany: Springer-Verlag.

Renkl, A. (1995). Learning for later teaching: An exploration of mediational links between teaching expectancy and learning results. *Learning and Instruction*, *5*(1), 21-36.

Schunk, 1995. Self-efficacy and education and instruction. In J. E. Maddux (Ed.), *Self-efficacy, adaptation, and adjustment: Theory, research, and application* (pp. 281-303). New York: Plenum Press.

Silvervarg, A., Haake, M., Pareto, L., Tärning, B., & Gulz, A. (2011). Pedagogical agents: Pedagogical interventions via integration of task-oriented and socially oriented conversation. Paper presented at the *Annual Meeting of the American Educational Research Association* (pp. 1-16). New Orleans, USA.

Silvervarg, A., & Jönsson, A. (2011). Subjective and objective evaluation of conversational agents in learning environments for young teenagers. In *Proc. of the 7th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. Barcelona: Spain.

Silvervarg, A., & Jönsson, A. (2013). Iterative Development and Evaluation of a Social Conversational Agent. In *Proc. of International Joint Conference on Natural Language Procession* (pp. 1223-1229). Nagoya, Japan.

Silvervarg, A., Haake, M., & Gulz, A. (2013). Educational potentials in visually androgynous pedagogical agents. In *Proc. of International Conference on Artificial Intelligence in Education* (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 *LNAI, vol. 6738: Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 337-344). Berlin/Heidelberg, Germany: Springer-Verlag.

Tärning, B. (2018). *Review of feedback in digital applications – does the feedback they provide support learning?* Manuscript submitted for publication.

Tärning, B., Haake, M., & Gulz, A. (2011). Off-task Engagement in a Teachable Agent based Math Game. Paper presented at the *19th International Conference on Computers in Education*. Chiang Mai, Thailand.

Tärning, B., Silvervarg, A., Gulz, A., & Haake, M. (2018). *Instructing a teachable agent with low or high self-efficacy – does similarity attract?!* Manuscript submitted for publication.

Uresti, J. A. R. (2000). Should I teach my computer peer? Some issues in teaching a learning companion. In *Proc. of International Conference on Intelligent Tutoring Systems* (pp. 103-112). Berlin/Heidelberg, Germany: Springer-Verlag.

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.

Veletsianos, G. (2012). How do learners respond to pedagogical agents that deliver social-oriented non-task messages? Impact on student learning, perceptions, and experiences. *Computers in Human Behavior*, *28*(1), 275-283.

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.

Paper IV

G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.) (2017). *Proc. of the 39th Annual Conference of the Cognitive Science Society* (pp. 1169-1174). London, UK, July 26-29, 2017.

Supporting Low-Performing Students by Manipulating Self-efficacy in Digital Tutees

Betty Tärning (betty.tarning@lucs.lu.se)
Lund University Cognitive Science, Lund University, 222 22 Lund, Sweden

Magnus Haake (magnus.haake@lucs.lu.se)
Lund University Cognitive Science, Lund University, 222 22 Lund, Sweden

Agneta Gulz (agneta.gulz@lucs.lu.se)
Lund University Cognitive Science, Lund University, 222 22 Lund, Sweden

Abstract

Educational software based on teachable agents has repeatedly proven to have positive effects on students' learning outcomes. The strongest effects have been shown for low-performers. A number of mechanisms have been proposed to explore this outcome, in particular mechanisms that involve attributions of social agency to teachable agents. Our study examined whether an expression of high versus low self-efficacy in a teachable agent would affect low-performing students with respect to their learning outcomes and with respect to a potential change in their own self-efficacy. The learning domain was mathematics, specifically the base-ten system. Results were that the learning outcomes of low-performers who taught a low self-efficacy agent were significantly better than the learning outcomes of low-performers who taught a high self-efficacy agent. There were no effects from the manipulation of self-efficacy expressed by the teachable agent on changes of the low-performing students' own self-efficacy.
Keywords: social agency; educational software; teachable agent; math self-efficacy; math performance

Introduction

A *teachable agent* (TA) is a graphical computer character in a tutee role. The basic idea is that *the student* instructs and guides the TA (Brophy, Biswas, Katzlberger, Bransford, & Schwartz, 1999). In essence, TA-based educational software implements the pedagogical approach *learning by teaching*, (Bargh & Schul, 1980).

To date a set of TA-based learning games targeting the STEM areas have been developed and evaluated, and repeatedly proven to have positive effects on students' learning outcomes. Some studies have compared effects of TA-based software with ordinary teaching (regular classroom practice) (Pareto, Haake, Lindström, Sjödén, & Gulz, 2012; Chin, Dohmen, & Schwartz, 2013). Others have compared educational software versions with and without a teachable agent included (Chase, Chin, Oppezzo, & Schwartz, 2009; Pareto, Schwartz, & Svensson, 2009).

An observation from several of the studies is how readily the metaphor of the computer figure as a tutee (*digital tutee*) is accepted by students. They express engagement for the task of teaching the character, although it is in fact nothing but a computer artifact (Chase et al., 2009; Lindström, Gulz, Haake, & Sjödén, 2011.) They also make more effort to learn in order to teach their digital tutee than to learn for themselves (Chase et al., 2009). In effect, students attribute mental states and responsibility to the digital tutee as if it were a social agent (Chase et al., 2009; Lindström et al., 2011). They see the agent as a socio-cognitive actor that can learn (respond to being taught by them) and that can be ascribed traits such as 'brave', 'slow', 'smart', 'forgetful' etc.

TA-systems and Low-Performing Students

Several studies show that the students who benefit most by educational software with teachable agents – whether compared to equivalent software without TA or compared to ordinary classroom teaching – are the low-performing students. When comparing eleven year olds who used an educational game in biology with or without TA, the former spent more time on learning activities and also learned more, with the effects most pronounced for lower performing students (Chase et al., 2009). In a study by Sjödén and Gulz (2015), 9-10 year-olds used a TA-based educational math game in school over a period of eight weeks. Thereafter, the students were divided into two groups, matched according to their pretest scores, and randomly assigned to a post-test with or without the TA present (the TA did not act in order to influence the test but was merely present). Results showed that low-performers (according to the pretest) improved significantly more than high-performers but only when tested with the TA. Pareto et al. (2009), likewise found a considerably stronger improvement for low ability students than for high ability students when they used a math game with a TA feature compared to the math game without the TA.

Mechanisms in TA-Systems that may Support Low-Performing Students

A number of explanations for the pedagogical power of TA-based games have been proposed, including some that also provide possible rationales for why the effect is often larger for low-performers.

First, in a TA-based game, the student is positioned as the one that is most able, the one who can teach someone else that knows less. This experience – being someone who is

capable, who knows more than someone else – can potentially affect a student's view on her own competence in a positive way. This will likely benefit low-performers more than high-performers, since the latter are more likely to already have experienced the role of 'teaching someone else' and 'knowing more'. High-performers are more likely than low-performers to spontaneously take a teacher role (or be assigned this role in class). Acting teacher can potentially strengthen the student's belief in her own capability in the domain in question, and this may in turn have effects on performance.

Second, a teachable agent can be a model of learning behaviors (Blair, Schwartz, Biswas, & Leelawong, 2007). A TA is often designed to model fruitful and productive student behaviors, such as being curious, asking questions, reasoning, being explicit about parts of 'knowledge'. It is, however, more likely that high-performing students already have such behaviors on their repertoire compared to low-performing students, and that the latter therefore are more helped by being inspired by productive learning behavior in a TA.

Thirdly and crucially a TA is *teachable.* More specifically a TA models someone who from the beginning has little or no knowledge but learns incrementally or step-by-step. In other words, a teachable agent (re)presents or models an incrementalist theory of competence in contrast to an entity theory of competence according to which some individuals are held to be gifted and others non-gifted. This latter view is quite common among students (Dweck, 2006). Specifically it holds in the domain of mathematics, where it has also been shown that teachers to a larger extent than for other subjects used terms such as 'talented' and 'not talented' (Rattan, Good, & Dweck, 2012). In principle both high- and low-performing students can have an entity view of competence, and potentially benefit from viewing competence (in this study competence in math) as something that can be changed with effort. However, it is more likely that low-achievers with an entity view of competence are trapped in a circle, where they don't think they are talented and see no meaning in making an effort; therefore make little effort; therefore don't achieve and thus confirm they are not talented. In other words, they create a self-fulfilling prophecy.

Fourthly, Chase et al. (2009), propose a mechanism named *ego-protective buffer*. In TA-system it is the TA that is tested for its knowledge. When the TA fails at a test, the failure or non-success does not come as close onto the student as when she takes a test herself. Even if students are aware that the TA's knowledge reflects how the TA has been taught by themselves, the responsibility for failing is not only theirs. Instead of bearing the full burden of a failure, the responsibility of failure can be shared between the TA and student. Even though this may benefit high-performers as well, low-performers are more used at failing at school and thus the *ego-protective buffer mechanism* may explain why in particular low-achieving students perform better when working with a TA.

In sum, there is a set of proposed mechanisms that may explain why low-performers benefit more than high-performers from using teachable agents. All mechanisms involve the tendency of students to attribute social characteristics and agency to the agent, and interact intellectually and socially with it. For instance, to view the TA as someone that it is possible to share a failure with; to view the TA as someone who can accomplish a task (or not), as someone whose knowledge is different from mine and that I can influence by teaching it; to view that TA as someone that can learn – and as learner be slow, quick, smart, forgetful, etc.

In view of the above, we found it plausible that students would also tend to attribute *high or low self-efficacy* to an agent, if designed in an adequate manner. Spelled out, they would tend to attribute to an agent high or low belief in its own capability to learn and be successful – in our case with respect to math and base ten problems. The present study thus approaches the trait of self-efficacy, which to our knowledge has not been studied before in teachable agents.

Does TA Self-Efficacy Matter for Student Progress

Having an ability to learn, i.e. being *teachable*, is the very essence of a digital tutee or teachable agent. However, whether other kinds of properties are attributed to a TA depends in the first place on how the TA is designed and implemented, and also on the student interacting with the TA. For instance, depending on how it is implemented, a TA can be (perceived as) a quick learner or a learner that needs many rehearsals. A TA can be (perceived as) more or less challenging or questioning (Kirkegaard, 2016).

In our study the TA was designed to express either high or low belief in its own capacity to learn and perform in a math game. We will soon present our predictions but first discuss the phenomenon of *self-efficacy* in real human students. For human learners we know that there is a relation between self-efficacy and actual performance (Bandura, 1997) in that self-efficacy predicts subsequent performance. Low self-efficacy predicts low performance, and high self-efficacy predicts high performance. Proposed mechanisms are that student's self-efficacy influences how much effort she puts into a task, her tendency to persist, how high she sets her aspirations and her tendency to persevere when being challenged by the task. Individuals with high self-efficacy often achieve more in intellectual terms (Bandura, 1997). Importantly, however, the relations are correlational and on a group level. There are no causal or absolute relations between individual's self-efficacy and her performance; students may over-estimate as well as under-estimate their own capacity.

We now return to self-efficacy in teachable agents. The central research question in the present study was whether a teachable agent expressing low or high self-efficacy,

respectively, would have different impact on low-performing students in terms of their learning and progress. In addition we explored whether there would be any effects on students' own self-efficacy in either of the conditions.

Research Questions and Predictions

Research Question 1 (RQ1) Will learning and progress differ between low-performing students who teach a TA expressing low self-efficacy (**lowSE-TA**) and low-performing students who teach a TA expressing high self-efficacy (**highSE-TA**)?

As a basis for our predictions we used two different theories: (i) role-modeling theory by Bandura (1977) and (ii) the theory of the TA protégée effect by Chase et al. (2009). This resulted in two alternative predictions that point in opposite directions. As such this is not surprising since the predictions are generated from theories not related to one another.

The first, alternative, prediction in line with Bandura´s idea of role modeling focuses on teachable agents as behavioral models, as discussed in the introduction. A **highSE-TA** models a learner with a strong belief in her own abilities to learn, a willingness to persist and not give up, etc. Together with the TA:s incremental progression (given that it is reasonably taught by the student) this is likely to be a positive model for low-performers, that often themselves have low self-efficacy. Thus we predict that low-performers will make larger progress if they teach a **highSE-TA** than if they teach a **lowSE-TA**.

The second, alternative, prediction is based on the protégée-effect mentioned above: in general, students seem to take responsibility for a TA and make an effort to teach it. Now, a **lowSE-TA** expresses uncertainty in its own capacity, and seems in considerable need for support and engagement from the teacher (i.e. student), whereas a **highSE-TA** expresses confidence in its own capability to learn and manage and seems in less need for help from the teacher. Therefore low-performers may be more motivated to take responsibility and make an effort to teach a **lowSE-TA** compared to a **highSE-TA**. Consequently they will also themselves make more progress. Thus we predict that low-performers will make larger progress if they teach a **lowSE-TA** than if they teach a **highSE-TA**.

There is also third possible result, namely that whether the TA expresses low or high self-efficacy will not matter for low-performers progress.

Research Question 2 (RQ2) Will a potential change in self-efficacy in low-performing students differ between those students who teach a TA expressing low self-efficacy and those who teach a TA expressing high self-efficacy?

If the TA functions as a behavioral model with respect to self-efficacy, low-performers are more likely to increase their own self-efficacy if they teach a **highSE-TA** than if they teach a **lowSE-TA**. The reason is that they may be inspired to model the TA along the line "*If this character, my digital tutee, believes strongly in its capability, why shouldn't its teacher, that is me, do so too?*"

From the protégée effect no straightforward prediction can be derived on potential self-efficacy change in students, depending on TA self-efficacy. As discussed under RQ1, if the protégée effect is at work, participants will put particularly large effort into teaching a **lowSE-TA**, since such a TA signals a greater need of help and support than a **highSE-TA** that signals that can learn on its own. But whether students that take more responsibility and make a larger effort to teach their TA also change their belief in their own capacity to learn is not obvious. On the one hand, an interplay between performance and self-efficacy is likely but such influences may take time.

Again there is a third possible result, namely that whether the TA expresses low or high self-efficacy does not matter with respect to low-performers potential self-efficacy change.

To sum up, the present study made use of a learning game in math including a TA, where we manipulated the TA:s expressed belief in its own capability to perform and learn math as expected in the game. Our two research questions were: RQ1: Would the manipulation of TA self-efficacy have an effect on low-performing students' progress in the game (i.e. their learning math)? RQ2: Would the manipulation of TA self-efficacy have an effect on potential change in self-efficacy in the low-performing students?

Method

Participants

Participants were 166 students (83 girls and 83 boys) aged 10-11 years from 4 schools and 9 classes in Southern Sweden from areas with relatively low socio-economic status and school performance below average. Students were randomly assigned one of the conditions: teaching a digital tutee that expressed high self-efficacy (**highSE-TA**) or teaching a digital tutee that expressed low self-efficacy (**lowSE-TA**). Out of the initial set of participants, 24 were excluded due to missing data points or low attendance. Next, out of the 142 remaining students, the 62 students who performed below the median on a math performance test were selected for further analysis. The math test was based on a representative part of the national tests in mathematics and consisted of 21 problems relating to place value. Thus, in the final data set, there were 28 students in the **lowSE-TA** condition and 34 in the **highSE-TA** condition.

The Educational Game

The TA math game, developed by Lena Pareto (Pareto, 2014), targets basic arithmetic skills related to the place value system, where the student teaches a digital tutee

named Lo, so that Lo can compete against other students' digital tutees or against a computer actor in different digital board games. Lo's knowledge – based on the system's knowledge domain (Pareto, 2014) – develops entirely on the basis of what the student teaches her (and if taught wrong, Lo will learn wrong).

A central part of the student's teaching consists of answering questions from the digital tutee about the math content, specifically regarding place value, via multiple-choice for answering (see figure 1). The other main interaction between student and digital tutee takes place via a free text chat (Silvervarg & Jönsson, 2011). This is also where Lo, the TA, expresses her self-efficacy (see figure 1).

Figure 1: The math game with multiple choice conversation and 'free text chat' conversation (overlay).

Self-Efficacy in the Teachable Agent

High or low self-efficacy in or study was defined as high or low belief in ones capability to make progress and perform well in the math game. In turn, this requires making adequate moves and answering questions regarding the place value system correctly. The definition can be compared to a more general definition of self-efficacy in mathematics as the belief in ones capability to successfully learn mathematics (Bandura, 1997).

After each round of the game where Lo (the TA) has been active – observing and posing questions to the student or being guided by student – the chat conversation starts. The chat begins with Lo commenting on the previous round saying for example: "*Awesome! We won! I have a good grip now of tens and hundreds and all that you teach me.*" (reflecting high self-efficacy), "*Oh I won, did I? Nice. But I feel very uncertain about how to play well.*" (reflecting low self-efficacy).

The chat conversation also contains other comments and reflections from Lo on her own learning, for instance: "*I'm learning the rules in the math game slowly. I'm not a very brilliant student.*" (reflecting low self-efficacy), "*It's going to get better and better. I have so quickly learned so many things about how to play the game.*" (expressing high self-efficacy), and "*I am not sure I can learn these things.*" (expressing low self-efficacy).

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

Lo's utterances had previously been evaluated with regard to whether they sounded as uttered by someone who was confident, not confident, or neither nor in her ability to learn and perform. The evaluators were 22 fourth graders from a school not participating in the study. The evaluation resulted in the removal of a few sentences and slight modifications of others, resulting in a set of 136 sentences, 68 reflecting a digital tutee with high self-efficacy and 68 reflecting a digital tutee with low self-efficacy.

In addition the manipulation – low and high self-efficacy in the TA – was validated within the present study by participating students. At the end of the last study session they were asked to evaluate Lo's belief in her/his own capability to play the math game on a Likert scale. A Mann-Whitney test showed a significant difference ($Z = -4.85$, $p < .001$, $r = .39$) between the low SE-TA and the high SE-TA, confirming that the manipulation had intended effects on the perception of the TAs self-efficacy.

Procedure

All study sessions took place in ordinary classrooms and lasted about 30 minutes. At the pre-test session, students completed a math pre-test targeting the place value system, and a pre-questionnaire targeting their self-efficacy in math with respect to the place value system. The students' math pre-test scores were used to identify the target group for this study's research questions, i.e. low-performers (in math).

Thereafter students participated during seven game-playing sessions, once a week. At the post-session, students again filled out the questionnaire targeting their self-efficacy in math and the place value system and were debriefed about the two different types of digital tutees and the purpose of the study.

Measurements

Performance During Game Play Students' performance while teaching the digital tutee is a reflection on how well they perform themselves. In line with this we calculated a performance score for each student on the basis of the data-logging. Through the game the digital tutee poses questions to the student that concerns the conceptual model and principles of the place value system. For instance: "*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?*" The tutee

posed three such questions during each game session, and the student had to choose one out of four alternative answers (one correct, two incorrect and the alternative "*I don't know.*"). The performance score was calculated as the percentage of correct answers minus the percentage of incorrect answers. Additionally, a study by Pareto (2014) showed that in-game performance in this math game correlated with standard paper-and-pencil tests on the place-value system.

Self-Efficacy Change To measure this we used a self-efficacy pre- and post-questionnaire based on Bandura, Barbaranelli, Caprara, and Pastorelli (1996); for this study translated into Swedish

The seven items targeted the students' self-efficacy with regard to the place value system and the question "*How good are you at solving this type of task?*" Item one to five regarded calculation tasks such as "*1136 + 346*", and item six and seven targeted place value concepts, such as: "*Which digit has the highest place value in the number 6275?*" All items were graded in five steps from "*Not good at all*" to "*Very good at*".

Results

Statistical analyses were conducted in R v3.2.4 (R Core Team, 2016). Of the 142 participants with complete data, the 62 performing below the median on the pre-test in math were included in the analysis.

Effects TA Self-Efficacy on Low-Performing Students' Performance During Game Play

An unmatched two sample *t*-test showed a significant difference ($t(60) = 3.40$, $p = .0012$, Cohen's $d = 0.87$) of TA self-efficacy on student performance with the students in the **lowSE-TA** condition ($M = 54.8$, $SD = 13.7$) outperforming the students in the **highSE-TA** condition ($M = 43.7$, $SD = 12.0$).

Effects of TA Self-Efficacy on Low-Performing Students' Self-Efficacy Change

An unmatched two sample *t*-test showed no significant difference ($t(60) = 0.35$, $p = .73$) of TA self-efficacy on student self-efficacy change between the students in the **lowSE-TA** condition ($M = 1.18$, $SD = 3.81$) and the students in the **highSE-TA** condition ($M = 1.53$, $SD = 4.00$).

Discussion

Teaching a **lowSE-TA** compared to teaching a **highSE-TA** made the participants perform significantly better, as measured by their in-game performance scores. But the two conditions did not differ with respect to whether the participants changed their own self-efficacy. Changes were small and did not differ between the conditions.

These results contribute to our knowledge about mechanisms in a TA-based educational game with respect to why low-performers tend to benefit more than high-performers from these games. First, we showed that a manipulation of expressed self-efficacy in a TA can influence performance for low-performers: a TA that expressed low self-efficacy was more beneficial than a TA that expressed high self-efficacy. The effect as such, regardless of direction, confirms that at least some of the pedagogical power in a TA-based game derives from attributions of social agency to TA:s, in this case attributing to the TA a weak or strong belief in its own capability. Consequently this is one of the traits that a TA designer ought to be aware of; a trait that can explain why low-performers benefit more than high-performers from TA-based games.

With respect to student performance, we based our predictions on two different theoretical models: role modeling according to which a **highSE-TA** should have the most positive influence on the performance of low-performers, and the protégée effect according to which a **lowSE-TA** should have the most positive influence on the low-performers performance. The latter theory was supported and can be further elaborated on by means of the results of our study. According to the protégée-effect students tend to make more effort and take more responsibility for the task of teaching a TA than for the task of learning for themselves (Chase et al., 2009). In our study the outcome was better when low-performers taught a **lowSE-TA** compared to a **highSE-TA**. It is near at hand that they made an even larger effort and took even more responsibility for a TA with low self-efficacy since this TA expresses a low trust in her own ability to learn, and likely comes across as someone who is more in need of help than a TA with high self-efficacy. A **highSE-TA**, on the other hand, indicates that s/he is capable to learn and perform, and is in less need of help.

The lacking effect on students self-efficacy change, depending on high or low self-efficacy in the TA, means that the role-modeling hypothesis proposed above was not supported. Students were not inspired by a **highSE-TA** as a model to increase their own self-efficacy. Neither did teaching a **lowSE-TA** lead to an increase in the students' self-efficacy. However, it did lead to an increase in their performance, and we can thus conclude that the increased performance was not caused by an increased self-efficacy, at least not as measured in our study. It should also be pointed out that an increase in self-efficacy is not always desirable, in particular not for students who overestimate their capabilities. At the same time, given the interactions between self-efficacy and performance, it is often a good thing when students with low self-efficacy in a domain gain more confidence in their abilities to make progress. What is desirable *in general* is that as many students as possible have an incrementalist rather than an entity view of intellectual capabilities – something that the use of TA-

based educational games may contribute to (Chase et al., 2009).

Limitations of the Study and Future Research

The study should be seen as a first examination about how the manipulation of self-efficacy in a digital tutee can influence student performance. Some limitations should be kept in mind when interpreting the results. One is that there was no group of students who taught a digital tutee that expressed a neutral mode of self-efficacy. In future research such a condition should be included. Furthermore, rather than aiming to be conclusive, the present study opens up for associated studies. For instance, one relevant question is whether the results will replicate or not with other age groups than 10-11 year olds. Another interesting line of research could be to explore a TA with adaptive self-efficacy that reflects the rate at which it actually learns, which in turn reflects the proficiency of the student that is teaching it.

Acknowledgments

This research was funded in part by Marcus and Amalia Wallenberg Foundation and the research environment Cognition, Communication and Learning. The authors thank all students and teachers who participated in the study.

References

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

Bandura, A. (1977). *Social learning theory*. Englewood Cliffs, NJ: Prentice-Hall.

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.

Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. *Journal of Educational Psychology*, *72*(5), 593–604.

Blair, K., Schwartz, D., Biswas, G., & Leelawong, K. (2007). Pedagogical agents for learning by teaching: Teachable agents. *Educational Technology*, *47*(1), 56–61.

Brophy, S., Biswas, G., Katzlberger, T., Bransford, J., & Schwartz, D. (1999). Teachable agents: Combining insights from learning theory and computer science. In S.P. Lajoie & M. Vivet (Eds.), *Frontiers in Artificial Intelligence and Applications, Vol 50. Proc. of AIED 1999*. Amsterdam, The Netherlands: IOS Press.

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

Chin, D. B., Dohmen, I. M., & Schwartz, D. L. (2013). Young children can learn scientific reasoning with teachable agents. *IEEE Transactions on Learning Technologies*, *6*(3), 248–257.

Dweck, C. (2006). *Mindset: The new psychology of success*. Random House.

Kirkegaard, C. (2016). *Adding challenge to a teachable agent in a virtual learning environment*. Licentiate Thesis in Cognitive Science, Linköping University. Linköping, Sweden: Linköping University Electronic Press.

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

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., Schwartz, D. L., & Svensson, L. (2009, July). Learning by guiding a teachable agent to play an educational game. In V. Dimitrova, R. Mizoguchi, B. du Boulay, A. C. Graesser (Eds.), *Frontiers in Artificial Intelligence and Applications, Vol 200. Proc. of AIED 2009* (pp. 662–664). Amsterdam, The Netherlands: IOS Press.

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*, 723–751.

Rattan, A., Good, C., & Dweck, C. (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.

Silvervarg, A., & Jönsson, A. (2011). Subjective and Objective Evaluation of Conversational Agents. In *Proceedings of the 7th Workshop on Knowledge and Reasoning in Practical Dialogue Systems* (pp. 65–72). Barcelona, Spain.

Sjödén, B., & Gulz, A. (2015). From Learning Companions to Testing Companions. In C. Conati, N. Heffernan, A. Mitrovic, & M.F. Verdejo (Eds.), *LNAI/LNCS: Vol. 9112. Proc. of AIED 2015* (pp. 459–469). Berlin/Heidelberg, Germany: Springer-Verlag.

R Core Team (2016). *R: A language and environment for statistical computing* [Software]. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org/.

Paper V

Looking into the black box of students' (not) handling feedback on mistakes

Since the publication of this thesis, a revised version of this article is published as:
Tärning, B., Joo Lee, Y., Andersson, R., Månsson, K., Gulz, A., & Haake, M. (2020). Assessing the black box of feedback neglect in a digital educational game for elementary school. *Journal of the Learning Sciences*, 1-39. Early online: 2 July 2020.
https://doi.org/10.1080/10508406.2020.1770092

Betty Tärning[1], Yeon Joo Lee[1], Richard Andersson[1], Kristian Månsson[1], Agneta Gulz[1], & Magnus Haake[1]

[1] *Div. of Cognitive Science, Lund University, Lund, Sweden*

Abstract. Previous research has shown that critical constructive feedback that scaffolds students to identify mistakes and to improve on a task often remains untapped in practice in the sense of not being used to revise and improve. Our aim was to illuminate where, in a feedback chain, students are lost: in noticing the feedback, processing it, or in attempting to use it in revising the task? The steps of noticing and processing were measured via eye-tracking. After that, behavioural data logging tracked students' actions.

As expected, few students reached the last step of making progress on the task. With our methodology we could also follow the preceding steps, where students were successively lost.

We also experimentally compared three different feedback framing conditions: signalling via a pedagogical agent, via an animated arrow, or no signalling (control condition). The pedagogical agent condition in comparison led to a significantly lower rate of feedback neglect in both the noticing and the reading step. Thus, it appears possible to influence students who are not initially inclined to notice and read feedback text into doing so. Future work will address how they can also be scaffolded to act according to the feedback.

Keywords: critical constructive feedback, feedback neglect, feedback processing model, learning, eye-tracking, data logging.

Introduction

It is clear that feedback is a central aspect of an educational context. Review studies by Hattie and Timperley (2007) and Shute (2008) show that feedback can help students to achieve their learning goals. In particular, feedback that aims at scaffolding students to identify errors and mistakes (hence, *critical feedback*) and to improve and make progress (hence, *constructive feedback*), is an important ingredient in teaching and learning. When a student provides a non-adequate solution to a problem, information that signals that the solution is inaccurate (critical feedback) and provides guidance for how to proceed (constructive feedback), can have a considerable impact on the student's continued learning and performance. This combination of critical feedback and constructive feedback, that we will term *critical constructive feedback* can be the support needed to make a student understand what went wrong and why, guiding her actions so that she improves her performance.

Black and Wiliam (1998) conducted a well-renowned meta-study, including 250 studies that, taken together, pointed to a considerable impact of formative feedback on learning. Not all reviewed studies encompassed critical constructive feedback, but some of the highlighted examples did. For instance, Elawar and Corno (1985) studied teachers that were educated in giving their students written critical constructive feedback (i.e. comments and suggestions on how to improve an error plus one positive remark). Results were that when given critical constructive feedback, students' achievement improved compared to students who did not receive this type of feedback. In another study, Butler (1987) compared four different feedback conditions (individual comments, praise, grade, and no feedback) and found superior performance for the student who received individual comments.

Even though it has been shown that students can gain from using critical constructive feedback, a set of basic conditions has to be fulfilled if a student should, in effect, profit from receiving such feedback. First, it must be possible for the student to understand the feedback (Lea & Street, 1998; Higgins, Hartley, & Skelton, 2001; Orsmond, Merry, & Reiling, 2005) and the amount of feedback provided must be reasonable. If the sheer amount is overwhelming, it will not help that the feedback is potentially comprehensible, useful, and supportive (Brockbank & McGill, 1998).

Second, it must be possible for the student to make the connection between the feedback received and what it can be used for. In other words, the student must be able to see the point of the feedback received (Orsmond et al., 2005; Wiliam, 2007; Segedy, Kinnebrew, & Biswas, 2013).

Student neglect of critical constructive feedback

Even if the two conditions described above are met, i.e. the critical constructive feedback (henceforth CCF) provided to students is both comprehensible and meaningful (in the sense that dealing with the CCF increases their chances to solve their tasks), there are abundant reports of neglect of CCF (Hounsell, 1987; Wotjas, 1998; Perrenoud, 1998; Clarebout & Elen, 2008; Conati, Jaques, & Muir, 2013). In practice this means that in spite of the amount of time teachers spend on providing CCF to their students, as well as the efforts from designers of educational software to provide CCF to students, the potential student gains to be had from the feedback frequently remain untapped. Note that 'neglect' in this sense does not necessarily involve intentionality. It is a broader term designating that students fall off at some point in the (ideal) chain of noticing feedback, processing feedback, and making use of feedback.

An illustrative case is presented in Segedy, Kinnebrew, and Biswas (2012). In a previous study the authors had realized that 77% of the CCF statements delivered in an educational science learning game for 10- to 11-year-olds seemed to be ignored by the students (although the reasons for this could not be separated in the study). For the new study (Segedy, Kinnebrew, & Biswas, 2012), the researchers refined the CCF so that it aligned clearly with the students' current goals within the game and provided useful explanations and examples. In spite of this, a large proportion of the feedback was still ignored by the students. The explanation proposed by the authors was that even though the feedback provided was potentially both comprehensible and useful for them, students were unwilling to deal with the feedback and therefore neglected it.

Again, unwillingness in this sense does not have to be intentional but can involve an intuitive way of responding. Even if it is constructive and potentially helpful for the student to get on with a task, the presence of CCF still points at the fact that the student failed at her previous attempt. When failure implies an ego threat, it may be rational to shut down rather than heed CCF in order to avoid a feeling of inadequacy. Indeed, it has recurrently been shown that students often ignore CCF because they interpret it as evaluative punishment (Hattie & Timperly, 2007). This area of research has, however, not addressed when neglect or avoidance occurs; does it occur already at the step of noticing the CCF, in the step of processing (e.g. reading) the CCF, or in the step of attempting to act upon the CCF (i.e. attempt to follow the hints and instructions provided) – and can this be investigated? These are the topics of our study.

Little is known about the process of neglect of critical constructive feedback

To recapitulate, even in situations where CCF is *potentially* comprehensible, accessible, and useful for students, many still neglect it in the sense of falling off at some point in the feedback chain (Hounsell, 1987; Wotjas, 1998; Perrenoud, 1998; Clarebout & Elen, 2008; Conati et al., 2013; Segedy et al., 2012). However, most previous studies have identified neglect by identifying students not reaching the last step in the feedback chain, i.e. using CCF to make progress in their task. These studies cannot tell at what point in the feedback chain the students may have fallen off; already at the point of noticing the CCF, at the point of processing (reading) the CCF, or at the point of making use of the CCF.

Likewise, many studies on feedback focus on the *type of feedback* provided in relation to *whether the learner makes progress or not* (performs better on a task, makes progress in a pre/post-test setup, etc.) (Kluger & DeNisi, 1996; Black & Wiliam, 1998; Mory, 2004 & Shute, 2008). In other words, the measurements concern the very last step in a feedback chain with the steps in-between treated as black boxes.

The goal of our study was to investigate *where in the feedback chain* neglect of CCF occurs and also whether different *signalling conditions* with respect to the presentation of CCF can influence neglect. Thus, in terms of method, our goal was to show that it is possible to study CCF-neglect throughout a chain of noticing, processing, and making use of feedback. To address this goal, we developed a CCF-processing model that we used for describing and analysing the behaviour of 11- to 12-year-olds that played a digital educational game in which CCF was provided. Three signalling conditions with respect to the CCF were part of the educational game, and potential differential effects of these were also studied.

The digital educational game, targeting history for 11- to 12-year-olds, was used for the presentation of the experimental signalling conditions, and as an instrument for the collection (logging) of behavioural data. The game (see also section "The Digital Learning Environment", p. 19) has been developed by our research group (Educational Technology Group; https://www.lucs.lu.se/etg/), and the reason for choosing our own educational platform was twofold. First, we were able to design and manipulate the tasks, the feedback texts, and the visual signalling for the study. Second, the game has previously been used in classroom settings in other studies (Kirkegaard, 2016; Silvervarg, Kirkegaard, Nirme, Haake & Gulz, 2014; Kirkegaard, Gulz & Silvervarg, 2014). Thus, we knew it could be used for regular lessons in history, which was important for our aim to obtain ecological validity.

Modelling the feedback process

For our development of a CCF-processing model, we found initial inspiration in a paper by Timms, DeVelle, and Lay (2016), that called out for more knowledge about what happens when a learner has made an error and then (potentially) notices, processes, understands, and makes use of feedback related to the error provided by the learning environment. Timms et al. (2016) present a model based on a cognitive psychology and neuroscience perspective with the following steps (figure 1): (i) the learner detects (or does not detect) an error that she has made, (ii) the learning environment detects the error and provides feedback related to the error, (iii) the learner notices (or not) the feedback provided by the learning environment, (iv) the learner decodes the feedback (or not), (v) the learner makes sense of the feedback (or not), and finally (vi) the learner corrects the error and a learning event occurs.

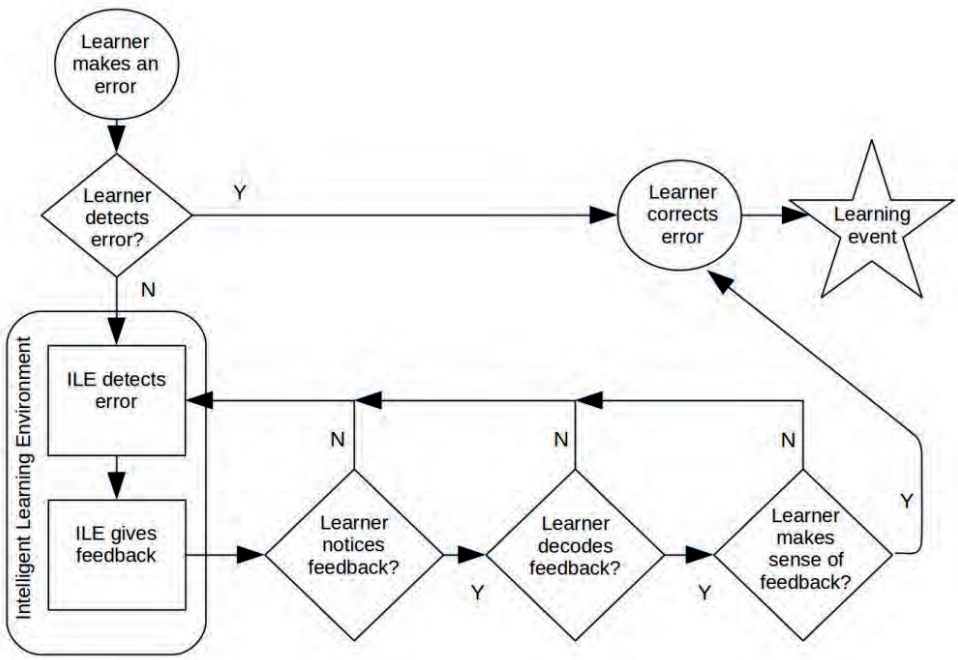


Figure 1. The model of feedback processing proposed by Timms et al. (2016).

Timms et al. (2016) also discuss how the steps could be measured and suggest eye-tracking for measuring whether or not the student notices the feedback and EEG as well as fMRI for measuring the 'decoding' and 'make sense' steps.

The CCF-processing model and its use in the study

For the purpose of our study targeting critical constructive feedback we designed a model, see figure 2, that has similarities with that of Timms et al. (2016), but also significant differences.

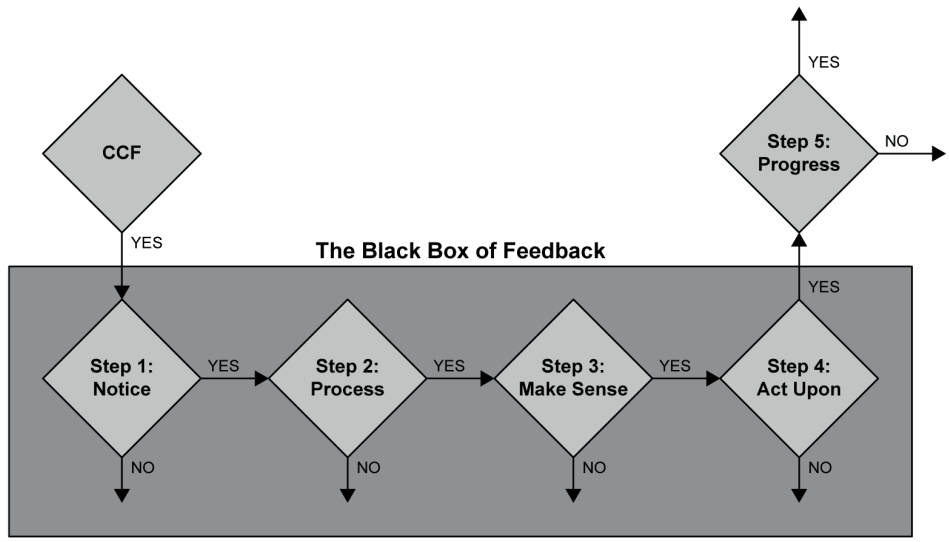


Figure 2. The CCF-processing model developed to study feedback used in this study: 1 (notice), 2 (process), 3 (make sense), 4 (act upon), and 5 (make progress). (*NB: Step 2 'process' corresponds to Timms et al.'s (2016) step 'decode'.*)

In brief the differences between the models are: (i) our model starts after the stage where a student receives CCF from the learning environment; it does not include the case where a student on her own detects and then (automatically) corrects the mistake; (ii) we use the term 'process feedback' for the step that Timms et al. (2016) call 'decode feedback'; (iii) in our model 'make sense of the feedback' is followed by the step 'act upon the feedback' and thereafter the final step 'make progress with the task', whereas 'make sense of the feedback' in Timms et al.'s (2016) model is directly followed by the final step 'correct the error'. We now turn to explicating the reasons for these differences.

The situation when a student detects a mistake on her own (without receiving any CCF) and then immediately knows how to correct the mistake is not applicable to the kinds of tasks provided to the students in our study. If it is a true mistake made by the student (and not what Norman (2013) calls "a slip", i.e. the person actually knows what is right but happens to 'slip') there is no possibility that she can correct it (except via brute and cumbersome trial-and-error) without guidance. In

other words, the 'critical' part of CCF, i.e. telling the student that the answer was not correct, will not as such make her know what the correct answer is. This applies to all tasks of a certain complexity, where it is not the case that knowing that an answer is not the right one per se means a possibility to bring up the correct answer. For this reason, we have left out this part of the model by Timms et al. (2016).

Both models include the step 'notice feedback'. For the step following upon 'notice feedback' we have chosen the term 'process feedback', whereas Timms et al. (2016) uses 'decode feedback'. With our choice of term, we wish to emphasize that this can be a rather complex step, e.g. a capacity to decode letters is not sufficient for being able to read. The two research groups also have different views on how this step could be measured. According to Timms et al. (2016), the 'decode' step is hard to capture with behavioural data measurements and they instead suggests EEG or fMRI, whereas our study makes use of eye-tracking.

The subsequent step is in both models termed 'make sense of feedback'. In the model of Timms et al. (2016), however, this step is directly followed by the step 'correct the error'. In other words, if the learner has made sense of the feedback she will per se correct the error. Our model has a middle step between 'make sense of feedback' and 'make progress with the task' (which corresponds to Timms et al.'s (2016) 'correct the error'), namely that of 'acting upon the feedback'. By this we address the step of 'making use of the CCF by attempting to follow the hints and instructions provided'.

This too, is a step where a student may or may not fall off. Even if a student has made sense of the CCF, she can refrain from acting upon it. In such a case, i.e. when the student does not attempt to use the hints and instructions provided in the CCF, she has small chances of correcting her mistake. On the other hand, if she does attempt to use the hints and instructions, she considerably increases her chances of correcting the mistake. Yet, even when hints and instructions are understood, they can be hard to follow, and an 'attempt to use hints and instructions' does not per se equal 'succeeding in using them to correct the mistake'. This is why our model has the dual outcome of 'yes' as well as 'no' at Step 4 of 'acting upon CCF' and at Step 5 of 'making progress'; contrasting the model of Timms et al. (2016) with a single self-fulfilling step of 'correcting the mistake'.

To illustrate by an example from the educational game used in our study. A student gets CCF that informs her it is not correct that Émilie du Châtelet criticized (Isaac Newton's) Principia and suggests that she should re-examine the book in du Châtelet's room. Even if the student makes sense of the suggestion, she may or may not actually act upon it and visit Châtelet to click on the book and read the text provided. Secondly, if the student does follow the suggestion (looks

for, finds, and reads the information needed to solve the task) she still may or may not be able to apply this information in order to 'make progress' (correct her former mistake and improve on the task).

The reason that the latter steps in Timms et al.'s (2016) model are fewer and of less complexity than in our model possibly reflects the lesser complexity of information brought to the students in their study. With tasks of less complexity the correction of a mistake, as soon as one is aware that there is one, is more straightforward. Yet another possible difference lies in Timms et al.'s (2016) focus on neuro-processes versus our focus on behaviour.

CCF-neglects in the CCF-processing model

According to our model (figure 2) CCF-neglect can appear at any of the following steps:

1. First, the student may not notice the CCF; in our study meaning that she does not notice the CCF-text on the screen.
2. Second, the student may not process the CCF; in our case meaning that she does not read the CCF-text that she has noticed.
3. The student may not make sense of the CCF; in our case meaning that she reads the CCF-text, but does not comprehend the meaning of the CCF-text.
4. The student may not act upon the CCF that she has read and comprehended; in our case meaning that she does not attempt to make use of the hints or instructions that she has read and understood.
5. The student, finally, may not make progress and solve the task at her next attempt – even if she has noticed, read, understood, and acted upon the CCF.

In our study of the different steps in the process of potential CCF-neglect, we also asked whether or not the presentational framing of the CCF would make a difference for the respective steps. All students were presented with two different forms of so called 'visual signalling' (a pointing arrow and a pointing and gazing pedagogical agent, respectively) as well as a control condition without any visual signalling. The research question was: "*Would visual signalling in one or both of these forms influence feedback neglect with respect to the different steps in the process?*" By studying these three different conditions, we would shed more light on the process of CCF-neglect than if we had studied only one condition. Previous research (discussed below) on the role of signalling for attention and learning indicates that signalling can be a modulating factor. Arrows, as well as gaze and pointing, are signalling methods that have already been studied in other, partly related, contexts.

Research questions

Referring to our CCF-processing model (figure 2) we addressed the possibility of opening the black box of feedback to study the entire chain of events and the extent of CCF-neglect in the different steps.

Research Question I

To what extent will students:

i) Neglect the CCF in the sense of not noticing it?
ii) Neglect the CCF in the sense of noticing but not processing it (in our case, notice but not read the text in the feedback boxes)?
iii) Neglect the CCF they noticed and read in the sense of not attempting to act upon it?
iv) Not be helped by CCF they attempted to act upon for improving their next result on the task?

Research Questions II

Will the data, with respect to Research Questions I (i–iv), differ between the three conditions: (i) a virtual agent looking/gazing towards and pointing to the CCF-text, (ii) a dynamical arrow pointing to the CCF-text, and (iii) the CCF-text as such (control condition)?

Related previous studies

In approaching neglect vs. uptake of *constructive critical feedback* from an *information processing perspective* (figure 2) the present study is, to our knowledge, a pioneer study. Yet there are two areas of previous research that are related our Research Questions II. First, there are studies in the area of 'visual signalling' in an educational context that have specifically addressed effects on learning outcomes comparing pointing arrows and pointing pedagogical agents and no signalling control condition. Second, studies addressing 'gaze cuing' and 'social attention' have investigated effects by human gaze compared to a pointing arrow on observer's attention and gaze direction.

In this section, we discuss previous results from these two research areas, and conclude by discussing how the studies differ from our study.

Signalling using a pointing pedagogical agent vs. a pointing arrow

A set of studies on students using a digital learning environment have addressed the two visual signalling conditions of pointing pedagogical agents and pointing arrows. None of the studies, however, focused on how students handle (or neglect) *feedback* depending on the visual signalling conditions. Instead they studied how visual signalling conditions influence students' handling (or neglect) of 'relevant information' in general – as measured via 'performance' or 'learning outcome'.

The results from the studies are mixed, some showing no effects on learning outcomes of neither signalling pedagogical agents nor arrows compared to non-signalling control conditions (Ozugul, Reisslein, & Johnson, 2011, experiment 2; Van Mulken, Andre, & Muller, 1998). Other studies present positive effects on learning outcomes from pedagogical agent as well as arrows compared to control conditions (Moreno, Reisslein, & Ozogul, 2010; Ozogul Reisslein, & Johnson, 2011, experiment 1). Yet other studies have showed positive effects on learning outcomes from signalling pedagogical agents compared to both signalling arrows and control conditions – but only for students with low prior knowledge of the learning domain (Choi & Clark, 2006; Johnson, Ozogul, Moreno, & Reisslein, 2013; Johnson, Ozogul, & Reisslein, 2015).

Taken together, in the context of a digital learning environment, a pointing pedagogical agent or a pointing arrow, used for guiding or highlighting certain information, seem to have varied effects on learning outcomes.

Concerning the diverging results on learning effects from signalling agents, Veletsianos (2007) suggests that a pedagogical agent's contextual relevance – or non-relevance – is key to its potential effects on students' learning outcomes. The agent has to be relevant in the context, e.g. it has to make sense that it is there and points at things, for a positive effect on students' learning to occur. A study by Veletsianos (2010) further supported this claim.

Zooming in on the question why, in some studies, a *pointing arrow* turns out to have less positive effects on learning outcomes than a *pointing agent*, a number of different answers have been proposed. A first answer (relevant primarily for the second step of our CCF-processing model) is that an agent makes the purpose of visual signalling more explicit than an arrow. Students are accustomed to teachers' pointing gestures in a learning situation, and it might therefore be easier to comprehend the intentions of an agent's pointing gestures than those of a symbolic dynamic arrow. In other words, one may grasp more clearly that the agent's pointing gestures are intended to guide attention to relevant areas of the visual display, whereas the purpose of the arrow may be more ambiguous (Koning & Tabbers, 2013).

A second answer (connecting to both the first and second step of our CCF-processing model) grounded in developmental psychology and neuropsychology, is that humans give priority to *social stimuli* (Gamé, Carchon, & Vital-Durand, 2003; Pinsk et al., 2009; Taylor, Wigget, & Downing, 2007). Therefore, they may follow an agent more attentively than an arrow, as well as prioritize what an agent points and looks at more than what an arrow points at. Relatedly, Mayer and DaPra (2012) showed that a fully embodied agent – using gestures, facial expressions, and eye gaze – led to better learning outcomes than the same agent without such embodied actions. The authors propose a social agency theory for explaining the results, suggesting that social cues in the form of gestures, gaze, facial expressions "[...] prime a feeling of social partnership in the learner, which leads to deeper cognitive processing during learning, and results in a more meaningful learning outcome as reflected in transfer test performance." (Mayer & DaPra, 2012, p. 239).

Third, it has been suggested that it is the persona effect (Lester et al., 1997) – the visual presence of an agent as such, irrespective of its pointing and/or gazing – that makes a student more motivated to focus on information and work with a learning material and therefore results in increased learning outcomes. For instance, Johnson et al. (2013) saw a significant benefit for students with low prior knowledge when exposed to visual signalling by an agent, but not when exposed to an arrow. On the basis of this, they proposed that a combination of an agent's action (pointing) and its social presence, i.e. a persona effect, enhanced the participants' motivation towards the learning task. Another study (Johnson et al., 2015) involved a comparison between a pointing agent and a non-pointing agent. This time, results supported a beneficial impact on learning outcomes for students with low prior knowledge for the pointing agent only.

Effects of gaze-cuing and arrow-cuing

Turning to the other body of literature on gaze-cueing and social attention, Birmingham and Kingstone (2009) wrote a research review on gaze as a cue to information. Their main conclusion was that a growing collection of studies suggests that *any* cue with a directional component, e.g. an arrow pointing, a hand pointing, or eyes gazing towards something, may produce reflective orientation of attention. However, the authors argue, when observers are left free to select what they want to attend to, they focus on people and their eyes. The bottom line of the arguments is that eyes and gaze is more likely than an arrow to be attended to in the first place – but that once attended to, they have similar power in making people attend to the information/object that is looked at or pointed at.

Becchio, Bertone, and Castiello (2008) did not compare arrow cueing and gaze cueing but explored how processing of an object in the environment can be influenced by someone else looking at the object. Their conclusion is that we seem to process stimuli differently when there is another gazing agent (person) around compared to when this is not the case. The authors suggest that if another person gazes at an object and you observe that the person does so, the object in question gets loaded with more meaning than if no one had been gazing at it. An arrow pointing does not have this effect, which possibly is related to the fact that a person (agent), in contrast to an arrow, can potentially act with respect to the object (Risko, Richardson, & Kingstone, 2016). In turn, this relates to theories on so called *joint attention* – a developed form of (simple) gaze following. It creates a shared space of common psychological ground that enables collaborative activities with shared goals unique for human cognition (Moll, Koring, Carpenter, & Tomasello, 2006; Tomasello & Carpenter, 2007).

Important differences between previous studies and the present study

Before we reason about how these previous studies can guide our predictions, we will bring forth the differences between our study and the described, previous studies.

First, most studies within the gaze-cuing domain are lab-studies that present single, isolated cues in the visual field of a participant. Likewise, most studies on signalling digital agents in educational contexts have been using more or less constrained or controlled digital materials compared to our study. In contrast, the tasks provided to the students in our study, and the possible strategies for solving them, are associated with a large freedom of choices. At each point in the game, the student is exposed to choices for action, which in turn influences what information appears next. This corresponds to a high degree of ecological validity with regard to students' use of educational software.

Second, our study focused on students' dealing with critical constructive feedback whereas none of the studies presented above did so. Third, the agent signalling condition in the present study involved a digital agent that both gazes and points towards the feedback text, whereas the previous studies described featured both human agents and digital agents, some both gazing and pointing, others only pointing or only gazing agents.

Finally, and most importantly, the reported studies deal with a particular aspect of the information process, whereas our study targets the entire process from detecting a specific piece of information, reading the information, understanding and making use of the information, to finally progressing with respect to learning outcome. The studies described above that involve pedagogical agents primarily

measure *learning outcomes* (Step 5 in our model) in relation to the information signalling (arrow vs. pedagogical agent), whereas social cueing studies primarily measure *detection of information* (Step 1 in our model). From such results, little can be inferred with respect to the different intermediate steps of the information processing or the process as a whole.

Predictions

Referring to our CCF-processing model (figure 2), we aimed to identify instances of CCF-neglect in the different steps: (1) not noticing the CCF, (2) noticing but not processing (in our case reading) the CCF, (4) not acting upon the read CCF, and (5) not making progress, (i.e. fail at improving task performance, even though one has acted upon the CCF). We also explored potential differences at the different steps of the CCF-processing model between the three signalling conditions: agent visual signalling, arrow visual signalling, and no signalling (control). Below follow the predictions we made with respect to our research questions.

Prediction 1: Large CCF-neglect as measured by the last step in the model

Previous studies have shown that neglect of CCF, measured as behavioural output in terms of performance, is a common phenomenon (Hounsell, 1987; Wotjas, 1998; Perrenoud, 1998; Clarebout & Elen, 2008; Conati et al., 2013; Segedy et al., 2012). We therefore predicted a large proportion of CCF-neglect as measured when students reached Step 5 in our model. In other words, a small proportion of students would make it through all the way and improve on the task by making use of the CCF provided.

When it comes to the series of steps leading up to the final step, and the degree of 'fall off' at each of these steps, there are no previous studies to base predictions on. Thus, we left the proportion of 'fall off' *at the different steps of the CCF-process* – detection, processing, acting upon, and progress – as open questions.

Prediction 2: Effects of the three signalling conditions' on noticing the CCF

Our prediction was that the students would be more inclined to notice the feedback text in the agent condition than in the arrow condition, but also that they would be

more inclined to notice the feedback text in the arrow condition compared to the control condition.

A main reason for this prediction with respect to the signalling conditions' potential effect on noticing the CCF, is that the visual stimuli in our study were not strictly controlled and that the students had a high degree of freedom to choose different actions. This, in turn, may promote prevalence towards 'eyes' as a main cueing agent. The comprehensive research review by Birmingham and Kingston (2009) concluded that cues with a directional component such as an arrow pointing, a hand pointing, or eyes gazing towards something, all produce reflective gaze orientation. In our study, this would mean that an agent gazing and pointing towards a piece of information and an arrow pointing at the same piece of information will have equivalent effects in terms of guiding someone's direction of gaze. However, if the visual environment is complex and the freedom to act large, the same review (Birmingham & Kingston, 2009) states that eyes are more likely to attract attention in the first place vis-à-vis other visual stimuli.

Prediction 3: Effects of the three signalling conditions on processing the CCF

We predicted that the agent condition compared to the two other conditions would make students more inclined to process (in our case read) the CCF they had noticed. As related above, developmental psychologists and neuropsychologists suggest that humans give priority to social stimuli once they are detected (Gamé et al., 2003; Pinsk et al., 2009; Taylor et al., 2007). Since an agent is more likely than an arrow to prime a social schema, students ought to be more inclined to read a detected text in the agent condition than in the arrow condition. The prediction is also supported by Koning and Tabbers' (2013) argument that learners are accustomed to teachers' pointing gestures as signals to something of importance. The purpose of an agent's pointing gesture may therefore be less ambiguous than the purpose of an arrow and more powerfully indicate that: "You should bother to read, this is relevant!". In addition, the conclusion by Becchio et al. (2008) that humans seem to process stimuli differently when there is another gazing agent around, points towards the same prediction. If another person (agent) gazes at an object and you observe this, the object in question gets loaded with more meaning. The agent pointing and looking at the piece of CCF-text (the object in our case) may load the text with more meaning than an arrow pointing to the same piece of CCF-text.

Likewise, Veletsianos (2007; 2010) proposal that contextual relevance is important for a positive learning effect of a pedagogical agent, supports our prediction of a higher proportion of CCF-processing (reading) in the agent

condition than in the two other conditions. Notably the agent in our study has a central and highly relevant role in the educational game.

Study

Participants

A total of 46 students (22 boys and 24 girls) from two fifth-grade classes at a Swedish municipal school participated in the study. The students were 11- to 12-years-old with a mean age of 11:6 years and had middle-class socio-economic and socio-cultural background. All students were good or very good readers of Swedish text, according to teacher evaluations.

The digital learning environment

The Guardian of History (GoH) is an educational game targeting history, based on the pedagogical approach 'learning by teaching' (Bargh & Schul, 1980; Chase, Chin, Oppezzo, & Schwartz, 2009) where the student takes the role of a teacher and potentially learns by doing so. GoH features a digital tutee or, in technical terms, a *teachable agent* (Blair, Schwartz, Biswas, & Leelawong, 2007; Chase et al., 2009) whom the student is supposed to teach about historical discoveries and inventions and their consequences. In the game narrative Professor Chronos, who is the Guardian of History watching over the passage of time, is about to retire. When the student first enters the learning environment, the digital tutee Timy explains that s/he would like to become the successor of Professor Chronos but unfortunately suffers from 'time travel sickness'. This means s/he cannot travel with the time machine to learn about the past. Instead, Timy suggests, the student could make the time travels and return to teach her/him about the things the student has learned. Thus, the students learn by travelling with the time machine to different historical environments where they engage in text conversations with historical persons and explore the surroundings via interactive objects. In the customized version of GoH used for the present study, students were given six different missions to complete and teach Timy about.

Teaching activities

Having completed a time travel, the students´ task was to teach Timy. The teaching activities were of three different formats: the construction of a *conceptual map*, the completion of a *sorting task*, and the pairing of concepts and placing the

paired concepts on a *timeline*. To complete a task correctly, students needed to make use of the information they had gained from the time travels.

For the conceptual map, a starting node was provided, e.g. Émilie du Châtelet (figure 3). The students were then to choose the adequate concepts from a box and then select the adequate relation out of three possible alternatives.

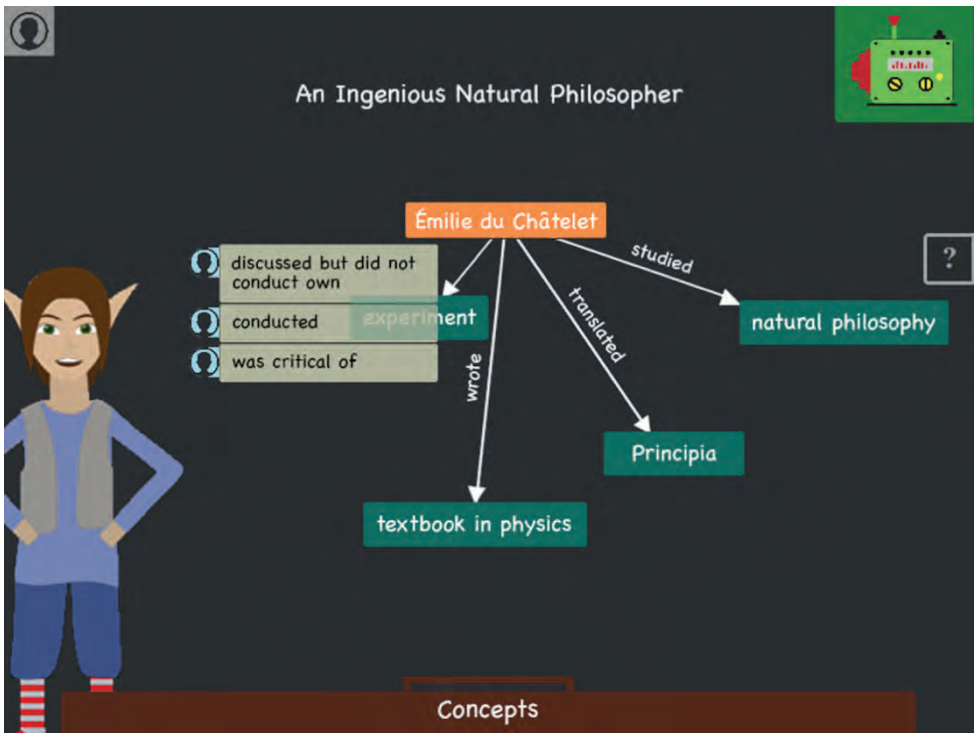


Figure 3. Classroom activity; arranging a conceptual map.

For the sorting task, students were presented with cards presenting statements that apply to one or more of the historical female scientists from the 17th and 18th centuries as well as some statements that apply to women in general during these periods of time. The statement cards were to be sorted correctly into one or several archive boxes (figure 4).

Figure 4. Classroom activity: sorting task categorizing statements.

The timeline required students to match puzzle pieces together and then drag them to the adequate time slot on the timeline. One mission targeted the scientific work and inventions of Galileo Galilei, Isaac Newton, and Émilie du Châtelet. Another mission targeted Johannes Gutenberg and the printing revolution with its preconditions and consequences. Yet another mission dealt with the factors that affected the possibilities to work with science and spreading results during this historical period of time (figure 5).

Figure 5. Classroom activity: constructing a timeline.

Feedback provided within the digital learning environment

Having completed a teaching activity, the student would click on the 'correction machine' (figure 6) for the machine to provide the result of the teaching activity by highlighting incorrect items in red and correct items in green. At the same time as the correcting machine presented this information, a textbox with critical constructive feedback was shown for one randomly chosen incorrect item (if any). Note that this CCF-text was coded to appear automatically when the correcting machine provided the results. That is, all students were presented with one first, automatically generated, CCF-text. After the automatic presentation of this CCF-text, the students were allowed to generate as many (or few) additional CCF-texts as they wanted by clicking any of the remaining red-marked incorrect items.

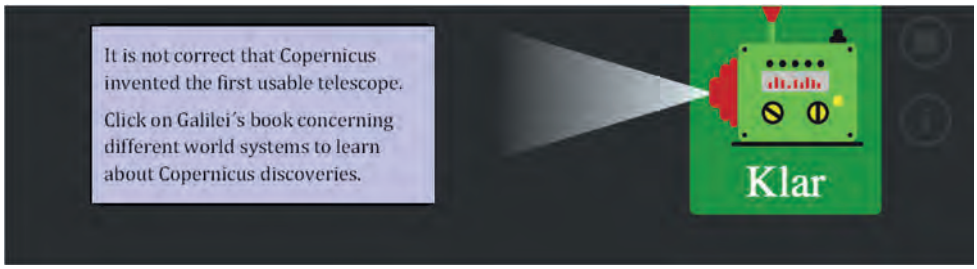


Figure 6. The correcting machine providing the CCF.

The design rationale for having a first CCF-text appear automatically was to counteract the risk that students – if it was completely optional to receive CCF-texts – by default would read (almost) all attained CCF-texts. We reasoned that, if a student makes the active choice of clicking a red-marked incorrect item, she might be primed into looking at the CCF-text. In particular, such an effect would have interfered with our comparison of the three experimental conditions with respect to their inducement of noticing and reading the CCF-text. In retrospect, we could not identify any priming effect – clicking on an error did not automatically lead to students' looking at the CCF-text – but we did not know this beforehand. In effect, the equivalence in students' inclination to notice and read CCF-texts, whether automatically appearing or chosen, enabled us to use all occurrences of CCF-events – both automatically generated and chosen by the students – in our analysis.

The content of the CCF-texts was designed to be contextualized, which has been shown important for students to find feedback meaningful (Segedy et al., 2013). The CCF-texts were composed of two parts, the 'critical' part and the 'constructive' part – which aligns with what Black & Wiliam (1998) put forth as the two main functions that feedback should fulfil. In the critical part, the feedback points out what was incorrect, while in the constructive part, the feedback gives suggestions to guide students to improve and revise the answer. For example, if a student erroneously paired the piece 'Gutenberg' with the piece 'Typewriter', the following critical feedback text appeared: *"'Gutenberg' and 'typewriter' do not belong together."* This was followed by the constructive feedback text: *"Find information to solve this by visiting Gutenberg and look more closely at the bible in his office."* That is, students were provided a hint on where to travel and where to look to find the right answer.

For each error, three different phrasings with the same central information were prepared and randomly provided in order to prevent students from getting bored by recurring identical text phrases. The length of the CCF-texts varied according to the corresponding contents, ranging from 12 to 33 words.

It should be pointed out that the main rationale for our design of the CCF in the present digital learning environment was to *enable a methodological study* of how critical constructive feedback is handled (or not handled) by students. The purpose was not to design pedagogically optimal CCF-text s and evaluate student behaviour in relation to this. The CCF-texts used in the study are according to this objective designed in a standard manner with regard to other kinds of critical constructive feedback in text format.

Design of experimental conditions

The CCF was presented in one of three different ways corresponding to three experimental conditions: (i) the teachable agent (TA) pointing and gazing to the CCF textbox (condition TA); (ii) an animated arrow pointing to the CCF textbox (condition AR); (iii) a control condition with no signalling (condition CN). Given their visual forms, the agent and the arrow were designed to have similar size and to appear at the same position of the screen (figure 7). Importantly, the animations of the hand and the arrow were identically designed. The goal was that the two ways of signalling would not differ in terms of salience, where it is known that motion is powerful in attracting attention. A small pilot study was conducted to verify this objective. Finally, the correction machine was located on the opposite side of the agent/arrow to avoid interfering eye-gaze data.

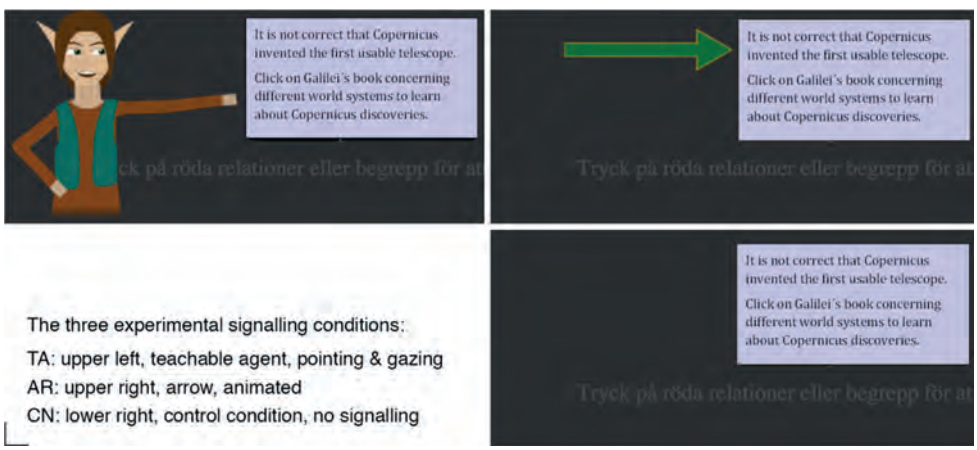


Figure 7. The three experimental signalling conditions: TA, AR, and CN.

The study followed a within-subjects design where all participants encountered all three CCF formats according to a randomized scheme; in a series of three CCF occurrences, each format was chosen once in random order. That is, we used randomized sampling without replacement. A benefit of this method was that we

could ensure that the three formats would occur equally often for each individual student as well as in the entire material.

The manipulation of CCF formats or signalling of CCF occurred all through the students' gameplay, i.e. both during the two initial training sessions as well as the final experimental session.

Procedure

The two participating classes used the learning game during three sessions that each lasted approximately one hour. The students played the educational game individually using their own login and were instructed to focus on their own game play. The first two sessions took place during scheduled history lessons at the students' school using the school's iPads. Researchers were present in order to assist with technical and interaction related issues but did not help out in solving any game tasks. The third study session took place in a lab classroom at Humanities Lab, Lund University, where each student was seated in front of a desktop computer with an integrated eye-tracker. After a short introduction and presentation of the task ahead of them, the students were instructed to go through an individual calibrating procedure for the integrated eye-trackers. Following the rather swift and smooth calibration procedure (running through a set of fixations on the screen), the students logged into the game. In order to secure the eye-tracking data collection, researchers occasionally gave gentle encouragement to children to keep an upright position when they tended to slump down in front of the computer screen.

During the two first sessions the students familiarized themselves with the educational game and learnt how it worked with respect to interaction and the kinds of tasks provided. In addition, they familiarized themselves with the four researchers. The benefit of this was that all students knew what was expected of them and were sufficiently relaxed (they understood the game and were familiar with the researchers) at the third sessions when they went to the university lab where the eye-tracking took place. In this way we could maintain a relatively high ecological validity even though the third session was taking place elsewhere than in the students' regular classrooms. If we would have had access to portable eye-trackers we would have preferred to conduct the third session as well in regular classrooms at their school, but this was not an option.

During the two first sessions, when the students used the game in their regular classrooms, four of the six missions were available. Regardless of how far individual students had come during the two first sessions, they all started at mission five at the third session when the eye-tracking took place. This way we could control that the content and tasks at this data collection session were similar

for all students. With the exception of the eye-tracking data collection, everything concerning the educational game and the game-play was the same during all three sessions, including the exposure to the three signalling conditions.

Data collection method

Methodologically the study used a combination of eye-tracking and behavioural data logging. For the two first steps in our CCF-model – noticing CCF (Step 1) and processing (reading) CCF (Step 2) – data was collected by means of eye-tracking. The two last steps in our CCF-model – whether student acts on CCF or not (Step 4) and whether the student makes progress or not (Step 5) – were measured via behavioural data logging, i.e. logging of the students interacting with the learning tool. In addition, contextual data was gathered via a questionnaire that the students were asked to complete at the end of the third (experimental) session.

Eye-tracking data collection

The eye-tracking equipment was composed of a remote eye-tracking system with a SMI REDm eye-tracking camera integrated in a desktop computer screen. The rationale for choosing this remote eye-tracker system was to provide a less intrusive and more comfortable environment considering that the students were using the educational game for as long as 50 minutes. This type of remote eye-tracking system captures eye-movement data with less accuracy, precision, and sampling frequency, but allows for simultaneous recording of multiple students, thus making it possible to emulate a classroom environment. Given the resolution of the eye-tracking system, the analyses did not focus on word-level analyses, but rather on the noticing of particular elements and a more general reading activity. Prior to the eye-tracking recordings, the participants went through a calibration procedure with an animated 9-point calibration method available in the SMI iViewX software. During the eye-tracking recordings, eye movement events such as fixations, saccades, and blinks were detected using the SMI BeGaze 3.6. software. For more details on the eye-tracking methodology, we refer to Lee, (2017).

Log data collection

For the data collection on students acting (or not acting) upon CCF, i.e. attempting to act in accordance with the hints provided by the CCF-text, we used behavioural data logging integrated in the game code. The evaluation of student performance (progress or not progress) – with respect to the provided CCF – was likewise studied by means of behavioural data logging. The following data was extracted from the log data for each instance where a CCF-text was presented to a student:

- if the CCF was automatically provided or chosen,
- the CCF framing (agent, arrow, or control),
- the content of the CCF textbox,
- the student game activities (after the CCF textbox had been closed), i.e. whether the student tried to pursue the hints in the CCF or not,
- the scores on each teaching activity before and after CCF provision.

Measurement

Noticing CCF

Noticing, in the sense we use it, involves awareness at a very low level of abstraction and occurs when allocating attention resources to features in visual input. As measurement of noticing CCF, we chose detection of looking or glancing at the CCF textbox, making use of fixation-based areas of interest (AOI) hits provided by the eye-tracking data (Holmqvist, Nyström, Andersson, Dewhurst, Jarodzka & Van de Weijer, 2011). In the present study, the CCF textboxes were defined as AOIs and a fixation (hit) holding a coordinate value inside the AOI was counted as 'noticed'. In this way, every CCF-instance was categorized as either 'noticed' or 'not noticed', i.e. CCF-neglect See figure 8 for an example of typical eye-movement patterns.

Reading CCF

The identification of reading behaviour by means of the eye-tracking data exploited a support vector machine (SVM) for data analysing and pattern recognition in order to separate reading and non-reading events. In view of the large differences among individuals with respect to reading behaviour, the support vector machine was applied on an individual level. Furthermore, three different eye-movement measures relevant to reading (fixation duration, saccadic amplitude, and regression) were considered in the evaluation model. Every CCF-text instance was labelled with the binary value of 'read' or 'not-read' (i.e. CCF-neglect) based on an intrinsic threshold set by a pilot study. For further details, see Lee, (2017).

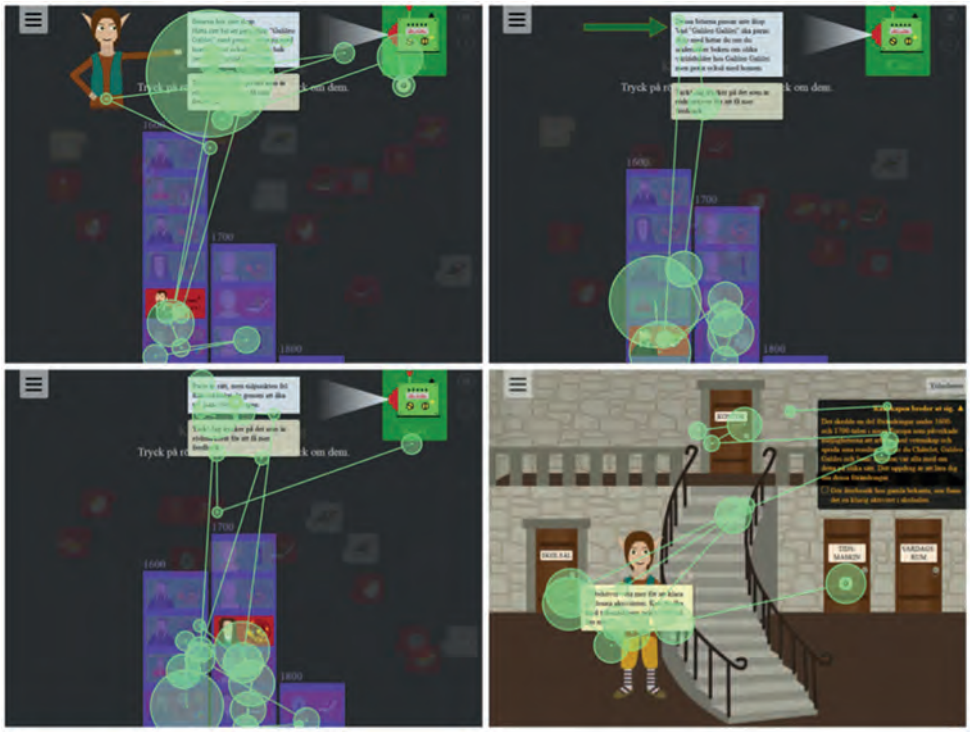


Figure 8. Examples of a participant's eye movements during the game play superimposed on the corresponding game scenes: the tree different signalling conditions; TA (upper left), AR (upper right), CN (lower left), and the central castle hall right after having received a new mission (lower right).

Making sense of CCF

This step was only indirectly measured in our study since the direct measurement of whether a student makes sense (or not) of a text requires other methods than eye-tracking and behavioural data logging. In our case, however, there are close relations to the preceding step of 'process (read)' and the subsequent step of 'act upon'. In the context of the present study it is very unlikely, but not entirely impossible, that a student *acts on a given* CCF (Step 4) without having *made sense* of the CCF (Step 3). On the other hand, when a student in our study has *read* a given CCF-text (Step 2), it is likely that she also has *made sense* of the CCF, given the teacher's assurance that the CCF-texts were comprehensible for the age group and that all participating students were good or very good readers. Yet, for a given student and CCF-instance, we do not have sufficient information to tell for sure whether a student who did not act upon a given CCF (Step 4) also did not understand (Step 3) what she had read (Step 2) – or actually had understood the CCF, but still not acted upon it.

Acting upon CCF

To measure whether students acted (or not acted) upon an instance of a CCF-text that they had read, the interaction history logged in the data (e.g. time travelling and text information items the student had clicked on) was extracted. These interaction data logs were then evaluated with respect to the instructions and hints provided in the CCF-text. Depending on the extent to which the student followed the CCF-instructions, she received an 'act upon' score between 0 and 3. Score 0 means that a student travelled with the time machine to a place and time that did not relate to the hints and instructions provided in the CCF-text. Score 1 was given when a student time-travelled to the right place and time, but 'clicked on' the wrong informative item (e.g. if instructed in the CCF to visit Newton and investigate his manuscript to 'Principia'; the student did visit Newton but only 'clicked on' Newton's diploma on the wall). Score 2 was given when a student time-travelled to the right place and time, as hinted in the CCF, and 'clicked on' the suggested informative item together with other items. Finally, a score of 3 was given when a student travelled to the right place and time and only 'clicked on' an object or person that had been suggested in the CCF-text. For the analyses, only scores of 2 and 3 were counted as 'act upon' (i.e. the student had followed the instructions/hints provided in the CCF appropriately) and the scores of 0 and 1 were noted as 'not acted upon', i.e. CCF-neglect.

Progress following acting upon CCF

The progress score in our model was calculated by evaluating whether students had corrected the errors on which they had received CCF in their next attempt to complete the teaching activity. Since each error was due to one or several incorrect choices made by the student, the percentage of correct choices constituted the progress score. For example, the two identical (puzzle) pieces (cf. figure 5) 'the printing press' should each be paired with either 'mass-production of brochures' or 'mass-production of books'. If neither of these pairings were made by the student, she would receive CCF such as: "*These pieces do not belong together, you will find the right piece to pair the 'printing press' with if you check the Bible in Gutenberg's workshop or if you visit the Priest in Wittenberg.*" If the student during the next round in the teaching activity correctly paired 'the printing press' with 'mass-production of books', but did not pair the other piece of 'the printing press' with 'mass-production of brochures', this would count as a progress of 50%.

Students' answers to some explicit questionnaire questions about the feedback texts

At the end of the third (experimental) study session, all students were asked to fill out a questionnaire about the educational game, including four question items

probing their experiences and thoughts about the feedback (CCF): To what degree would they say that they read the feedback texts? To what degree would they say that they clicked to receive more feedback (more than the automated one)? To what degree would they say that the feedback was helpful for correcting their errors? What did they think about the amount of feedback?

Since this information is contextually relevant for the present research questions, we included this data in our result section. (Other parts of the questionnaire are used for other research purposes.)

Results

All statistical analyses were performed with the statistical software R (R Core Team, 2016). The logistic mixed-effects linear analysis made use of the R method *lme4* (Bates, Mächler, & Bolker, 2012).

Of the 46 students participating in the study, four students did not participate in the third (experimental) session, i.e. the session where the eye-tracking took place. An additional six students were excluded from the analyses due to technical problems resulting in loss of eye-tracking data.

The remaining data set for analyses thus consisted of 36 students, 20 girls and 16 boys. For these 36 students, 451 CCF-instances were recorded, i.e. situations where a text box with critical constructive feedback was presented on the computer screen. An additional twenty-seven incidents of data loss in the data logging associated to Step 4 and 5 of the CCF-model further reduced the number; resulting in a final data set of 424 CCF-instances, 218 automatically presented and 206 opted for, distributed over 36 students. Out of these, 143 instances occurred with arrow signalling, 142 with agent signalling, and 139 with no signalling (control condition).

Research questions I: At which steps does CCF-neglect occur?

Out of the 424 CCF-instances in the data set presented to students, only 4% (figure 9) remained after the last step (Step 5), i.e. in only 4% of presented CCF-texts students made progress with the task after first noticing, reading, and then acting upon the CCF. This aligns with our prediction (Prediction 1) of a large amount of CCF-neglect as measured by the last step (progress) of our model.

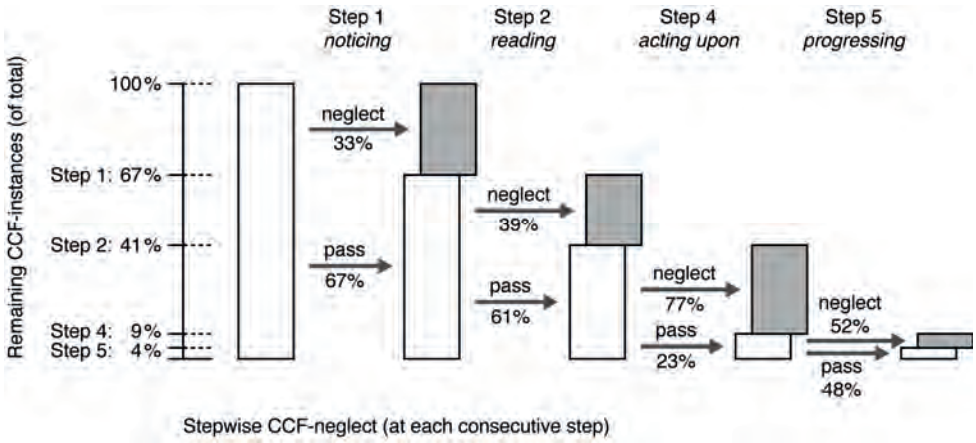


Figure 9. Y-axis showing the remaining CCF-instances (out of the total initial CCF-instances) after Step 1, 2, 4 and 5 of the model; x-axis showing the stepwise CCF-neglects at each and one of the four consecutive steps (Step 1, 2, 4 and 5) of the mode.

Figure 9 also presents the percentages of CCF-neglect in Steps 1, 2, 4 and 5 of our model. In Step 1, 33% of the CCF-instances presented to the students were neglected (not noticed). Out of the CCF-instances that passed Step 1, 39% were neglected (not read) in the second step, and of the CCF-instances that passed both Step 1 (notice) and Step 2 (read), a total of 77% invoked no CCF-related actions in Step 4 (act upon). Finally, 52% of the remaining CFF-instances, that had passed the steps of noticing, reading, and acting upon, did not lead to any CCF-prompted progress in the last, fifth step (progress) of our model.

Figure 9 finally reveals that the largest 'falling off' of students occurs at the fourth step (acting upon the CCF). Pairwise Chi-square tests comparing the proportions of pass/neglect (see figure 9) shows a significantly larger 'falling off' (portion of feedback neglect) for Step 4 (acting upon) compared to Steps 1, 2, and 5 (table 1).

Table 1. Pairwise Chi-square test comparing the proportions of 'pass - neglect' (figure 9) for the four Steps 1, 2, 4, and 5 of the FB-model. The alpha level was divided with a factor 6 to adjust for multiple comparisons.

fall off	**Chi-square** (χ^2)	***p*-value**		**Effect Size**	
				Phi (φ)	**Strength**
notice > read	0.462	0.50		0.058	(very) small
notice > act upon	36.7	1.4e-09	***	0.438	medium to large
notice > progress	6.80	0.0091	.	0.194	small to medium
read > act upon	28.1	1.2e-07	***	0.385	medium to large
read > progress	3.21	0.073		0.137	small (to medium)
act upon > progress	12.0	0.00054	**	0.255	(small to) medium

Significance levels: . 0.0167 * 0.0083 ** 0.00167 *** 0.000167

Research questions II: Does the extent of CCF-neglect differ in relation to the signalling conditions?

The focus of Research Questions II was on whether neglect of CCF – in the senses of not noticing the CCF-text, not reading the CCF-text, not acting upon the CCF-text, not making progress by using and acting upon the CCF-text – would differ between the three conditions of: (i) a virtual agent looking towards and pointing to the CCF-text, (ii) a dynamical arrow pointing to the CCF-text, and (iii) the CCF-text as such (control condition).

With respect to this set of questions we made two further predictions. The students would be more inclined to notice the feedback text in the agent condition than in the arrow condition and more inclined to notice the feedback text in the arrow condition compared to the control condition (Prediction 2). In the agent condition students would be more inclined to read the CCF they had noticed compared to in the two other conditions (Prediction 3).

The overall result is presented in figure 10, showing the consecutive 'falling off' with regard to feedback neglect. The figure suggests a positive effect of the 'agent' signalling condition on CCF-neglect for the two first steps of our model (compared to the arrow and control conditions). After that, this 'agent signalling effect' is eradicated for the two last steps ('act upon' and 'progress'). Figure 10 also parallels the large 'falling off' in the step 'act upon' presented above, a result overriding any effects of signalling condition. Next follows a more detailed analysis of the differences between the three signalling conditions for the first two steps of the model, thereby addressing Prediction 2 and 3.

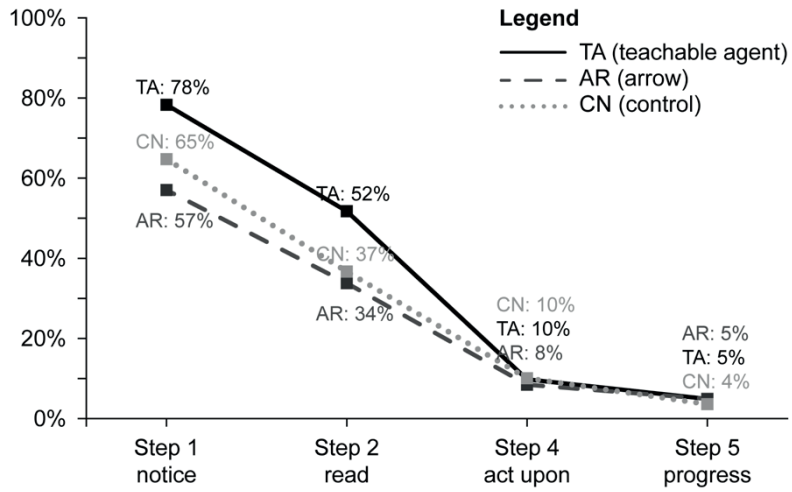


Figure 10. Consecutive 'falling off' (remaining CCF-instances out of the initial amount of instances) for each of the three signalling conditions: teachable agent (TA), arrow (AR), and control (CN) for Steps 1, 2, 4, and 5 of the FP-model.

Effects of the three signalling conditions on noticing and reading CCF

For a more thorough analysis of the differences between the three signalling conditions in the first two steps of the model, table 2 shows means and standard deviations calculated from the means of each individual participant and experimental condition in order to balance the influence of a few participants alone increasing the total mean. On average, each student noticed 67% (*SD* = 26%) and read 43% (*SD* = 28%), of the total amount of feedback boxes presented to them, which means that on average 64% of the previously noticed feedbacks were actually read. (Note that the 'Total Mean' (Tot.) for 'Reading' in table 2 slightly deviates from the corresponding number in figure 9, as the means are calculated in different ways.)

Table 2. Means and standard deviations for the different experimental conditions.

	Tot. (%) *M* (*SD*)	TA (%) *M* (*SD*)	AR (%) *M* (*SD*)	CN (%) *M* (*SD*)
Notice	67 (26)	75 (34)	61 (35)	66 (33)
Read	43 (28)	52 (35)	38 (35)	40 (35)

To compare the effects of the three signalling conditions on the 'noticing' and 'reading' steps in the model and account for the within-subject factor of the participants, we used logistic mixed-effects linear regression analysis for each of the two steps respectively. The regression models featured the two respective steps of 'notice' and 'read' as binominal outcome (dependent) variables and signalling

conditions as an independent variable with the individual students (participants) as the random factor:

notice/read[yes/no] ~ *signalling_condition*[TA/AR/CN] + *participant*(random factor)

As this analysis only included the first two steps of the CCF-model (relying on eye-tracking data), the larger data set of 451 CCF-instances could be used.

Table 3. Logistic mixed-effects linear analysis for the TA (teachable agent) and AR (arrow) conditions against the CN (control) condition on effects of visual signalling on CCF-neglect in the 'noticing' and 'reading' steps of the model.

	Noticing					Reading				
	Est.	*SE*	*Z*	*p*		Est.	*SE*	*Z*	*p*	
(Interc.)	0.474	0.288	1.645	0.100	.	-0.802	0.287	-2.792	0.005	
TA	0.959	0.283	3.383	< 0.001	***	0.979	0.265	3.691	< 0.001	***
AR	-0.139	0.262	-0.531	0.595		0.132	0.270	0.487	0.626	

Significance levels: . $p < 0.10$ * $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$
Model fit. Likelihood ratio tests of the full model with the effect in question against the model without the effect in question showed that the independent variable contributed significantly to explaining the observed variance in noticing and reading (notice: $\chi^2(2) = 17.3$, $p < .001$; read: $\chi^2(2) = 16.3$, $p < .001$).

The logistic regression analysis (table 3) provided partial support for our predictions displaying significant effects during the 'noticing' and 'reading' steps for the TA (teachable agent) condition (notice: $Z = 3.38$, $p = .001$; read: $Z = 3.69$, $p < .001$), but not so for the AR (arrow) condition, compared to the CN (control) condition. Thus, the TA condition seems to encourage noticing and reading positively compared to both the AR (arrow) and CN (control) conditions, whereas there were no significant differences between the AR (arrow) and CN (control) conditions.

Effects of the three signalling conditions on Step 4 and 5 (act upon CCF & make progress with task)

No predictions were made with respect to these two steps. Furthermore, the percentage of falling-off was considerably large in Step 4 (see figure 9 & 10) and the resulting data set in terms of CCF-instances became critically small for these last steps (CCF-instances (act upon): TA = 14, AR = 12, CN = 14; CCF-instances (progress): TA = 7, AR = 5, CN = 7). We consequently deemed statistical comparison between groups inappropriate and only use descriptive numbers.

In the fourth step (act upon), 23% of the CCF-instances that had been both noticed and read were also acted upon (figure 9). Separated on the three experimental conditions, 19% of the CCF-instances were acted upon in the agent condition,

25% in the arrow condition, and 27% in the control condition (cf. figure 10). In the fifth step (progress), 48% of the CCF-instances that had been noticed, read, and then acted upon led to progress. For the three experimental conditions, 50% of the CCF-instances led to progress in the agent condition, 58% in the arrow condition, and 36% in the control condition (cf. figure 10).

Questionnaire data

The questionnaire data was primarily used to gain contextual knowledge and to address the basic conditions (discussed in the introduction), that need to be fulfilled for students to be able to profit from CCF. For this analysis, we used the collected data from the 42 students present during the third (experimental) session, i.e. including the 6 participants with technical issues as to the recording of eye-tracking data.

First, for CCF to be potentially useful, it must be comprehensible and the amount of feedback has to be adequate. Regarding comprehensibility, the class teachers had been asked to evaluate the text and had affirmed that the text was on an adequate level of difficulty. As for the amount of feedback, one 3-level questionnaire item explored the students' views on this by asking: *"Was there an adequate amount of text in the feedback boxes? [too much text, adequate amount of text, too little text]"*. A clear majority, 30 (71%) of 42 students, answered that they found the amount of text in the text boxes adequate, 7 (17%) answered *"too much text"*, and 4 (10%) *"too little text"*. (One student did not answer the question.)

Second, the students need to be able to understand the point of the feedback. The questionnaire data provides some information on this. It included one 5-level item asking, *"How often did you read the feedback texts? [always, often, sometimes, seldom, never]"* followed up by a free-text question asking, *"Why?"* Out of the 42 students, 23 (55%) answered that they *"always"* or *"often"* read the feedback texts, 12 (29%) that they *"sometimes"* read the feedback texts, and 7 (17%) that they *"seldom"* or *"never"* read the feedback texts. For the free-text follow-up question *"Why?"*, 18 (46%) of the 39 students who provided a written answer reported that they were helped by the CCF whereas 9 (23%) answered that it was of 'no help' or that they 'did not know' (figure 11).

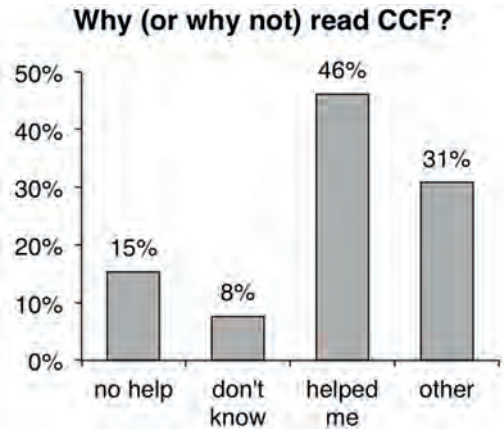


Figure 11. The results on the free-text follow-up question: "*Why [did you read/not read feedback]?*"

Another 5-level questionnaire item directly probed the students' understanding of the point of the feedback: *"Did the feedback texts help you? [1 = not at all, 3 = somewhat, 5 = a lot]"* (figure 12). One third (33%) of the students stated that they were helped by the feedback texts, another 33% that they were somewhat helped, and the remaining 33% that they were not helped by the feedback texts.

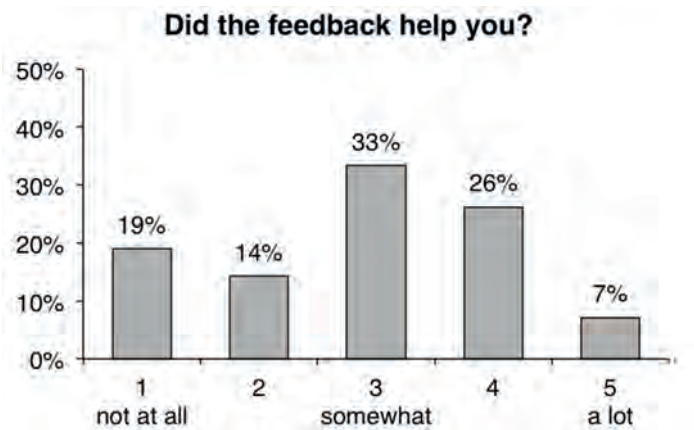


Figure 12. The results on the questionnaire item: "*Did the feedback texts help you?*"

In sum, the questionnaire data indicates that it was possible for most of the students to understand and process the CCF and that the amount of feedback was adequate or reasonable. But the questionnaire data also indicates that one third of the students found that there was too much feedback or that the feedback was not helpful, and that another third of the students found them only partially helpful.

Discussion

The goal of the study was to enter the black box of feedback processing. We wanted to show that it is possible to study the phenomenon of feedback neglect, specifically neglect of critical constructive feedback, CCF, other than by solely looking at a final outcome in terms of students' improvement (or not) on a task.

Using a novel model, we addressed initial detection of CCF, followed by cognitive processing (in our case reading) and acting upon CCF, until final progress in terms of solving a task. By doing so, we were able to study disruptions along the way in the form of CCF-neglect.

In studying CCF-neglect throughout the different steps in our model, we also explored whether or not the *presentational framing of CCF* could have an impact on the respective steps of the model. The study compared two different forms of visual signalling – a pedagogical agent and a pointing arrow – as well as a control condition with no visual signalling.

In the following, we will summarize and discuss the principal findings of the study.

CCF-neglect at different steps of the CCF-processing model

In line with previous research and our first prediction, we saw a large proportion of CFF-neglect as measured by progress at the final step (Step 5) of the CCF-processing model.

Looking closer at Steps 1, 2, 4, and 5 of the model, they all revealed losses in term of feedback neglect with the largest loss in the fourth step (act upon CCF). One third (33%) of all CCF-texts were neglected already in the sense of not being noticed (Step 1). Out of the noticed texts, a little more than one third (39%) were neglected in the sense of not being read (Step 2). Next, out of the CCF-texts that were noticed and read, about three quarters (77%) were neglected in the sense of not being acted upon and finally, of those acted upon, around half (52%) did not lead to progress, i.e. the task of correcting an error following hints provided by CCF was not successfully solved at the student's next attempt.

The largest falling off occurred in the fourth step (act upon) of the CCF-processing model. One possible reason is that the CCF-texts were not sufficiently useful for sufficiently many students. Results from the questionnaire supports this explanation where one third of the students stated that the feedback texts did not help them, and another third reported the texts were only partially helpful. The fact that only half of the CCF-texts that were 'acted upon' resulted in progress also

indicates that the feedback was not useful enough for the target group as a whole. On the positive side, the relatively high proportions of non-neglects in the two initial steps of noticing and reading, indicate that the presence of feedback texts per se are *potentially meaningful* in educational software with the target group in question; they will be read rather than neglected.

Effects of the experimental conditions on noticing and reading CCF

With respect to the three framing conditions (agent signalling, arrow signalling, and no signalling), results were that the agent condition positively affected both the 'noticing' step and the 'reading' step, whereas the arrow condition did not show any significant difference compared to the control condition in either of these two initial steps.

The experimental conditions effects on noticing CCF

We predicted – based on previous related research – that students would be more inclined to notice the CCF-texts in both the agent and arrow conditions compared to the control condition as well as be more inclined to notice the CCF-texts in the agent condition than in the arrow condition. The prediction was only partly supported in that there was no difference between the arrow condition and the control condition with respect to CCF-neglect in this first step (noticing) – whereas the agent condition alone did have a significant positive effect in that the amount of CCF-neglects was reduced.

Why did the agent signalling, but not the arrow, have an impact on noticing the CCF compared to the control? One possible explanation is that the contextual difference between the agent and the arrow already at the step of noticing influenced to what extent students fixated the target (the CCF-text) for the gaze or pointing. It is still possible that both conditions affected *gaze direction* similarly, but that the observed difference in fixation between the two conditions is due to the gaze of the time elf (teachable agent), since another agent may influence object processing (Becchio et al., 2008).

An alternative explanation can be found in Birmingham and Kingstone's (2009) argument that even though humans react similarly to an arrow, an agent's gaze, and other directional cues, they are more likely to notice eyes in the first place. Possibly, in our study, the agent's eyes (synchronized with the pointing arm) were more likely to be noticed in the first place than the arrow – and once noticed, the students used it as a directional cue and also noticed the feedback text. Although we can't resolve this possibility in our study, this could be explored in a follow-up study.

The experimental conditions effects on reading CCF

Turning to the second step, reading the CCF-texts, we predicted that students would be more inclined to read the noticed CCF-texts when in the agent condition compared to when in the two other conditions. The results support this prediction and also lend support to the theories and previous studies on which we based our prediction. Thus, the results align with the theory that humans give priority to social over non-social stimuli also once detected (Gamé et al., 2003; Pinsk et al., 2009; Taylor et al., 2007) in that the participating students seemed more inclined to care to read the text in the agent condition than in the arrow condition). The results also align with the argument that an agent's pointing gesture – more clearly than an arrow's pointing – may indicate that the student 'should bother to read' (Konig & Tabbers, 2013). In other words, the agent's pointing and gazing towards the CCF-texts may have loaded these texts with more meaning than the arrow pointing to the same piece of CCF-texts, whether from a general communicative perspective or because the students are primed that teachers' or peers' pointing gestures signal something of importance in a learning situation.

Relatedly, with respect to the use of a teachable agent (TA) in our experimental design, the positive effect of signalling may arise from the contextual relevance of the agent in question (Veletsianos, 2007; 2010). The time elf Timy's role in the game, as someone asking the student for help in a joint mission, strengthens the student's willingness as well as responsibility to care about the object of Timy's interest – in this case to read the text. Actually, the specific effects of teachable agents instigate further elaboration. The use of a TA has been shown to influence students' attitudes toward errors. Chase et al. (2009) found that students who taught TAs were more likely to acknowledge errors than students who learned for themselves, suggesting that teaching a TA protects students' egos from the psychological ramifications of failure – and fear of failure can be a reason for feedback neglect (Nicholls, Cobb, Wood, Yackel, & Patashnick, 1990). In the present study, the reduced neglect as to the reading of CCF-texts in the agent condition may relate to the fact that it is the (teachable) agent that needs help to pass the tests in the game and thus risks failing these tests. In other words, the critical feedback provided by a teachable agent poses less of an ego threat.

Yet another mechanism that may have been in play is that of *shared gaze*. Becchio et al. (2008) explored how processing of an object in the environment can be influenced by someone else's looking at the object. They concluded that we seem to process stimuli differently when there is another social agent around, compared to when this is not the case.

The authors suggest that if another person gazes at an object and you observe that the person does so, the object in question gets loaded with more meaning than if no one had gazed at it. A pointing arrow does not have this effect, which possibly

is related to the fact that a person in contrast to an arrow can potentially act with respect to the object (Risko et al., 2016). This, in turn, relates to theories on so called *joint attention* – a developed form of simple gaze following. Joint attention creates a shared space of common psychological ground that enables collaborative activities with shared goals (Moll et al., 2006; Tomasello & Carpenter, 2007). Notably, teachable agents show pronounced social characteristics in engaging learners in a teacher-tutee metaphor (Chase et al., 2009). It can be assumed that the participants in the study have constructed a social relationship with their TAs which did not only influence simple gaze following as shown by enhanced noticing behaviour – but also advanced joint attention, as demonstrated by encouraged reading behaviour.

Effects of the experimental conditions on acting upon CCF and on progress

We made no predictions on whether the different CCF framing conditions would have any effect on the two last steps of the CCF-processing model ('act-upon' and 'make progress') and the behavioural data logs showed no significant differences between the conditions. Thus, in contrast to the two preceding steps (noticing and reading), agent signalling did not affect the steps of acting upon and progressing compared to the two other conditions. This means that the step where the largest falling-off or neglect of CCF occurred – the step 'act upon' – was not affected by the different CCF framings.

A possible explanation, supported by our questionnaire data, is that the CCF provided was not sufficiently useful or helpful for some of the students. When feedback text is designed in a single format, following a 'one-size-fits-all' approach, there is a risk that it will not be adequate for all students in a group, e.g. a school class. In the experimentally adapted version of the educational game used in our study there were, for example, no options to choose a simpler text and/or to look up difficult words, something that would be desirable from an educational perspective.

It is worth to point out, however, that the pedagogical value of the TA-condition in increasing the percentage of students that notice as well as the percentage of students that then read CCF, is not nullified by the absence of any effects in the act-upon-step. The TA-signalling seems better at making students notice and read CCF and is therefore more likely than the other conditions at influencing students that are not from the start inclined to notice and read. It is thus possible that we in a more large-scale study would have obtained a larger cumulative net effect in a TA-condition.

But the crucial result is that it appears possible to influence students who are not initially inclined to notice and read feedback text into doing so. Future work will address how they can also be scaffolded to also act upon what the feedback says and how the design and presentation of feedback texts can be improved to exploit the potential of the increased number of student who read the CCF-texts. For the specific game that was used in the study, an implication is that more work is required on the design of the CCF-texts as well as scaffolding.

Distracting or Not?

Many studies on signalling by means of digital agents in educational contexts are conducted within a framework that highlights cognitive (over)load as a potential hindrance for learning. From a cognitive load perspective, the visual presence of an agent may increase the extraneous cognitive load by adding a visual display component that distracts the learner from what is central or relevant information on the display (Johnson et al., 2015; Moreno, 2005). In our study the visual presence of the time elf Timy would, according to this theory, guide the visual attention of the learner away from the CCF-text, thus interfering with the desired information selection process on the part of the learner. The same would apply to the arrow.

Our results do not support this view. For the arrow, we saw no differences from the case with no signalling. As for the agent, the potential benefits from encouraging and scaffolding the learning process rather seem to trump the possible detrimental effects of increased cognitive load.

From a practical-pedagogical perspective, this result can be used by educational game designers. Pedagogical agents can be engaged to signal central information, such as feedback, in order to prompt the likelihood that students notice and also process the information. It should be noted, however, that contextual relevance of the agent probably is a prerequisite to yield an effect.

Limitations of the study and future studies

Limitations

The lack of a controlled environment can be deemed as a limitation. For instance, the large freedom in how to take on a task produced a variety in the number of CCF-instances generated for individual students. Constructing a more controlled

environment by providing all the participants with the same amount of CCF-instances would yield a more balanced data set, but this would be at the expense of ecological validity.

Other limitations are the small sample size in terms of the number of students and the exploratory nature of the study. Yet another limitation is the large number of errors that some students made within a mission and that all generated CCF-texts. Even if the CCF provided for each error or mistake was adequate and comprehensible, the total amount of information to deal with in these cases may have been too large to handle. For future studies of feedback neglect, the amount of information is an apparent parameter for closer examination.

In terms of measurements and technical apparatus, the type of eye-tracking equipment used has limitations as described above in the section on data collection measures. In addition, the data analysis of the two last steps in our CCF-processing model ('act upon' and 'progress') might be constructed in alternative ways.

A potential limitation lies in the study design randomly presenting the students with the three CCF framing conditions (and not one). Potentially, this design could have been perceived as odd or disturbing by the students. However, a semi-structured interview with ten students revealed that none of these students seemed to have thought about the fact that there were three different framing alternatives. In other words, they had not found the set-up odd and neither had they reflected upon it. This suggest that the study design mixing the three CCF conditions didn't introduce any confounding elements by participants being aware of what was manipulated in the study.

As to the presentational format, we only studied CCF provided as pure text. Other relevant and interesting presentational formats, not the least within educational settings, are different graphics, animations, and audio formats – and combinations of these.

Finally, in this study, it is a digital learning environment that provides critical constructive feedback to students. This differs in some respects from when a student gets CCF by a teacher who, for example, comments on an error made by the student or marks her essay. In the latter situation there are additional variables involving relational and communicative factors between teachers and students that are not covered in the present work. Therefore, results from our study cannot straightforwardly be generalized to the situation of a human teacher providing CCF to a student. However, findings from a computer environment about factors involved when students handle – or don't handle – critical constructive feedback may be later explored in a classroom environment to determine if they have similar influence there. Simultaneously, given the growing ubiquity of computer-based learning environments, understanding how the uptake and use of critical

constructive feedback can be increased in these environments is valuable in its own right.

Future studies

The results of our study provide novel information about when students fall off in the process of handling critical constructive feedback. A next step is to continue the investigation by probing more into why they fall off. There are individual variables that can influence feedback neglect (as well as uptake), such as different goal orientations (Dweck, 2000). There is also a potential, very prosaic explanation for CCF-neglect: if there is an alternative to processing and using CCF in order to be able to move on, for instance retrying the task in a trial-&-error style, a student may prefer this alternative strategy in an attempt to minimize her effort even if it is not an efficient way of making progress. Future studies on the process of feedback neglect may scrutinize this. Likewise, future studies may investigate why students do not notice the feedback text in the first place and what they are doing and/or looking at instead.

With regard to the signalling conditions used in this study, we propose futures studies to learn more about *when*, *why*, and *how* pedagogical agents can be fruitfully used to signal (and help structuring) feedback. With a focus on individual factors, participants' prior knowledge is likely to be one variable of interest. Several studies have shown that pedagogical agent signalling may have limited beneficial effect on learning outcomes when averaging across all learners, but large beneficial effects on learners with low prior knowledge (Choi & Clark, 2006; Johnson et al., 2013).

Yet another line of future investigation, relating to traits and effects of pedagogical agents, could compare the signalling of a teachable agent with the signalling of other kinds of pedagogical agents. Such a line of research may tell us more about the effects of different social attributions.

Finally, the CCF-processing model proposed in this article, can also with appropriate modifications, be used to study the phenomenon of CCF-neglect more broadly in different digital contexts.

Acknowledgements

This research was funded by Marcus and Amalia Wallenberg foundation. The authors want to thank the anonymous reviewers of this paper for their thorough

and extremely helpful reviews. Last, but not least, we thank all students and teachers who participated in the study.

References

Bargh, J. A., & Schul, Y. (1980). On the cognitive benefits of teaching. *Journal of Educational Psychology, 72*(5), 593-604.

Bates, D., Mächler, M., & Bolker, B. (2012). lme4: linear mixed-effects models using S4 classes [Computer software]. Retrieved from https://cran.r-project.org

Becchio, C., Bertone, C., & Castiello, U. (2008). How the gaze of others influences object processing. *Trends in Cognitive Sciences, 12*(7), 254-258.

Birmingham, E., & Kingstone, A. (2009). Human social attention. A new look at past, present, and future investigations. *Annals of the New York Academy of Science,* 1156, 118-140.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74.

Blair, K., Schwartz, D., Biswas, G., & Leelawong, K. (2007). Pedagogical agents for learning by teaching: Teachable agents. *Educational Technology & Science, 47*(1), 56-61.

Brockbank, A., & McGill, I. (1998). *Facilitating reflective learning in higher education*. Bristol, PA: Taylor & Francis.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79*(4), 474-482.

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.

Choi, S., & Clark, R. E. (2006). Cognitive and affective benefits of an animated pedagogical agent for learning English as a second language. *Journal of Educational Computing Research, 34*(4), 441-466.

Clarebout, G., & Elen, J. (2008). Advice on tool use in open learning environments. *Journal of Educational Multimedia and Hypermedia, 17*(1), 81-97.

Conati, C., Jaques, N., & Muir, M. (2013). Understanding attention to adaptive hints in educational games: an eye-tracking study. *International Journal of Artificial Intelligence in Education, 23*(1), 136-161.

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

Elawar, M. C., & Corno, L. (1985). A factorial experiment in teachers' written feedback on student homework: Changing teacher behavior a little rather than a lot. *Journal of Educational Psychology, 77*(2), 162-173.

Gamé, F., Carchon, I., & Vital-Durand, F. (2003). The effect of stimulus attractiveness on visual tracking in 2- to 6-month-old infants. *Infant Behavior and Development, 26*(2), 135-150.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Higgins, R., Hartley, P., & Skelton, A. (2001). Getting the message across: the problem of communicating assessment feedback. *Teaching in higher education, 6*(2), 269-274.

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

Hounsell, D. (1987). Essay writing and the quality of feedback. In J. T. E. Richardson, M. W. Eysenck, & D. Warren-Piper (Eds.), *Student learning: Research in education and cognitive psychology*, (pp. 109-119). London, UK: SRHE & Open University Press.

Johnson, A. M., Ozogul, G., Moreno, R., & Reisslein, M. (2013). Pedagogical agent signaling of multiple visual engineering representations: The case of the young female agent. *Journal of Engineering Education, 102*(2), 319-337.

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.

Kirkegaard, C. (2016). *Adding Challenge to a Teachable Agent in a Virtual Learning Environment.* Licentiate Thesis in Cognitive Science, Linköping University. Linköping, Sweden: Linköping University Electronic Press.

Kirkegaard, C., Gulz, A., & Silvervarg, A. (2014). Introducing a challenging teachable agent. In P. Zaphiris, & A. Ioannou (Eds.), LNCS: Vol. 8523. *Proceedings of HCI International 2014* (pp. 53-62). Heidelberg, DE: Springer-Verlag.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback intervetions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284.

Koning, B. B., & Tabbers, H. K. (2013). Gestures in instructional animations: A helping hand to understanding non-human movements? *Applied Cognitive Psychology, 27*(5), 683-689.

Lea, M. R., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education, 23*(2), 157-172.

Lee, Y. J. (2017). *Effects of a Teachable Agent on children's noticing and reading feedback in a digital educational game.* Master Thesis in Cognitive Science, University of Vienna.

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., & Bhogal, R. S. (1997). The persona effect: Affective impact of animated pedagogical agents. In CHI'97: *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems* (pp. 359-366). New York, NY: ACM.

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.

Moll, H., Koring, C., Carpenter, M., & Tomasello, M. (2006). Infants determine others' focus of attention by pragmatics and exclusion. *Journal of Cognition and Development, 7*(3), 411-430.

Moreno, R. (2005). Instructional technology: Promise and pitfalls. In L. PytlikZillig, M. Bodvarsson, & R. Bruning (Eds.), *Technology-based education: Bringing researchers and practitioners together* (pp. 1-19). Greenwich, CT: Information Age Publishing.

Moreno, R., Reislein, M., & Ozogul, G. (2010). Using virtual peers to guide visual attention during learning. *Journal of Media Psychology, 22*(2), 52-60.

Mory, E. H. (2004). Feedback research review. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745-783). Mahwah, NJ: Lawrence Erlbaum.

Nicholls, J. G., Cobb, P., Wood, T., Yackel, E., & Patashnick, M. (1990). Assessing students' theories of success in mathematics: Individual and classroom differences. *Journal for Research in Mathematics Education,* 109-122.

Norman, D. (2013). The design of everyday things: Revised and expanded edition. New York, NY: Basic Books.

Orsmond, P., Merry, S., & Reiling, K. (2005). Biology students' utilization of tutors' formative feedback: A qualitative interview study. *Assessment & Evaluation in Higher Education, 30*(4), 369-386.

Ozogul, G., Reisslein, M., & Johnson, A. M. (2011). Effects of visual signaling on pre-college students' engineering learning performance and attitudes: Peer versus adult pedagogical agents versus arrow signaling. *In Proceedings of the 2011 ASEE Annual Conference & Exposition: T544A – Engineering Education Research in K-12* (pp. 14569-14670). Vancouver, BC.

Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning processes. Towards a wider conceptual field. *Assessment in Education: Principles, Policy & Practice, 5*(1), 85-102.

Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology, 92*(3), 544-555.

Pinsk, M. A., Arcaro, M., Weiner, K. S., Kalkus, J. F., Inati, S. J., Gross, C. G., & Kastner, S. (2009). Neural representations of faces and body parts in macaque and human cortex: A comparative FMRI study. *Journal of Neurophysiology, 101*(5), 2581-2600.

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

Risko, E. F., Richardson, D. C., & Kingstone, A. (2016). Breaking the fourth wall of cognitive science real-world social attention and the dual function of gaze. *Current Directions in Psychological Science, 25*(1), 70-74.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189.

Segedy, J. R., Kinnebrew, J. S., & Biswas, G. (2012). Supporting student learning using conversational agents in a teachable agent environment. In The future of learning: *Proceedings of the 10th international conference of the learning sciences (ICLS 2012): Vol. 2* (pp. 251-255). Sydney, Australia.

Segedy, J. R., Kinnebrew, J. S., & Biswas, G. (2013). The effect of contextualized conversational feedback in a complex open-ended learning environment. *Educational Technology Research and Development, 61*(1), 71-89.

Silvervarg, A., Kirkegaard, C., Nirme, J., Haake, M., & Gulz, A. (2014). Steps towards a Challenging Teachable Agent. In T. Bickmore, S. Marsella, & C. Sidner (Eds.), LNCS: Vol. 8637. *Proceedings of IVA 2014* (pp. 410-419). Heidelberg, DE: Springer-Verlag.

Taylor, J. C., Wiggett, A. J., & Downing, P. E. (2007). Functional MRI analysis of body and body part representations in the extrastriate and fusiform body areas. *Journal of Neurophysiology, 98*(3), 1626-1633.

Timms, M., DeVelle, S., & Lay, D. (2016). Towards a model of how learners process feedback: A deeper look at learning. *Australian Journal of Education, 60*(2), 128-145.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science, 10*(1), 121-125.

Van Mulken, S., André, E., & Müller, J. (1998). The persona effect: How substantial is it? In H. Johnson, L. Nigay, C. Roast (Eds.), People and Computers XIII: *Proceedings of HCI '98* (pp. 53-66). London, England: Springer.

Veletsianos, G. (2007). Cognitive and affective benefits of an animated pedagogical agent: Considering contextual relevance and aesthetics. *Journal of Educational Computing Research, 36*(4), 373-377.

Veletsianos, G. (2010). Contextually relevant pedagogical agents: Visual appearance, stereotypes, and first impressions and their impact on learning. *Computers & Education, 55*(2), 576-585.

Wiliam, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1053-1098). Charlotte, NC: Information Age Publishing.

Wotjas, O. (1998, September 25). *Feedback? No, just give us the answers. Times Higher Education Supplement*. Retrieved from: https://www.timeshighereducation.com/news/feedback-no-just-give-us-the-answers/109162.article

Paper VI

Review of feedback in digital applications – does the feedback they provide support learning?

Since the publication of this thesis, a revised version of this article is published as:
Tärning, B. (2018). Review of feedback in digital applications – does the feedback they provide support learning? *Journal of Information Technology Education: Research*, *17*, 247-283.
https://doi.org/10.28945/4104

Betty Tärning[1]

[1] *Div. of Cognitive Science, Lund University, Lund, Sweden*

Abstract. Digital applications ("apps") have become commonplace in Swedish schools. To fulfil their potential as support for learners, they need to accomplish several things. One is the focus of this paper: do the different kinds of feedback they provide support learning? For the paper a total of 242 apps (including subgames) were reviewed with respect to feedback provided. The results show that 78 % provide nothing but *verification feedback:* i.e., they tell the learner only whether an answer was correct or incorrect. A tenth of the apps provide the correct answer in response to an incorrect one. Only 12 % provide feedback to guide the learner toward the correct answer: e.g., by providing hints or explanations. Previous research has shown that feedback of the latter kind, sometimes called *elaborated feedback,* is more beneficial for learning than verification feedback or providing the correct answer when a learner is mistaken. Although half of the apps reviewed encouraged the learner in some way after a successful task, not a single one offered encouragement for good effort or partial progress. The encouraging feedback focused on the learners' abilities or intelligence, not on the task at hand. This is contrary to what is recommended from a learning-science perspective. If the goal is to design apps that truly support learning, designers need to revise their present approach.

Keywords. Digital applications, Verification feedback, Corrective feedback, Elaborated feedback, Encouraging feedback, Result feedback.

Introduction

Digital applications are today part of everyday school life; the number of educational apps has grown immensely over recent years. However, evaluation of the applications with respect to their effects on learning has generally been neglected.

As a consequence, it is difficult for teachers, parents and students to select which apps to use. In the Swedish context, there are webpages such as *Pappas appar* (http://www.pappasappar.se/) or *Länkskafferiet* (http://lankskafferiet.org/), that describe and score apps. The scoring is done by a group of parents, some of whom are teachers. The criteria for scoring vary but include whether the app is free to download, whether it has nice graphics, whether one´s own children or class seem happy using it. The guidance is of value, but it is hardly a systematic evaluation of how the applications affect learning.

This paper targets one aspect in which educational apps can be of high value for learners and teachers: the feedback provided to the learners.

Feedback is a consequence of performance. The learner receives a response to what she does: a response that, in most cases, tells something about the quality of her action or answer. For example, learners can be informed simply whether their answer was correct – so-called *verification feedback* – sometimes accompanied by the correct answer. Other types of feedback provide learners with more information: why an answer was correct or incorrect, or a small hint pointing toward the correct answer.

It is well-established that feedback has a large effect on learning (Black & Wiliam, 1998; Hattie & Timperley, 2007; Shute, 2008). Given this, it is striking how relatively little attention feedback has received when it comes to educational apps. Even though a teacher is still unbeatable when it comes to providing individualized feedback, digital systems have a potential that teachers do not. A teacher cannot simultaneously place herself beside every student to provide individualized feedback.

The goal of this paper is to examine the kinds of feedback provided in the educational apps used in Swedish schools today and discuss whether that feedback supports learning.

Background

Guiding the learner in exploration

A debate about the best way to learn has raged for decades. At one end of the scale, we find those who recommend free play, where the learning environment is not structured or designed in any purposeful way (Gray, 2013: in Hirsh-Pasek et al., 2015). At the opposite extreme we find those who believe only in highly structured instruction, where the teacher explains how things work and what the learner needs to know (Hirsh-Pasek et al., 2015). In effect, there are pros and cons with both approaches, but the best solution lies somewhere in between (Schwartz, Tsang, & Blair, 2016).

In *free* or *discovery* learning (Mayer, 2004), the learner explores an environment with little or no guidance. It is up to the learner herself to select, organize and integrate information. An advantage is that the student is free to construct her own learning experiences and is forced to take an active role in the learning task. At the same time, free exploration in a complex environment can generate high cognitive load, detrimental to learning (Sweller, 1994). The problem is especially relevant for novice learners, who lack existing frameworks into which to integrate the new information and who therefore must search the problem space more thoroughly. Kirschner, Sweller, and Clark (2006) argue that free learning makes a poor fit with our cognitive architecture.

The idea that learners should construct their own knowledge is reasonable; however, leaving learners without guidance in that endeavour is not. Many, if not all, learners struggle at one point or another when left on their own (Chi, 2009). A learner may focus on the wrong information from the beginning. Then it becomes hard – if not impossible – to straighten oneself out. The learner needs someone or something to guide her in the right direction again. A meta-review from Alfieri, Brooks, Aldrich & Tenenbaum (2011) concludes that direct instruction results in better learning than free play, but that the best learning is achieved through assisted discovery, with the instructor taking a supportive "back seat" role. Mayer (2004) argues that overwhelming evidence should make anyone sceptical of the benefits of pure discovery learning, with experimental evidence all pointing in the direction of having guidance when exploring a learning environment.

Learners need help not to treat new information as something just to memorize and recite. Rote memorization typically does not lead to so-called transfer. The goal should be to train students to be self-regulating learners, taking control of their own learning. Students need to be able to recognize when they understand and when they do not (Bransford, Brown, & Cocking, 2000).

Guiding the learner to exploit her learning potential

Guiding and being guided are everyday experiences: we observe how others do things and we act as role models to others – often without knowing. Lev Vygotsky was one of the first to recognize the importance of guidance to learning. That guidance might come from a parent or more experienced peer. In either case, someone who is more experienced helps someone who is less experienced move from their current *performance* level to their *potential* level: what the individual can do with help. Vygotsky (1980) calls the gap between these the *zone of proximal development*. At first the learner may need help at every step. Gradually she is able to perform some steps independently. Finally, she performs the entire activity with no assistance at all. Assisted performance guides the learner toward achieving things she could not achieve on her own (Gibbons, 2002).

Children are intrinsically motivated to participate in many kinds of activities, but they may not always see why certain activities are important. This is up to adults to explain. Vygotsky observed optimal motivation in children when asked to perform just above their present abilities (their present performance level). This means that a child can be motivated to learn more and make further progress if we provide them with *scaffolding*.

Scaffolding is a broad concept, encompassing all kinds of support provided to a learner in order to back her up in her learning activities. Feedback is a part of scaffolding – scaffolding provided in response to what the learner does.

Scaffolding

Wood, Bruner, and Ross (1976, p. 90) define scaffolding as:

> "*[A] process that enables a child or novice to solve a problem, carry out a task or achieve a goal which would be beyond his unassisted efforts. This scaffolding consists essentially of the adult 'controlling' those elements of the task that are initially beyond the learner's capacity, thus permitting him to concentrate upon and complete only those elements that are within his range of competence.*"

As with any scaffolding, the scaffolding is removed over time, allowing learners to accomplish the same task on their own. Since then scaffolding has become a well-researched topic and researchers have discussed which factors or ingredients are important for it (Van de Pol, Volman, & Beishuizen, 2010; Bransford et al., 2000).

Bransford et al. (2000, p. 104) list as factors[1] the following six activities or tasks as possible ingredients in scaffolding:

1. Making sure the learner keeps up interest in the task.
2. Reducing the number of steps needed to solve the task.
3. Motivating and directing the learner to pursue the goal.
4. Pinpointing the differences between the learner's current work and the desired outcome.
5. Reducing frustration and risk.
6. Demonstrating what an ideal performance looks like.

Feedback provided while a learner is working on a task seems to be a key ingredient for successful learning. Lepper, Aspinwall, Mumme, and Chabay (1990) further examined how expert tutors scaffold their learners. They conclude that experts tend to draw the learners' attention to an error, then provide a second chance at the solution – instead of offering corrective feedback. They usually ask the learner questions and avoid explicit directions. Fox (1991) reports a similar pattern.

Scaffolding and feedback intertwine. Scaffolding is the wider concept, including all forms of support given throughout the learning process. Scaffolding can also be a way of preventing a situation before it occurs (i.e. before the learner does something unwanted), and it can also be used to provide targeted support for particular learners or to deliver general instructions to a whole group of learners. On the other hand, feedback is information brought to the learner in response to something she has done.

Feedback

In general terms feedback can be said to be information coming back to a person in response to her performance, thoughts or ideas. According to Hattie and Timperley (2007), feedback can provide the learner with corrective information, provide alternative strategies, bring information to clarify ideas, provide encouragement, and provide the learner with correctness regarding their response.

Review studies by, Black and Wiliam (1998), Hattie and Timperley (2007) and Shute (2008) show that feedback can help learners to better achieve their learning goals. With that said, feedback *per se* does not ensure good performance. If learners can peek at what is designed as feedback before they have constructed their own

[1] Although Bransford et al. (2000) does not discuss feedback by name, it clearly is what he and his colleagues have in mind.

answer, there is little effective 'feedback'. The learner could merely copy-&-paste the answer without reflecting at all (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991). Feedback in the form of grades or other markers telling the learner how they are doing in comparison to others is usually not beneficial (Kluger & DeNisi, 1996; Wiliam, 2007; Butler, 1987). In contrast, feedback that contains information about the task and how to do it more effectively supports learning (Hattie & Timperley, 2007).

There are many different forms of feedback. First of all, we have *positive* and *negative* feedback. Positive feedback is information that tells the learner that there is no need for further learning or action; they already made a correct response. Negative feedback tells the learner that there is a discrepancy between her performance and her learning goal (Schwartz, Tsang & Blair, 2016). Unfortunately, negative feedback can have a threatening effect on learners. Learners who are *performance oriented* rather than *mastery oriented* are not likely to see negative feedback as something positive. The feedback tells them that they failed, which they may interpret as indication that their performance was not good enough and that they are not smart enough. Learners who are mastery oriented are more likely to see the feedback as a chance to improve their learning. To them, negative feedback is meaningful in that it helps them in their goal: to learn and make progress.

This article will concern both positive and negative feedback, but the focus will be on negative feedback.

The amount of information contained in feedback varies from "none" to "too much" (Schwartz et al., 2016). With respect to adequateness, both *amount* and *kind* of feedback vary between groups of learners. Novices generally need more information to correct their answers, compared to more knowledgeable learners.

Kulhavy & Stock (1989) write that good feedback should contain verification (*whether* the answer is right or wrong) and elaboration (*why* the answer is right or wrong). If a learner receives adequate feedback, this can reduce the uncertainty of where she stands in relation to the task. Uncertainty often takes attention away from the task itself. Adequate feedback can help reduce cognitive load and provide information useful for correcting misconceptions or inappropriate strategies (Shute, 2008). Good feedback should also be specific and timely: feedback should make clear the difference between learner performance and goal; and it should be delivered in reasonable time, so the learner can correlate the feedback to the task (Schwartz et al., 2016). Feedback should be understandable (Schwartz et al., 2016; Lea & Street, 1998; Higgins, Hartley & Skelton, 2001; Orsmond, Merry & Reiling, 2005), non-threatening (Schwartz et al., 2016), and reasonable (Brockbank & McGill, 1998). Lastly, the learner must be able to see the connection between the feedback and the task; otherwise she will not see the point in using the feedback (Orsmond et al., 2005; Wiliam, 2007; Segedy, Kinnebrew, & Biswas, 2013).

Verification feedback

A simple form of feedback gives the learner verification of whether her answer was correct or incorrect. I will call this *verification feedback* (sometimes also known as *knowledge of result* or *right/wrong feedback*). Verification feedback provides the learner with a sense of knowing whether she is on the right track: it can be more or less explicit. Examples are when the learner enters an answer, and the app indicates 'incorrect' or 'correct' via words or symbols, e.g. a red cross vs. a green checkmark or a sad vs. a happy face (figure 1). Often such markers are accompanied by a negative or positive sound. A correctly spelled word may be read out load by the software.

Figure 1. An example of direct verification feedback. The student has answered incorrectly, as shown by the unhappy red faces.

Implicit verification looks a little different. Say that a learner solves a crossword and tries to spell a word correctly by dragging a letter to one of the squares. If she chooses incorrectly, the letter is automatically removed from the square without explicit sounds or other types of signals indicating that the choice was incorrect. When the learner drags a letter to its right place, the letter stays, indicating that the choice was correct.

Studies on verification feedback are not unanimous concerning its usefulness. Pashler, Cepeda, Wixed, and Rohrer (2005) let their participants learn English translations of Luganda words (e.g., "*leero*" means "*today*"). After an initial training session, participants took a test and received: (i) no feedback, (ii) verification feedback, or (iii) corrective feedback: i.e. they were provided the correct answer if their answer was incorrect. Participants then took a second test and a third a week later. The corrective feedback led to the best performance on both the second and third test, while the verification feedback was no more useful than receiving no

feedback. Other studies conclude that it is generally better to provide learners with more elaborated feedback than just feedback in the form of 'correct' or 'not correct' (Bangert-Drowns et al., 1991; Pridemore, & Klein, 1995; Shute, 2008; McKendree, 1990; Moreno, 2004).

However, some studies do show a beneficial effect for verification feedback. Hanna (1976) conducted a study in which participants answered multiple-choice questions on science, mathematics, and social studies and then received (i) no feedback, (ii) verification feedback, or (iii) answer-until-correct feedback (something I refer to as 'trial-&-error', see section "*Trial-&-error*"). The results show verification feedback tended to be sufficient for high-performing learners, who were able to deduce the correct answer when informed that their answer was incorrect. Low-performing learners, on the other hand, were less likely to deduce the correct answer when informed that their answer was wrong. This group benefitted more from the answer-until-correct feedback.

Marsh, Lozito, Umanath, Bjork, and Bjork (2012) compared the effects of (i) no feedback, (ii) corrective feedback, and (iii) verification feedback. A total of 48 learners answered a series of general knowledge multiple-choice questions and took a test immediately the feedback was received with a second test after two days. The verification feedback was more useful than no feedback for improving on the final test, but corrective feedback was the most useful overall.

It appears that the benefit of verification feedback depends on the type of test as well as learners' ability level. The tests in Hanna's (1976) study were multiple-choice (though this was also the case for Pridemore and Klein (1995), and Marsh et al. (2012)). With such a test, indication of an incorrect choice tells more compared to a free-recall test: after all, it is possible to exclude at least one of the answers.

When it comes to learners' ability level, the study by Hanna (1976) showed that high-performing learners receiving verification feedback were likelier than low-performing learners to deduce the correct answer. Similarly, Schwartz et al. (2016) point out that for learners who already have basic knowledge within a specific area, simple verification feedback regarding whether an answer is correct or not, or whether a certain choice is adequate or not, can be useful. Very knowledgeable learners completing a familiar task may only need verification feedback. However, for novices the sweet spot resides in more informative feedback.

Finally, the benefit of verification feedback depends on what other forms of feedback it is being compared to. Compared to no feedback, it at least provides a hint that the learner is heading in the right direction.

Trial-&-error feedback

One problem with verification feedback is that it is often accompanied by an opportunity to use trial-&-error. In principle, the learner could keep entering one answer after the other, without the need to put any thought into it. If the teacher only gets to see the final correct answer, she will have no information how the student got there. In *Josefins skolvärld* (English: "*Josephine's School World*", figure 2), the learner is supposed to click the number corresponding to the number of ladybugs in the picture. No matter how many times the learner clicks the wrong answer, she can continue until she gets the correct answer. Since each correct answer is rewarded by a point, there is no way to tell from looking at the scores whether a learner solved the task on the first, third, or twelfth trial.

Figure 2. An example of verification feedback where it is possible to use a trial-&-error strategy.

If systematic trial-&-error helps a learner move forward at low cost of time and effort, it is not hard to understand why the strategy can be the learner's first choice. From the perspective of the teacher though, trial-&-error behaviour usually falls under the heading of 'gaming the system' defined by Baker et al. (2006, pp. 392-393) as "*attempting to succeed in an educational environment by exploiting properties of the system rather than by learning the material and trying to use that knowledge correctly*"

Returning to *Josefins skolvärld*, the design offers an opportunity to systematically click each answer until the right one is clicked, and the learner scores a point. The trial-&-error strategy in this case is low cost in terms of effort and time spent – and there is no decrease in scores for errors. In fact, it can take less time to finish a task by repeatedly clicking on the different alternative answers than by really taking one's time and thinking the answers trough. This is particularly true for learners who

are about to learn the content in question and generally need more processing time (that is, learners who are in their proximal zone of development and clearly could benefit from some more instructions and help).

Of course, the degree to which trial-&-error pays off is related to the design of the app. I will distinguish between 'low-cost trial-&-error', 'risky trial-&-error', and 'time-consuming trial-&-error'.

Low-cost trial-&-error is described in the example above. The learner can click randomly without risk of losing lives or scores.

Risky trial-&-error involves situations with a limit on time or the number of trials ("lives") allowed. If the learner is unlucky, she will not move on in the game and will have lost time, 'lives', and perhaps a chance to reach a high score.

Third, *time-consuming trial-&-error* involves situations in which trial-&-error behaviour is likely to be very time consuming. Consider a learner who cannot read and write who must spell a word, with all the letters in the alphabet at her disposal. It is possible in principle to take each letter and try it out – but solving the task this way will take a very long time. Other tasks are impossible to solve using this strategy – at least, not without a great deal of luck. Say, that a learner is supposed to find the sum 23 + 42, with no alternatives presented. A systematic attempt with all the numbers from one upwards will be (almost) impossible without thinking the answer through.

The use of trial-&-error strategies has repeatedly been shown to correlate negatively with learning (Aleven & Koedinger, 2001; Baker, Corbett, & Koedinger, 2004; Baker, Roll, Corbett, & Koedinger, 2005; Walonoski & Heffernan, 2006; Baker et al., 2006).[2]

Baker, Walonoski, Heffernan, Roll, Corbett, and Koedinger (2008) describe an animated agent designed to reduce the incentive to game the system: e.g., through trial-&-error. When such behaviour is detected the agent displays increasing levels of displeasure. If the learner by chance arrives at the correct answer anyway, the agent gives her a set of supplementary exercises covering the material that she has just skipped over. The results show that gaming-the-system behaviour decreased overall, while learners who persisted in using such strategies increased their learning through the supplemental exercises they were given.

[2] Even though these studies have been carried out with middle-school students, there is reason to believe that the same applies to younger children.

Corrective feedback

We have already learned that corrective feedback can be more beneficial than verification feedback (Pashler et al., 2005; Marsh et al., 2012). In addition, Clariana (1990) compared the effects of verification feedback (where a trial-&-error strategy was possible) and corrective feedback on 32 low-performing learners. The results show that corrective feedback had a significantly greater effect on performance than verification feedback. Furthermore, Phye, and Sanders (1994) showed how more specific feedback in the form of providing the correct answer improved the performance on a retention task compared to more general feedback.

Corrective feedback provides the learner with more information than negative verification feedback in that it also provides the correct answer. Providing the learner with the correct answer can happen immediately after an incorrect choice, or it can be delayed until the end of the session. An example of corrective feedback can be seen in *Minilobes*; here the learner is asked to find the lower-case letter "*a*" (figure 3, left panel). When the learner clicks the erroneous letter (lower-case "*c*"), the app directly says "*cee*" and then shows the correct answer (figure 3, right panel).

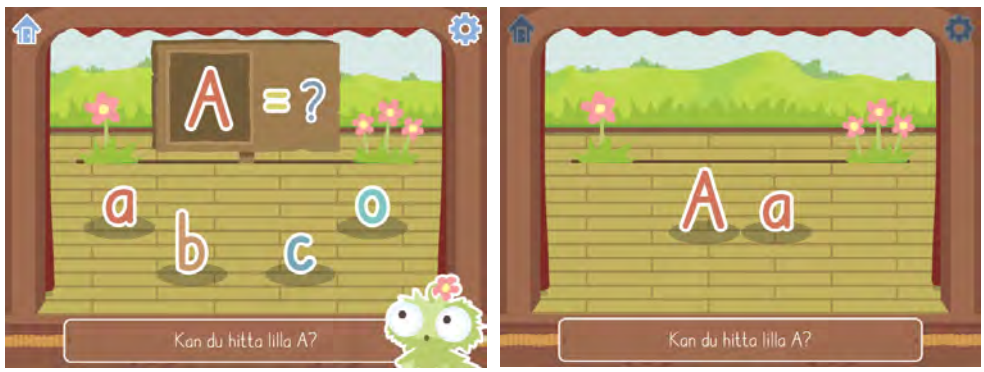


Figure 3. An example of corrective feedback: providing the learner with the correct answer after an incorrect one.

The effects of corrective feedback again depend on what other forms of feedback are being compared, as well as the task and ability of the learner. Hattie and Timperley (2007) argue that this simpler feedback is most powerful when it addresses faulty interpretations rather than total lack of understanding. Finn and Metcalfe (2010) found that corrective feedback seems to be beneficial for immediate testing, but not for delayed testing. Moreno and Mayer (2007), on the other hand, argue that novice learners learn better with explanatory feedback as compared to corrective feedback.

At the same time as corrective feedback provides the learner with more information than negative verification feedback, a potential drawback is that the learner, upon receiving the correct answer, just memorizes it without understanding. Rote learning is not a bad thing *per se;* but, in many cases, it is important first to have an understanding of what one is learning. Consider a child who learns that '$2 \times 3 = 6$' but has little knowledge what the numbers mean. She does not understand that '$2 \times 3 = 3 + 3$', which is the same as '$2 + 2 + 2$', and so on. Learning the multiplication table by heart only takes one so far.

Elaborated feedback

Elaborated feedback refers to any feedback that provides learners with more meaningful information. It comes in different forms and at different levels. Shute (2008) writes of elaborated feedback that it can choose to address the topic or the response; it can discuss specific errors, provide worked examples, give gentle guidance, or explain why a response was wrong and indicate the correct answer.

Elaborated feedback is generally associated with better learning (McKendree, 1990; Bangert-Drowns et al., 1991; Pridemore & Klein, 1995; Moreno, 2004; Shute, 2008). Further, Bangert-Drowns et al. (1991) argue that feedback is significantly more effective when it provides details of how to improve the answer instead of just indicating whether the learner´s work is correct or not.

Finn and Metcalfe (2010) conducted three experiments, comparing four situations: (i) corrective feedback, (ii) scaffolded feedback, (iii) answer-until-correct feedback, and (iv) minimal feedback, (participants given one additional try when their first answer was wrong). The scaffolded feedback offered small hints guiding the learner toward the final answer, step by step: first providing the first letter in a target word, then the second letter, and so on.

The first experiment showed that the corrective and scaffolded feedback gave the best test scores upon immediate testing. In the second experiment participants not only were tested immediately but also after 30 minutes. The third experiment was exactly the same but with a delay of one day. Corrective feedback was best again with immediate testing, but scaffolded feedback gave best results for the delayed tests. Minimal feedback consistently produced the weakest results. The experiments show that when a learner just has a short time to correct an error, corrective feedback can be the best option, but that scaffolded feedback works best for long-term retention.

Facilitative feedback

Facilitative feedback means that the learner is offered a comment or suggestion to help her find the right solution. In *Läskod* (English: "Access code", figure 4), the learner is to practice spelling. The task is to spell the word "*godis*" (English: "*candy*"). After a few mistakes, a question mark appears and the student is allowed to see the correct spelling briefly before it disappears again.

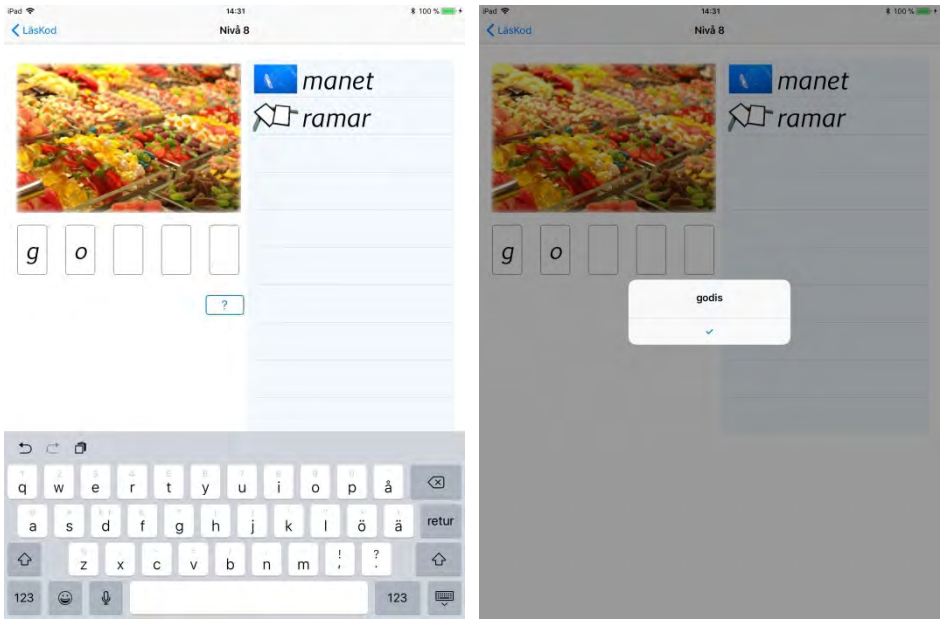


Figure 4. An example of facilitative feedback, showing the learner how to solve the task if she is struggling.

The problem with just showing the answer as a hint is that is allows the learner to copy-&-paste.

Explanatory feedback

If facilitative feedback provides information how the learner can solve the task, explanatory feedback provides more; e.g., why the answer was correct or not. With explanatory feedback the learner can build a deeper understanding of the task at hand and the foundations upon which to structure future tasks.

In the Swedish app *Särskrivning* (English: "*compound words written with a space between*" figure 5)[3] the learner is supposed to click on the expressions that are misspelled. In the left panel, the learner mistakenly clicks on an expression that is spelled *correctly* "*gröna bönor*" (English: "*green beans*"): the program responds "*No, that was incorrect. The beans are described as green*." In another task (right panel) the learner is asked to "*click on the right alternative"* for a common Swedish surname. The correct answer is "*Pettersson*". When the user clicks on "*Petters son*" (English: "*the son of Petter*"), the game responds "*Petter is a common Swedish first name – you just clicked on his son.*".

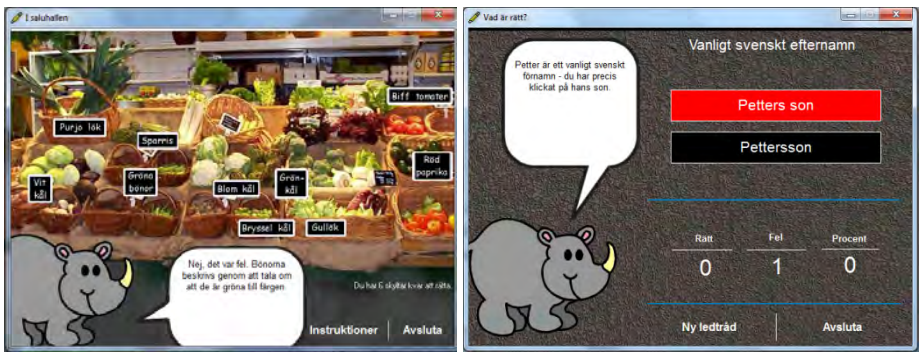


Figure 5. An example of explanatory feedback, where the learner receives information about why their answer was correct or incorrect. In the left panel the rhino says: "*No, that was incorrect. The beans are described by telling that they are coloured green*." In the right panel the rhino gives the explanatory feedback "*Petter is a common Swedish first name – you just clicked on his son.*" when the learner makes a mistake.

Moreno (2004) studied whether explanatory or corrective feedback worked best in a discovery-based learning environment, where novice learners were to design a plant capable of surviving under different weather conditions. Learners receiving explanatory feedback produced higher scores on a transfer test. Moreno argues that the explanatory feedback helped novices by decreasing their cognitive load, noting that benefits were found for cognitive but not affective outcomes; e.g., motivation or interest.

Implication feedback

Actions and choices have consequences. If one miscalculates how many tablespoons of yeast one needs to add to dough, one will experience first-hand that the dough

[3] In Swedish, 'compound words' are generally written without hyphen or space between them – and a misspelling can completely change the meaning, for example: "*sjuksköterska*" means "*nurse*" (in English), while "*sjuk sköterska*" changes meaning to "*sick nurse*" (in English).

does not rise well. If a child is told to give apples to each of four horses but has only three and starts giving an apple to each, she discovers that one horse is left without an apple. The child learns more than just that her solution was incorrect: she may understand that three apples are too few for four horses but not *far* too few.

Such *implication feedback* (Blair, 2009), found in the mathematics game *Magical garden* (figure 6) is meant to help pre-schoolers develop their understanding of number sense (Husain, Gulz, & Haake, 2015; Haake, 2018). Together with her teachable agent (a pedagogical agent whom the learner teaches at the same time as learning for herself), the learner creates a magical garden by collecting water drops, which she receives by solving math problems. In one game the learner is to help a hungry chameleon with weak eyesight catch ants. The learner can see if the chameleon´s tongue reaches too low or too high (or catches the ant if the answer is correct).

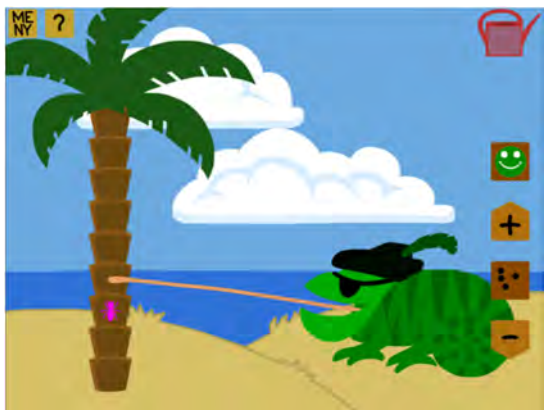


Figure 6. An example of implication feedback: the learner sees that her answer is incorrect by watching the chameleon aim too high and so miss its food.

Critter Corral (figure 7), aims to help pre-schoolers develop concepts for the numbers one through ten. In one exercise the learner's task is to fix a chair by choosing the correct leg size. If the learner choses a leg that is too short or too long (or correct) this will be reflected in the game as well as stated by the speaker voice. The thought behind this type of feedback is that the learner should develop a sense of magnitude and be provided with some guidance on how to revise their attempts (Blair, 2013). This in contrast with verification feedback, where the learner has to guess in which direction to go in order to fix a mistake. With implication feedback, in contrast, the learner is scaffolded by a hint.

Figure 7. An example of implication feedback from *Critter Corral*.

Feedback focusing on the learner

The kinds of feedback discussed so far all concern the task at hand and provide the learner with information about the task, the correctness of the task, and/or information about how to improve their solution to the task. In contrast, *encouraging feedback* and *result feedback* is information that concerns the learner rather than the task or how it can be solved.

Encouraging feedback

One role of feedback is to motivate the learner to continue with a task. *Encouraging feedback* is supposed to do this. Usually such feedback is displayed as visual and/or auditory encouragements such as cheering, clapping, rising stars or balloons (or something happy) or via text or voice expressing how well the learner does (for examples of both kinds, see figure 8).

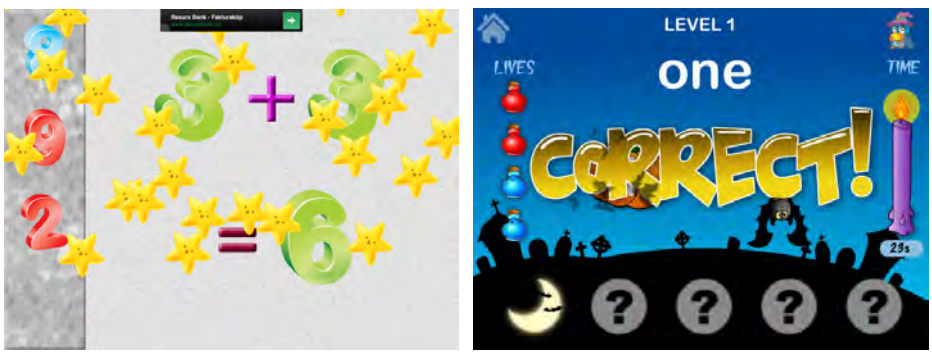


Figure 8. Two types of encouraging feedback.

This type of feedback contains little (or no) task-related information, and the effects of it are rarely converted into more engagement, commitment to the learning goals, an enhanced self-efficacy or understanding of the task (Hattie & Timperley, 2007). Feedback about the self can even be seen as meaningless, and meta-analyses on teacher praise have found small, if any, associations with learner achievement (Wilkinson, 1980; Kluger & DeNisi, 1996).

The problem with this type of feedback is that it targets the learner as a person, for example, by saying "*good girl*" or "*you are brilliant*". It does not say anything about *what* the learner did well (and perhaps less well). It does not contain any information about the learner's effort involved in trying to solve the task or in managing to solve it more effectively. Hattie and Timperley (2007) point out that the highest effect sizes with respect to learning were found in studies that involved learners who received feedback about the task and how to solve the task more effectively. Praise, rewards and punishments were associated with much smaller effect sizes. Feedback about the self, such as "*you are a great learner*", cannot really, as the authors point out, help the learner to proceed in her learning.

This is not to be confused with how praise regarding achievement and learning can sometimes assist in enhancing self-efficacy, which may, in turn, influence achievement (Hattie & Timperley, 2007). Nicol and Macfarlane-Dick (2006) further argue that praising effort and strategic behaviour leads to higher achievement compared to praising ability and intelligence. This is also supported by Black and Wiliam (1998), who recommend avoiding feedback that draws attention away from the task and towards self-esteem, since this can have a negative effect on attitude and performance. Learners are much better served by praise for the efforts they invest in a task than by praise earned by their innate abilities (Dweck, 2000). Praising only children's intelligence can lead them to avoid tasks in which they could potentially learn something due to the fear of looking stupid or loosing face (Dweck, 2000; Gunderson, Gripshover, Romero, Dweck, Goldin-Meadow, & Levine, 2013).

We should think twice before praising the learner (at least without thinking about what we are praising), which does not mean, however, that learners do not *like* to be praised – they most often do (Sharp, 1985; Burnett, 2002; Elwell & Tiberio, 1994).

Result feedback

A common form of feedback in school is results, such as a student's score or grade. This is also quite common in apps, and I will refer to this as *result feedback*. Just as with encouraging feedback, this type of feedback does not provide the learner with information about the learning process or a how the learner could improve. It is a mere evaluation of how well the learner has performed during, for instance, a game

session. Notably, this kind of information can be misguiding for someone who looks only at the result and who may, for instance, not be able to tell whether the presented result is an outcome of a low-cost trial-&-error strategy.

If the goal is for a learner to evaluate her own progress, such feedback can be a good parameter to use, but when used as a tool for comparing the performance of different learners, it may lead to stress and negative feelings for some students (but for some who like to compete it might also be encouraging). When result feedback is used in order to compare learners, the focus turns more to the learner than to the difficulties in a task and efforts to improve. Simply put, receiving a grade or a result can risk making the learner focus on the wrong thing, and if no other feedback is given, a simple result does not tell the learners how they could improve. This, in turn, has been shown to have a negative effect on motivation (Harlen & Deakin Crick, 2003; Craven, Marsh, & Debus, 1991; Butler, 1987).

Feedback in educational software

There is not a large amount of research on the types of feedback used in educational software. However, there are some recent studies that review educational software more broadly, and some consider feedback, even when the concept as such is not used.

Cherner, Dix, and Lee (2014) put forward a framework for how to choose educational apps based on their purpose, content, and value. Some researchers have examined different categories of apps (Handal, El-Khoury, Campbell, & Cavanagh, 2013; Highfield & Goodwin, 2013). Larkin (2015) evaluated apps for mathematics, analysing how many of them provided the learner with conceptual knowledge (i.e. information that involves understanding related to the meaning of mathematics), procedural knowledge (the ability to follow a set of sequential steps to solve mathematical tasks) and declarative knowledge (information that the learner retrieves from memory without hesitation). Highfield and Goodwin (2013) reviewed the pedagogical content within the most popular apps in Australia, UK, and the USA, and found that 74% of all apps had elements of 'drill and practice', tasks that require minimal cognitive investment on behalf of the learner. These types of tasks usually require minimal cognitive investment and frequently use extrinsic rewards. From this review it can be concluded that more apps need to be developed that also focus on children's ability to develop as self-regulatory learners, who do not only memorize things by heart without understanding. In Hirsh-Pasek et al. (2015) offers a way to define the potential educational impact of current and future apps. Along the same line, Sjödén (2017) evaluated what factors are important when evaluating an educational app, and here feedback is mentioned as one of the cornerstones.

The present study focuses on the types of feedback that are represented in apps commonly used in Swedish schools today. Based on findings from Highfield and Goldwin (2013), Blair (2013), Sjödén (2017), and my own experience with apps, I predicted that few of the apps would contain elaborated feedback – which according to the literature would be most appropriate in order to enhance learning.

Method

I distributed an email to different schools around Sweden asking them to send a reply regarding which apps they used. The email was distributed to approximately 40 schools, and 14 of them replied. The answering schools were distributed from Luleå in the north of Sweden, to Ystad in the south. The target software was apps used for children in primary school.

Number of apps reviewed

In total, I received the names of 164 different apps, of which several were used at more than one school. Since I had not asked for apps targeting a certain subject, the received apps targeting various subjects such as mathematics, Swedish, programming, learning the clock, biology, and geography.

In total, I removed 61 apps out of the 164; 25 because I considered them to be non-educational in that they did not cover any subject in the curricula (for example the camera, *Gmail*, the calculator, *iMovie*, etc.), 7 others because they did not give the learner any room for improvement (an example of this category was an app in which the learner could practice how to do different geometrical shapes with digitalized rubber bands). Seven apps were categorized as more general tools for the learner and/or teacher (such as an app translating a Swedish word into English). In addition, there were 24 apps that I found too complex to evaluate, since they were parts of a larger learning environment or because they could not be found in *App Store*.

Of the 103 apps several contained subgames, which in this study are treated as individual apps, since they touch upon different subjects or have a different gaming idea. When reviewing these, 29 were removed, since they did not contain any activity where the learner could do something categorized as wrong.

For example, in *Bugs and bubbles* (figure 9, left panel) the learner's task is to collect all green dots by tilting the tablet in different positions. The learner may miss a dot at one trial, but then she can just tilt the tablet so that the ball takes another round (preferably past the green dot). In *Siffermix 1* (figure 9, right panel) the learner is supposed to click any number; the number is then represented by a set of objects.

Figure 9. Examples of apps that were not included in the review, since the learner cannot do anything categorized as wrong.

One additional app was removed from the sample, since it did not provide the learner with any feedback. Having thus removed 30 subgames, I was left with 242 apps (including subgames)[4].

I played each app for as many times as it took to grasp the gist of it and establish what types of feedback were present. While playing, I was consciously trying to make as many mistakes as possible to see in what way the app would provide me with feedback and possibly guide me towards the correct answer. In addition, I also tried different trial-&-error strategies to see if any of my categorized strategies could be used. Approximate gaming time was between 20 minutes and 1 hour per app.

Measurements

Feedback categories

The reviewed feedback categories were the following:

Verification feedback: Feedback verifying whether the learner's answer was correct or not. Both implicit and explicit verification were categorized as belonging to this group.

When the app also provided an opportunity for learners to use a trial-&-error strategy, it was categorized as one of the following:

- *Low cost:* The learner can retrieve the correct answer without reflection and merely by clicking, without losing scores, 'lives', or time.

[4] Some apps were available as both a free and a paid version; in 13 cases there were both a free and a paid / commercial version, and in 8 of these the free version was used (this was in cases were the judgment was made that a payment would not bring anything extra).

- *Risky:* The same principle as ‘low cost’, but with a penalty in terms of time, life, or score limits. With some luck, the learner can succeed just by clicking and will finish the task without any drawbacks. But if unlucky, the time or ‘lives’ will run out or her scores will be reduced before the task is completed.
- *Time-consuming:* tasks that in theory can be solved by systematic try-outs, but also tasks that are (almost) impossible to succeed with by systematic try-outs.

Corrective feedback: The app provides the correct answer when entering an erroneous one.

Elaborated feedback: Feedback that gives the learner more information on how to solve the task or why their answer was incorrect or correct. Different amounts of information, ranging from giving a small hint to giving the whole answer or more explanatory feedback in text format.

- *Facilitative feedback:* The learner receives a small hint about how the task is supposed to be solved.
- *Explanatory feedback:* The learner receives more information regarding a possible erroneous or right answer.
- *Implication feedback:* The results or implications of the learner’s answer are presented in some form.

Encouraging feedback: Feedback directed at the learner herself after the completion of a task, generally in the form of praise. Visual elements, such as falling stars, auditory elements, such as clapping or cheering, as well as written encouragements belong to this category.

Result feedback: Gives the learner a ‘grade’ based on their performance, usually the proportion of correct answers.

For an overview of all reviewed feedback types, see figure 10.

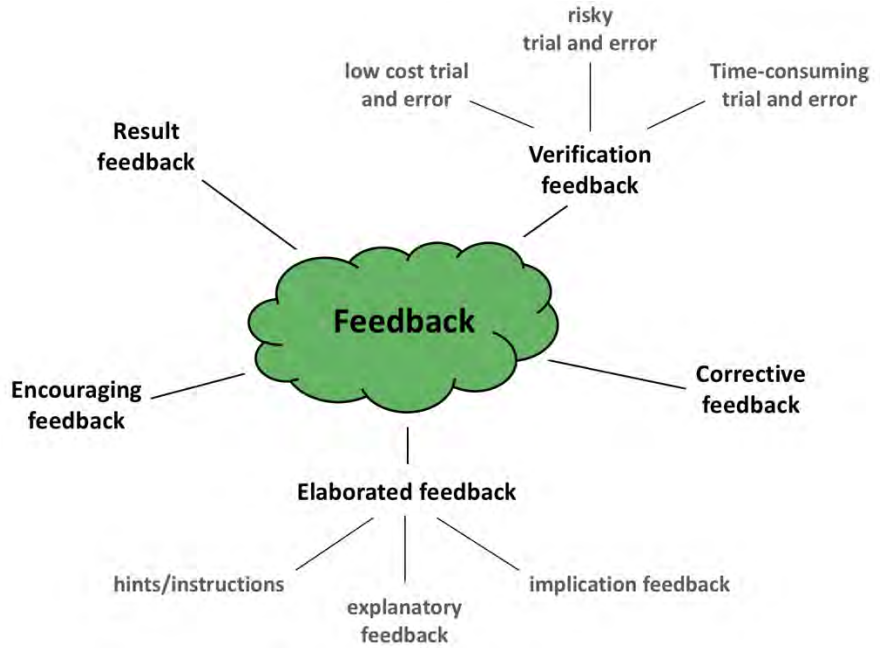


Figure 10. Reviewed feedback types.

This review only concerns feedback and does not look at other factors that could influence learning. This means that an app, in this review, might be portrayed as less satisfactory concerning the feedback it provides or how feedback is provided, while it may still have other, more positive features. For example, an app might help the learner to visually represent a number that she is supposed to calculate (figure 11). Here, the learner is supposed to add "*3 + 2*", but instead of only showing the numbers, which a learner at a given stage might find abstract and have a hard time grasping and finding meaningful, the number is also represented with cookies in different colours. Visualizing a number can make it easier for some learners to do the calculation, and instead of abstract numbers there are concrete objects to handle.

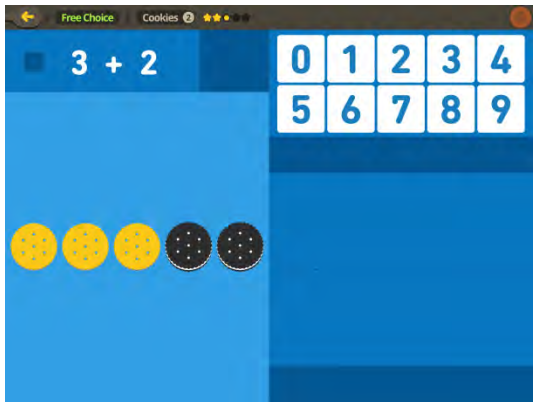


Figure 11. An example of visual representation in the app *Todo math – cookies*.

Results and discussion

Out of the 242 apps reviewed, 189 (78%) contained verification feedback only; that is, the app provided the learner with information as to whether an answer was correct or incorrect. Twenty-five apps (10%) contained corrective feedback in that the right answer was displayed after the player had provided an incorrect answer. Further, 31 apps (12%) contained information in the form of elaborated feedback (facilitative, explanatory, or implication feedback), see figure 12.

Figure 12. The proportion of apps providing each of three types of feedback.

These results confirm my prediction that few apps would contain elaborated feedback, and the percentage was as low as 12%. From what we know from the

literature, more elaborated feedback is preferable if learning is to be supported (McKendree, 1990; Bangert-Drowns et al., 1991; Pridemore & Klein, 1995; Moreno, 2004; Shute, 2008). Most of the apps reviewed do, thus, not fulfil that requirement.

Verification feedback

A potential pitfall with verification feedback is that it may encourage different trial-&-error strategies, which the learner can use in order to complete a task. I therefore analysed possibilities of using low-cost trial-&-error, risky trial-&-error, and time-consuming trial-&-error (figure 13).

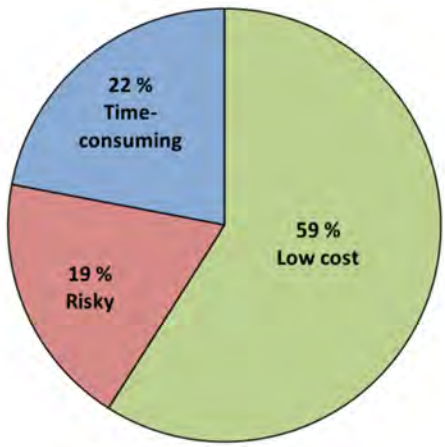


Figure 13. Verification feedback and its division in different trial-&-error strategies.

Low-cost trial-&-error feedback

Fifty-nine percent of the apps that only included verification feedback were designed in such a way that low cost trial-&-error strategies could be used. That is, more than half of the reviewed apps made it possible for the learner to get all answers correct without actually having to pay attention to the task at hand. Typical examples are provided in figure 14. In the app *ABC-klubben*, the task is to drag the card that starts with the letter M to the empty square (figure 14, left panel). When the wrong card is drawn (here, the picture of a fire), the card simply returns to its starting position, and a 'negative' sound can be heard. The learner can then try again for as many times as she wants until she chooses the correct card, and the app confirms the correct answer by saying "*monster*".

Another typical example can be seen in *Lola's Mattetåg* (figure 14, right panel), in which the learner is supposed to click on the number three. When clicking the

incorrect answer, Lola (the panda) shakes her head, and the erroneously clicked answer is highlighted with a red ring. Again, the learner can try for as many times as she wants until she gets it right.

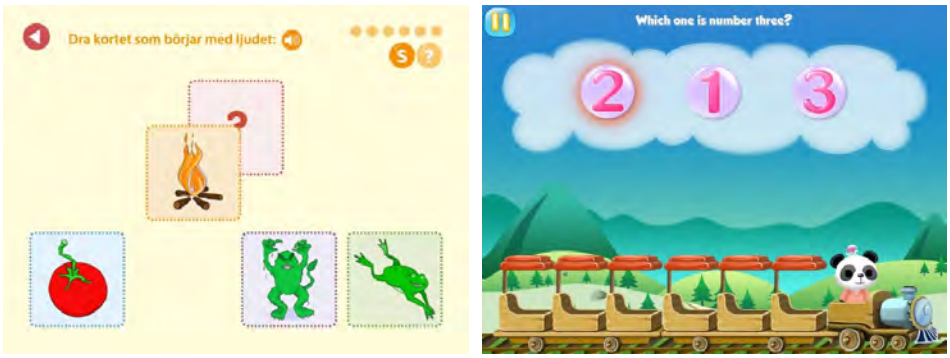


Figure 14. Examples of apps in which the learner can use a low-cost trial-&-error strategy.

As already mentioned, low-cost trial-&-error strategies are not beneficial for learning (Aleven & Koedinger, 2001; Baker et al., 2004, 2005, 2006; Walenoski & Heffernan, 2006). The learner could very well be thinking about other things and not the task at hand, but still gain a high score. Clicking without paying attention to the task is not likely to lead to any good learning. However, from an outsider's perspective, a high score and a fast response time indicate that the learner is good at the task, something that might be false.

Risky trial-&-error feedback

In 19% of the apps that contain only verification feedback the learner can apply a risky trial-&-error strategy. This means that with a little luck the learner may provide the correct answer by chance. But there are also other elements, such as time spent and 'game lives' lost, that need to be taken into account. The chances to get a high score or a fast time by just guessing are lower than in low-cost trial-&-error, and if the learner replies by chance every time, the chances of her getting a high score or moving on to the next level are slim, since every incorrect answer is 'punished' in some way such as, for example, losing a life or points or not levelling up.

The left panel in figure 15 (*Happi läser*) shows an example in which the learner needs six watermelons (as can be seen at the bottom of the left panel) in order to move on to the next level. Typing in an incorrect answer provides a 'negative' sound and the pictures disappear and are replaced by new ones. That is, clicking an incorrect picture will not give the learner any disadvantages, but at the same time she will not be able to move on to the next level unless she provides six correct answers.

Another example can be seen in the app *Math king* (figure 15, right panel), in which the learner is supposed to sum up the numbers represented by the fingers. The learner only has three 'lives', and each time an incorrect answer is provided, she also loses score, which can be used to climb a 'career ladder'. Here the learner has more to lose compared to low-cost trial-&-error apps (at least if the learner wants to progress within the app).

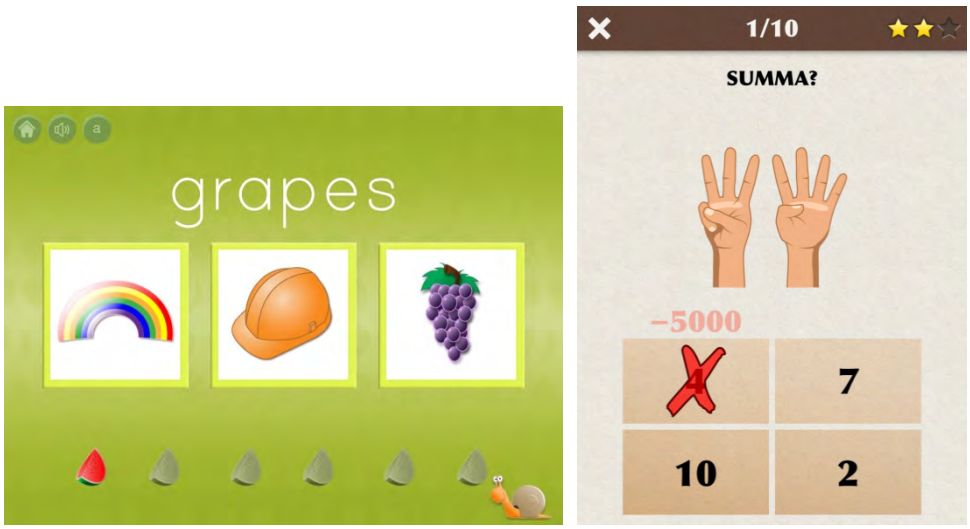


Figure 15. Two examples of apps where a risky trial-&-error strategy can be used.

Time-consuming trial-&-error feedback

The last 22% of the apps that contain only verification feedback were apps in which the task could be solved by using a time-consuming trial-&-error technique and also apps in which it is (practically) impossible to solve the task in this way if the learner has no idea of how to solve it. Examples of this can be seen in figure 16.

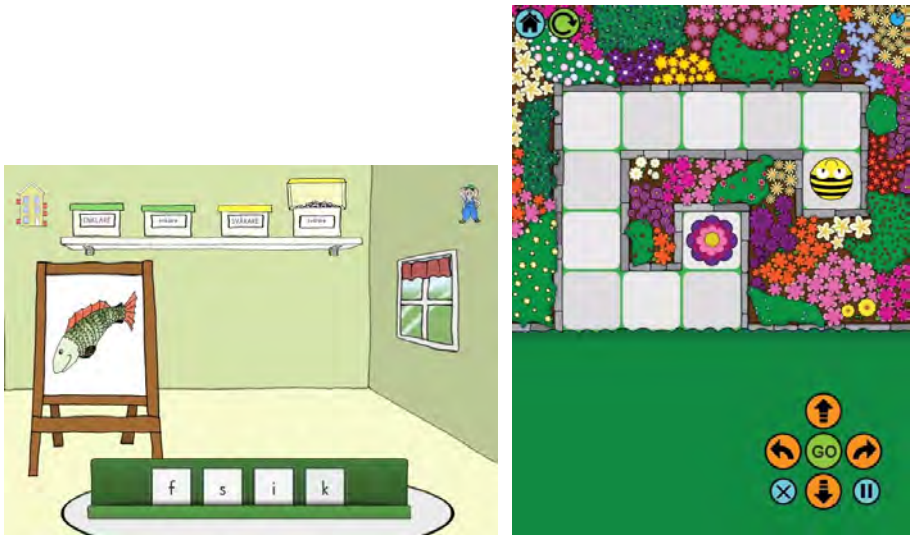


Figure 16. Two examples of apps where a time-consuming trial-&-error strategy can be used.

In *Bornholmslek – Bygga ord* (figure 16, left panel) it is possible to find the correct answer by using the strategy of trying out every possible combination of letters provided. Here the learner is supposed to spell "*fisk*" (English: "*fish*"). Even though it is time-consuming, it is not impossible to try different combinations of letters until it is correct. The learner is also provided with feedback in the form of sounds telling her how a certain letter is pronounced. By using this information, it is possible to find the correct solution without knowing it from the beginning.

In the app *Bee-Bot* (figure 16, right panel) it is, on the other hand, practically impossible to find the correct solution, unless the learner has an idea from the beginning about how to solve the task. The learner is here supposed to guide the bee to the flower by using programming commands. In addition, the difficulty of the problems increases considerably, in that the learner has to keep every command in their working memory – there is no visualization of commands already ordered.

In tasks like this the learner would probably be helped by, first of all, receiving some type of command tracing, so that they would not have to keep their commands in their head. In addition, if the learner could also trace their commands in combination with the bee's path, it would make it more visible for the learner where a possible error in their coding had occurred. What is also troublesome is that the bee always starts from the beginning of the commands, if not, it would be possible for the learner to take it one step at a time, like they can in another app called *Lightbot* (figure 17).

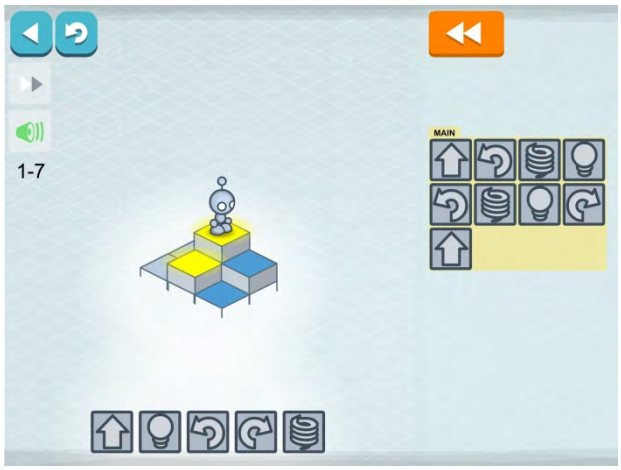


Figure 17. An example of visual feedback showing the learner which commands she has pressed.

In Lightbot the learner receives better scaffolding to trace their programming, since the commands are visualized and don't need to be kept in one's working memory. It can still be difficult for the learner to see exactly where their programming went wrong. If it was possible to slow the robot down even further, as well as making the robot walk at the same time as the corresponding command was lit up, the feedback would be even clearer, at least at the beginning, when the task might still be new and challenging.

Common for all apps that contain only verification feedback is that a learner who does not know the correct answer from the beginning can solve the tasks, yet still be left with knowledge gaps. They can also be stuck on a task without knowing how to fix it, since no further feedback is provided. For example, in Lightbot again, if the learner cannot figure out by testing how to make the robot light all the blue boxes, this can cause frustration, since there is no help available for each step the learner should take in order to reach the goal.

If the aim of the app is to teach something and for a learner to develop knowledge, skills, or understanding she did not have before, there should be some feedback helping the learner if she needs it. If the goal of an app, instead, is to test knowledge, understanding, or skills that one believes are in place, the demands on the app are different.

Corrective feedback

Twenty-five apps contained more information than only verification feedback in that they also provided the *correct* answer when the learner typed in an *incorrect* one. Two examples are shown in figure 18.

In the app *Math bingo* (figure 18, left panel) the correct answer, in this case "*1 + 1 = 2*", is shown after the learner has provided an incorrect answer. The learner then receives a new task to solve, here "*7 + 4 = ?*" In the app *Geoexpert*, the learner is shown a flag (right panel in figure 18, top left corner) and the name of the corresponding country. The learner's task is to find the country among those marked on the map and click it. After two incorrect answers the correct country is circled. An addition in this app is that if the learner clicks an incorrect country (such as clicking at Brazil) the app displays Brazil's flag as well as types the name 'Brazil'. Hereby the app provides the learner with information that she may use later.

Figure 18. Two examples of corrective feedback, in which the learner is provided with the correct answer after an incorrect one.

Another example can be seen in the app *Räkneapan* (figure 19) in which all numbers that the learner replied incorrectly to are summarized at the end of the game. By getting all the incorrect answers summarized at the end, the learner is given the opportunity of studying them further.

Figure 19. Corrective feedback shown at the end of the game.

Although not many explanations are provided in the analysed apps, the learner is not left with a complete blank as to what was wrong with their answer, since they are provided with the correct one. By being presented with the correct answer, they gain some information that may be used for learning.

Elaborated feedback

Twenty-nine out of 242 apps (12%) contained some type of elaborated feedback: facilitative feedback, explanatory feedback, or implication feedback.

Facilitative feedback

Most of the elaborated feedback has the form of being facilitating, providing the learner with some type of hints on how to solve the task at hand. Out of the 29 apps that contained elaborated feedback, I categorized 23 as providing facilitative feedback. An example of an adequate or useful hint provided to a learner can be seen in figure 20, showing the app *Mattebageriet 2*. If the learner is not able to solve the task "*12 + 20*" (figure 20, left panel) there is a lightbulb in the upper left corner, which can be clicked, and the app then provides a hint that asks the learner to count "*How many single cookies are there on the plate?*" (figure 20, right panel). This hint provides the learner with information that is useful for solving the task and also tells her where to start. If the learner has little experience of solving these types of tasks, just knowing where to start can be a problem and the task can seem overwhelming. Feedback that guides a learner towards the correct answer can be helpful for many. This app provides the learner with further hints if she doesn't know how to move forward.

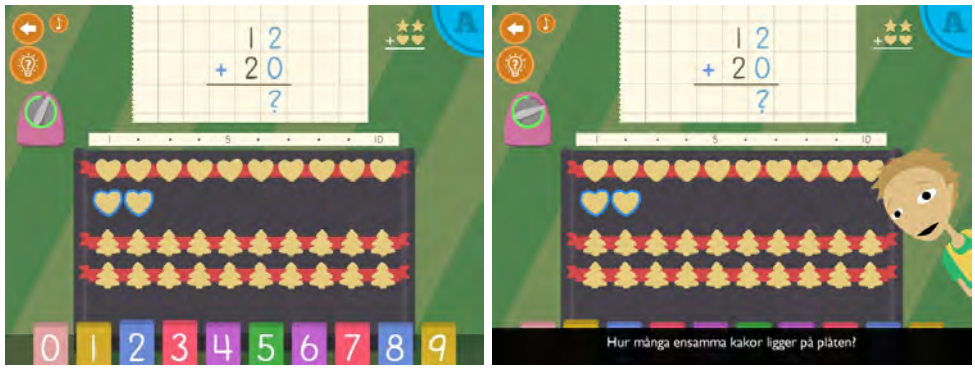


Figure 20. An example of facilitative feedback where the learner is provided with hints guiding her towards the correct answer. The hint appears at the bottom in the right panel saying: "*How many single cookies are there on the plate?*"

Another example, from the app *Bokstavspussel*, can be seen to the left in figure 21 in which the learner is supposed to spell the word "*giraff*" (English: "*giraffe*"). After a first incorrect try one letter is revealed, after two incorrect tries a second letter is revealed, and so forth.

This type of feedback provides a small part of the solution in order to help the learner spell the word correctly. It can potentially be problematic if a learner mindlessly drags whichever letter to a random place just to learn where to put one letter. After seven tries (in this case with "*giraff*") the answer will be shown and could just be copied. In cases when learners actually do make an effort and try to spell the word correctly, the feedback provided can, however, provide an adequate support for learning. It can be compared to the beneficial effects of scaffolding feedback that Finn and Metcalfe (2010) found in their studies.

It is more likely that the strategy of copy-&-paste, which is not desirable from a learning perspective, is applied by learners who use *Happi stavar* (figure 21, right panel). This game aims to let learners practice spelling with cross puzzles. The learner can try on her own, but if she gets stuck there is a lightbulb in the upper left corner, which makes the correct spelling appear in the background. Again, for a learner who doesn't know how to spell a certain word or for a learner who doesn't want to make an effort, the task can easily be solved by just copying the correct answer (after clicking the light bulb).

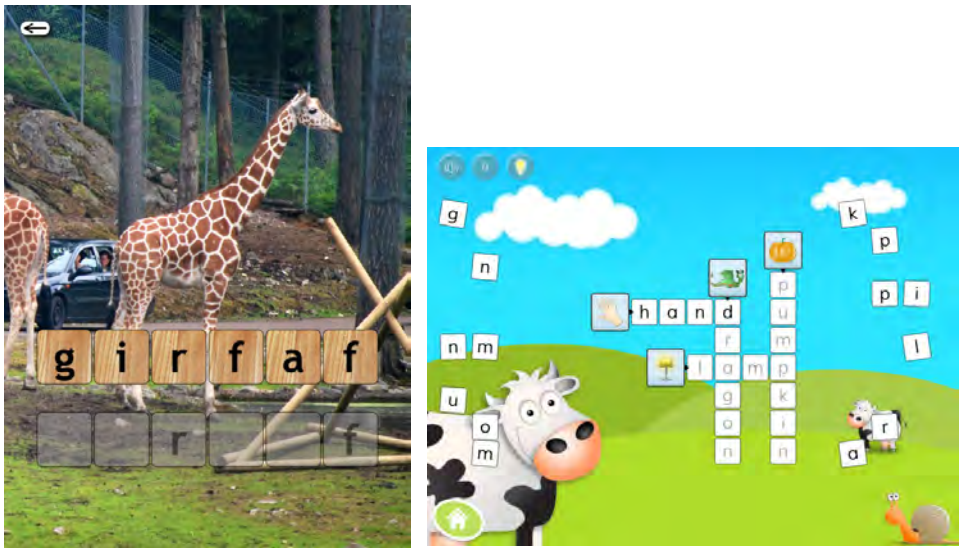


Figure 21. Two examples of facilitative feedback in which the learner *could* chose a copy-&-paste strategy.

Another type of hint can be seen in figure 22. Here the learner's task is to find the numbers that equal 10. After a few incorrect tries a hint appears in the upper left corner, providing the learner with an example of such numbers ("*8 + 2 = 10*"). Providing the learner with hints like this may remind her what she is supposed to do, and also what one possible solution could look like. Showing examples of performance can make explicit to the learner what is required; in addition, it can define a standard (Orsmond, Merry, & Reiling, 2002). However, in this game incorrect answers can also originate from the fact that the learner does not remember where certain numbers are situated; in other words, it is not merely a mathematical task, but also a memory task.

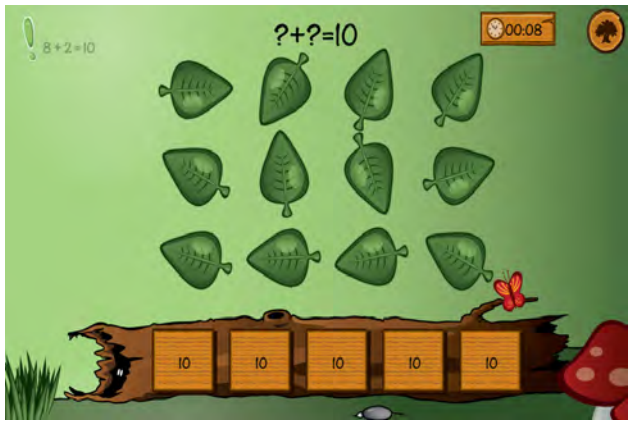


Figure 22. An example of facilitative feedback, showing a type example.

Further, there are apps that do provide good facilitative feedback, but still suffer from certain problems. In the app *Todo math – light it up* the learner is supposed to practice counting. As demonstrated in figure 23, the learner is supposed to solve "*10 + ? = 15*". The starting position shows ten blocks on the number line, but also a yellow triangle showing the final sum. In general, this is a very good example of facilitative feedback, since the learner receives support in the form of visual representations from the number line, and they get to see implications of their answers (implication feedback) when adding too few (figure 23, right panel) or too many (figure 23, left panel) boxes to the line. A problem, though, is that everything is shown from the beginning. This means that the learner does not have time to think the numbers through. Already from the beginning, the starting position and the end position are shown, and the learner only has to fill in the blanks. According to Bangert-Drowns et al. (1991), this type of information (where the learner will not have time to verbalize an answer of their own) can even have negative effects on learning.

On the other hand, the provided hints are good and would probably work great for learners who are struggling with these types of tasks. An alternative could be to let the learner have a go without the number line and blocks, and these could be added one at a time when the learner needs them.

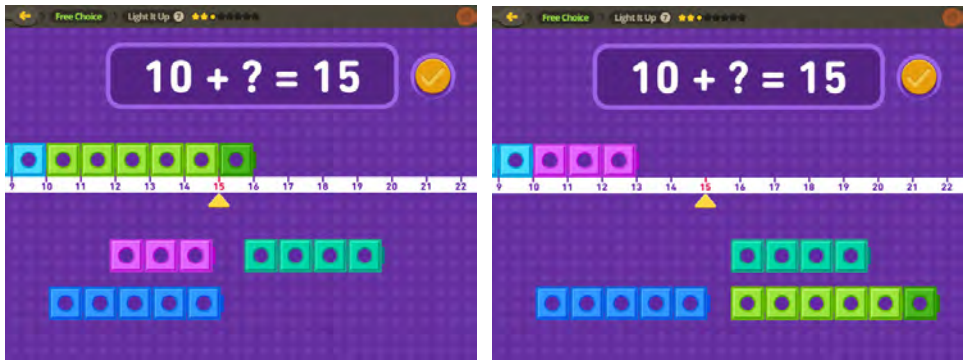


Figure 23. Example from *Todo math*, which guides the learner towards the task, but with the problem that all information is given from the start, without giving the leaner time to think the task through.

A similar problem is found in *Motion math – fractions*. Here the learner is supposed to tilt the tablet to make a bubble bounce at different fractions (¼ in the example in figure 24). After a first incorrect bounce an arrow appears, showing in which direction the learner should move the bubble. After a second missed bounce, lines appear (figure 24, second panel), displaying a visual representation of the whole number. After one more mistake the app displays the incorrect fraction the learner bounced the bubble on (figure 24, third panel). As one last hint, the app displays an arrow showing the learner the correct answer. Then the learner can try again for as many times as she wants.

Again, the facilitative feedback is well designed and aims at helping the learner reaching the correct answer. However, the problem is that the response has to be so quick that learners may have a hard time reaching the correct answer. The ball is bouncing at a predetermined rate, and if you are a slow thinker or are having trouble with how to tilt the device, you will not have the time to make a correct bounce. Removing the time factor or being able to choose at what rate the learner wants it to bounce, could possibly make the task easier. Or why bounce at all?

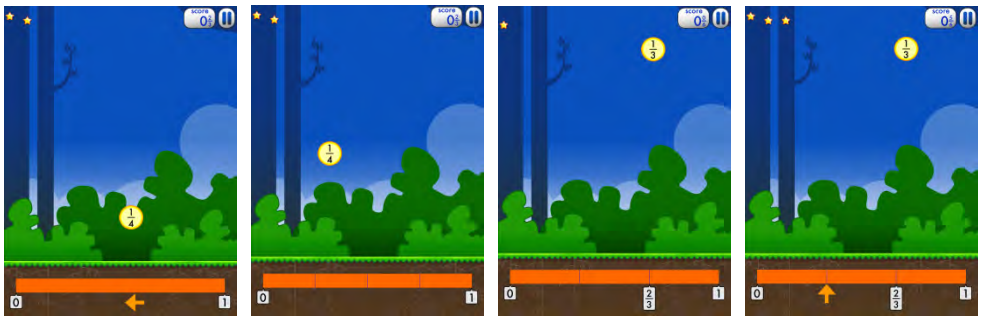


Figure 24. Example from *Motion math – fractions*, which guides the learner towards the correct answer. A problem might be that the hints are shown too fast, so that the student won't have the time to think the answer through.

A better example of facilitative feedback can be found in a subgame in the app *Vektor*, where the learner is asked to represent the number in the grey box (here number "5", see figure 25). The learner starts out with a timeline, where the numbers 0, 5 and 10 are visually shown. If she does not success in three trials, additional facilitative feedback is provided in form of more numbers shown on the line (figure 25, upper right panel). If the learner still doesn't succeed, the app displays additional hints in the form of an arrow showing the correct answer (figure 25).

Likewise, if the learner does succeed with the task, the hints are removed one at a time so that no numbers are shown in the end. One may argue that there is a possibility of applying a strategy of not succeeding on purpose, in order to be able to copy-&-paste the answer in the end, yet the learner has to try three times before a hint is displayed, and when she succeeds with three trials in a row, the hints are again removed.

Figure 25. A good example of facilitative feedback in that it provides the learner with hints on how to solve the task. The hints are removed when the learner succeeds with the task at hand.

Explanatory feedback

Only two apps out of the entire sample of 242 apps fulfil the criteria of providing explanatory feedback. In *Zcooly affären 2* (figure 26), the learner takes the role of a cashier with the task of providing the customer with the articles asked for. The learner also has to charge the customer the correct amount by putting money in the cash machine. When the learner does something wrong, the app tells her what was wrong, for instance, that the customer has received the wrong items or has been charged too much or too little.

Figure 26. An example from the app *Zcooly-affären 2*, in which the learner should provide the customer with the correct groceries and charge the customer. The learner is provided with explanatory feedback when she does something incorrect.

In the second example (figure 27), the app provides both explanatory and implication feedback. The learner owns a bakery, in which she bakes cupcakes to sell in order to make money. Then with more money she can buy more ingredients and bake more cupcakes, and so forth. In the upper left panel in figure 27, the customer requests x number of cupcakes. After having delivered the cupcakes, the learner is provided with information regarding incomes versus expenditures, and she can see the implication of her income and expenses in the form of earned money (figure 27, upper right panel). Further on, there are two stores from which the learner can buy her ingredients for the cupcakes. After such shopping the learner receives information about her purchase, telling her whether she could have saved money by going to the other store, or if she made the best available purchase (figure 27, bottom left panel). The app also provides the learner with an opportunity to find out more about her purchase (in this case overly expensive) by clicking "*Really?*" instead of "*OK*". Further explanations regarding her purchase are then provided (figure 27, bottom right panel).

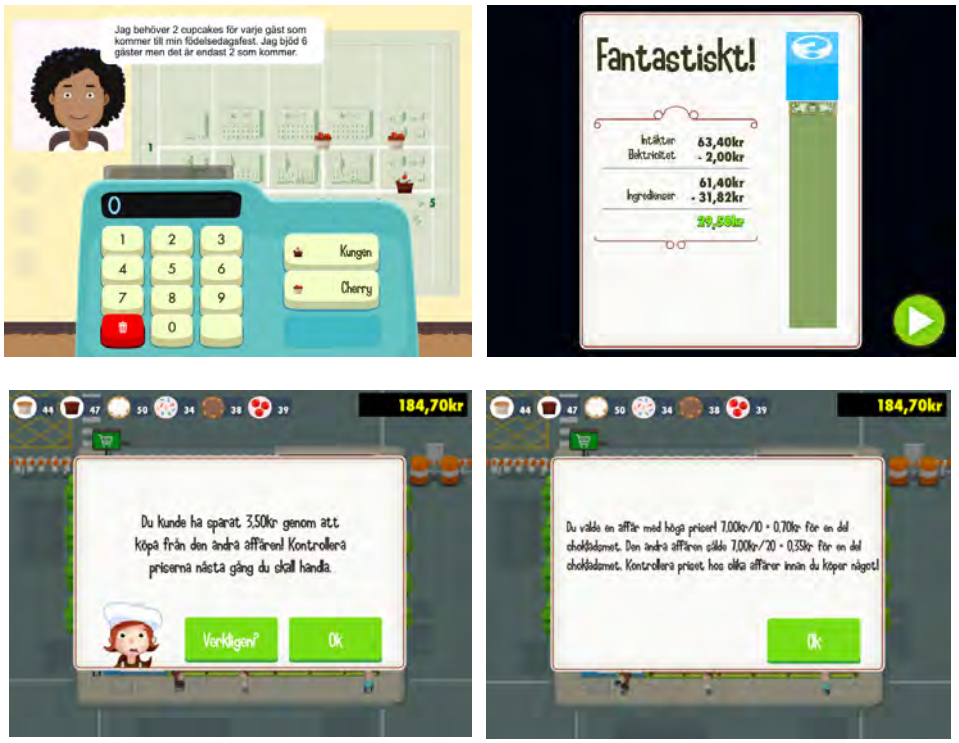


Figure 27. An example of both implication and explanatory feedback. Upper left: a customer telling the learner how many cupcakes she would like to buy. Upper right: the learner is shown the implications of her income and expenses. Bottom left and right: explanations to the learner regarding their purchase and why this was not the best purchase. In the bottom left panel it says: "*You could have saved 3,50 SEK by shopping in the other store! Check the prizes next time you are out shopping*.". In the bottom right, more explanations are provided "*You chose a store with high prices! 7.00 SEK/10 = 0.70 SEK for one batch chocolate dough. The other store sold one batch chocolate dough for 7.00 SEK/20 = 0.35 SEK. Compare prizes in different stores before you buy anything!*"

Here are two examples in which the student could have benefited from some explanatory feedback. In *Farm factor* (figure 28) the learner´s task is to fill in the number she thinks corresponds to the number of radishes in the basket. When the learner fills in "*5 × 3*" (figure 28, left panel), the app tells her that this is incorrect. She also receives a hint telling her that "*The multiplication symbol × means 'groups of'*." In the left part of the left panel of figure 28 it says: "*There are 3 groups of 5*.", but this is easily missed if the learner is just concerned with calculating the number. Also, the hint tells the learner nothing of how her answer "*5 × 3*" is the same as the requested "*3 × 5*" from a mathematical point of view. More informative feedback explaining to the learner in what sense it is adequate to equate "*3 × 5*" and "*5 × 3*" and in what sense it is not, which relates to what kind of answer is searched for in

this task, ought to be provided. Without this, the learner might in the worst case believe that "*5 × 3*" does not equal "*3 × 5*".

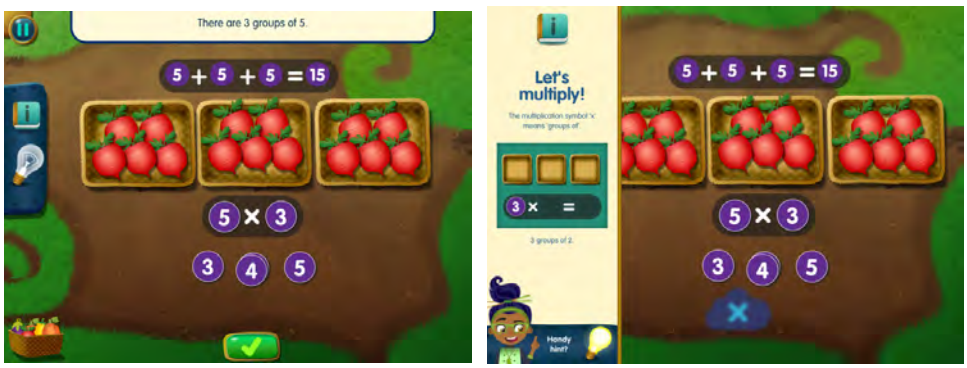


Figure 28. Example of a math problem in which the learner would have benefited from some explanatory feedback.

In a similar example (figure 29), the learner is supposed to spell "*ambulans*" (English: "*ambulance*"). When pressing "*a*" as the first letter, she receives feedback telling her that this is wrong. But why is it wrong? Well, there are two a's in ambulance and the student picked the 'wrong' one.

Figure 29. Example of a situation in which the learner makes an error but receives no explanation.

Encouraging feedback

Encouraging feedback often occurs in combination with some of the other types of feedback. In my review I have found two types of encouraging feedback. One is

encouragement in the form of applauses, cheering, balloons and stars, or other displays that appear after a task has been completed. In the other type, encouragement comes in the form of spoken or written utterances evaluating the learner's performance, such as: "*good work*", "*perfect*", or "*amazing, you did it*". Around half of all apps in this review (55%) contain some type of encouraging feedback. In turn, 53% of these make use of spoken or written utterances that comment on how well the learners perform, and the remaining 47% make use of balloons, cheering, etc. (See figure 31 for two examples.)

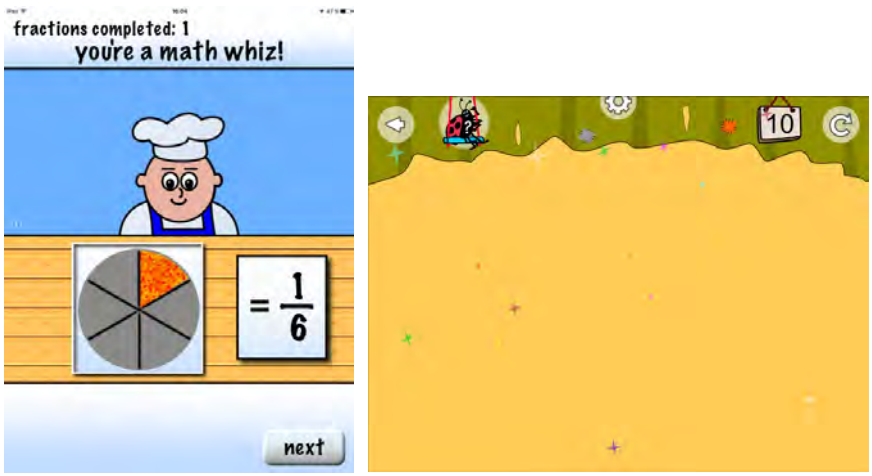


Figure 31. Two examples of encouraging feedback. To the left, encouragement towards the learner, and to the right, encouragement in the form of stars falling after a completed task.

As mentioned above, encouraging feedback rarely leads to higher performance or to higher self-efficacy (Hattie & Timperley, 2007). Nonetheless, learners do like to be praised (Sharp, 1985; Burnett, 2002; Elwell & Tiberio, 1994). More than half of the apps in the present review contained various kinds of encouragement and praise. Most likely, such feedback does not boost the learner's performance. On the other hand, if learners like it and it does not *decrease* performance, it should be fine.

Yet, the picture is more complex and worth digging into. This type of feedback can have a negative effect when it becomes obvious to the learner that there is no relation between the feedback and what actually goes on. For example, as illustrated in figure 32, the learner (me) correctly answered 21 questions out of 80 (making it 59 incorrect answers), but the sign still says that I "*did a great job*" and that "*this was awesome*". Even though it may not be wise to say, "*this was not so very good*", the response "*this was awesome*" might make the learner question the apps credibility, since most likely the learner has a sense of how well (or not so well) she has done on the task. Another known effect is that a learner risks thinking along the lines:

"*You don't think much of me if you say this was awesome.*" or "*So, no one expects more than this from me.*"

Figure 32. An example of encouraging feedback sending mixed signals to the learner. The top row says: "*4 correct answers and 11 incorrect*", second row: "*5 correct answers and 19 incorrect*", third row: "*9 correct answers and 15 incorrect*", and the bottom row: "*3 correct answers and 14 incorrect*". On the board to the right is says: "*Well done!*", "*Really good!*" and "*Awesome!!!*"

Almost all encouraging feedback in the entire range of apps was delivered after a *successful* trial. It is worrying how extremely unusual it was that an app contained any form of encouragement when the learner *did not* succeed with the task. Only 9 apps encouraged the learner to try again. In figure 33, the learner is encouraged to continue with the task by hearing things like "*not completely right, try again*" or "*there is a picture that fits the sound better, click the mouse with striped pants to hear the sound again*". This at least acknowledges that the learner clicked an incorrect answer and encourages her to go for another round. No app gave the learner any encouragement or praise for her effort, saying that the learner is doing a great job putting so much effort into the task or that she has fought well when doing something wrong. According to Hattie and Timperley (2007) and Dweck (2000), comments targeting the learner's intelligence and/or ability are problematic, since they turn the focus to the person and not the task, something that students can perceive as threatening. From a learning perspective it is preferable to comment on the efforts and/or steer the focus towards the task.

Figure 33. In the app *Bornholmslek – Ljud* the learner is encouraged to continue with the task when clicking the incorrect picture.

Result feedback

When it comes to result feedback, 94 apps out of 242 provided the learner with information presenting her results. This is information that can be used to compare to other learners or between own results, for instance to see whether one is making progress. Examples of result feedback can be seen in figure 34, where the learner receives the result and her personal high score. Often these results are received in combination with some type of encouragement (see figure 34 right panel) where the learner receives the comment "*CLOSE ENOUGH! You scored 5 out of 10 […]*".

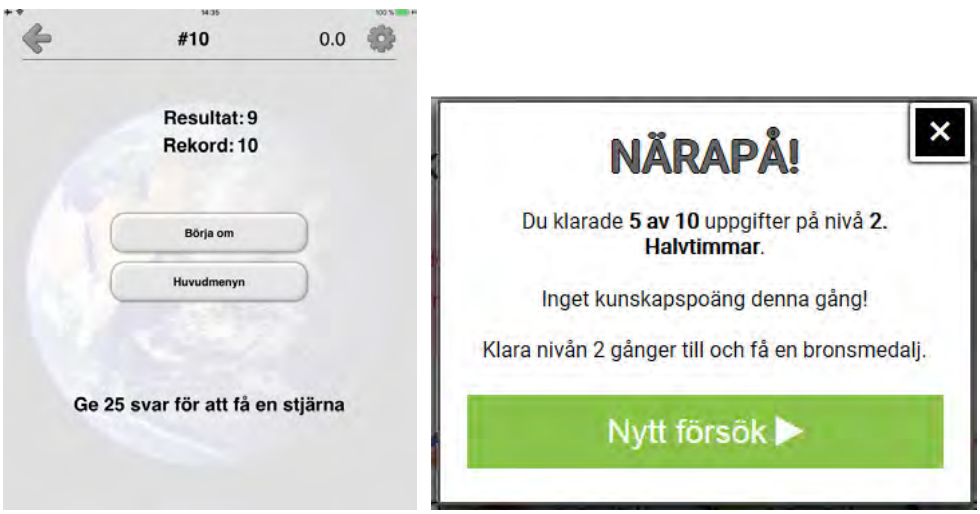


Figure 34. Two examples of result feedback. Left panel shows: "Results: 9; High score: 10". Right panel shows: "*Close enough! You scored 5 out of 10 tasks at level 2.*"

Eighteen apps showing result feedback presented the number of correct answers needed in order to move on to the next level. In figure 35 the learner needs five correct answers in a row in order to finish. The number of incorrect answers is not displayed, so a regular comparison to other results cannot be made. This could be positive, in that it is impossible for a learner to compare herself to others, which can potentially cause stress and negative feelings.[5]

From a teacher's perspective this feedback can be problematic, since the only result the teacher will ever see is the number of correct answers, and she will not know what types of questions the learner struggled with. Another example is when only the correct answers are summarized and displayed. There are several apps designed in this way.

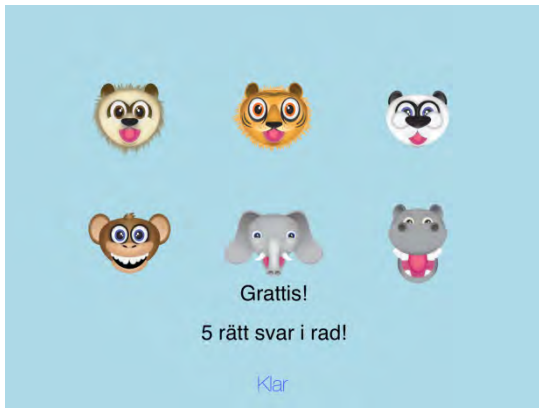


Figure 35. Example of result feedback, only shown after the goal is reached, saying "*Congratulations! 5 correct answers in a row!*"

Summary and conclusion

As predicted, a majority of the apps do not provide learners with elaborated feedback. In fact, only 29 out of 242 (twelve percent) provide anything more than only verification or corrective feedback. From a learning point of view, this is disappointing. If we look at the literature, most research emphasizes the importance of elaborated feedback for learning (McKendree, 1990; Bangert-Drowns et al., 1991; Pridemore & Klein, 1995; Moreno, 2004; Shute, 2008).

[5] To be noted is that this type of feedback can also be positive for some students, who see the results as encouragement and as an incentive to try harder.

Verification feedback

Verification feedback was the predominant type of feedback, and in 78% off all apps this was the only kind of feedback provided. With this type of feedback, the learner will know whether their answer is correct or not, as it provides some sort of guidance. For a learner with prior task knowledge such guidance can be sufficient as support towards the correct answer after they have made a mistake. But for a learner who does not have such prior task knowledge, just knowing whether their answer was correct or not will not help much.

Similarly, if the learner believes that her answer is correct, whereas the feedback says this is not the case, this may cause frustration and helplessness. Being told that you are wrong without any further guidance telling *why* or *how* can be problematic. Providing the learner with the correct answer (corrective feedback) will at least provide her with some information – but involves other disadvantages. Being presented with the solution instead of being allowed to actively come up with it yourself is often less powerful in terms of understanding and remembering.

A problem associated with verification feedback is that it encourages using trial-&-error strategies. In this review I categorized three different trial-&-error strategies. 'Low-cost trial-&-error ', in which the learner can move forward at a low cost in terms of time and effort, is the type of verification feedback that is the most problematic from a learning perspective. It is very common that apps allow learners to use this strategy – in this review 59% of all the apps provided verification feedback only. A low-cost trial-&-error strategy allows the learner to just click different answers until the correct one is hit, and there are no consequences when clicking an incorrect one.

That it is *possible* to use a low-cost trial-&-error strategy is not a problem with learners who actually try to solve the task and who are making an effort. But with learners who only want to 'get by' and would rather *not* make an effort, this possibility is troublesome. Research has shown that so-called 'gaming the system' is negatively related to learning (Aleven & Koedinger, 2001; Baker et al., 2004, 2005, 2006; Walonoski & Heffernan, 2006). Although this strategy is not used by all learners, the learners not using it are, in general, not the ones we have to worry about.

Similarly, if the learner does not know the correct answer, pure guessing – which is possible in apps that allow low-cost trial-&-error strategies – can let the learner finish the task with a good score and in good time. From a perspective from the outside, it might seem as if the learner knows what she is doing, whereas in fact little learning has occurred.

Pure guessing, which can be used in apps that allow for low-cost trial-&-error strategies, is constrained in an app that only allows for what I call risky trial-&-

error. In this case, if the learner uses the strategy of pure guessing, there is a cost in terms of 'lives', scores, or levelling. That is, every mistake the learner makes costs her, for example, a life or several points. Low-cost trial-&-error strategies are possible in 19% of all apps in the review that contain verification feedback only.

It is important to point out here that there is no ideal type of feedback, which will always work best for all learners in all situations. It is also not the case that verification feedback is always inferior to other kinds of feedback or a bad design choice. As already mentioned, verification feedback can be just what a high-performing learner with sufficient prior knowledge needs to work on a given task and to learn from it. Also, if the purpose of an app is to test or evaluate knowledge or skills, verification feedback is adequate. The mismatch may arise if the app is advertised as an app that supports learning.

Corrective feedback

Extending verification feedback by adding corrective feedback provides the learner with somewhat more information: at least they won't have to wonder what the correct answer should be. Potentially, they may also use the provided correct answers for further learning. In this review 10% of the 242 apps contained corrective feedback. That is, they provide the learner with the correct answer when she proposes an incorrect one. Previous research has shown that corrective feedback is more beneficial for learning compared to verification feedback (Pashler et al., 2005; Marsh et al., 2012).

There is a caveat: if the learner only memorizes the provided correct answers without reflecting on them and, if appropriate, trying to understand why a particular answer is correct, the resulting learning may be shallow. A piece of information learned by heart, with no knowledge of how and in what situations to use it, will not lead anywhere. Knowing *that* this was the right answer does not equal knowing *why* this was the correct answer. Again, learners with adequate prior knowledge are more likely to figure out why an answer is correct, whereas for students who are less knowledgeable this will be harder or impossible.

Elaborated feedback

When it comes to *elaborated feedback*, 23 out of the 29 apps that provide more than only verification or corrective feedback contained *facilitative feedback*. This refers to feedback that provides some kind of hint on how to solve or proceed with the task. This can be a good way of guiding the learner towards the correct answer if she is stuck with a task. Yet, in some cases, such as, for example, in *Happi stavar*

(figure 21, right panel), it opens up for the use of a copy-&-paste strategy if the learner wants to avoid making an effort (or completely mistrusts her own abilities to learn and to solve tasks).

Explicit explanations as to why a certain answer is correct or not were only provided by two apps. In *Motion math cupcakes* the learner is provided feedback on whether and why she made the best purchase she could when buying ingredients for her cupcakes. This information pinpoints an important feature, namely that comparing prices between the two stores could save her some money in future purchases. Even though this app contains somewhat more complex tasks, so that it may be more obvious that explanatory feedback is an adequate feature, other apps could very well benefit from it as well. For example, in a spelling app, if a learner is spelling the word "*träd*" (English: "*tree*") with two ä's, the app could tell the learner that this was almost correct, but that in the Swedish language we seldom use two vowels in a row, with a few exceptions like "*zoo*" and "*leende*" (English: "*smile*").

Another way to provide the learner with more information about how to reach the correct answer is to provide implication feedback. This is also the way we often encounter feedback in our everyday life. Four of the apps reviewed contained this type of feedback, and in three of them the task concerned math and a balancing scale. Even though it is encouraging to see that this type of feedback is prevalent in this specific domain, it should also be possible to provide this kind of feedback for many other types of tasks (cf. *Critter Corral*: Blair, 2013). This could for example be applied in *Lolas mattetåg* (figure 14) where the learner sometimes has to solve a math task by adding two numbers. Such an addition could be as follows. If the learner answers correctly, Lola's train will reach the train station, but if the learner proposes a sum that is too large, the train moves past the station, and if the learner proposes a sum that is too small, the train stops before the station.

Encouraging feedback and result feedback

Encouraging feedback almost always comes in combination with some other type of feedback, and 133 out of 242 apps (55%) contained either encouraging feedback in the form of cheering, balloons, etc., or messages in text or voice saying that the learner is awesome, is doing perfect or very well. In all apps that make use of written expressions, the feedback targets the learner and not the task. In other words, it is the learner who is praised for being smart, doing great, etc. The focus is on the child – not on the task. Addressing intelligence and/or ability in this manner has not been shown to be beneficial for learning. On the contrary, it can make the learner focus on the wrong things and lead them to avoid future tasks in which they risk failing (Dweck, 2000; Gunderson et al., 2013).

Yet, encouragements and praise are often appreciated by learners, and they can indeed be useful. The recommendation is also not to eliminate encouragement and praise, but to shift the focus from the learner to the task or to the effort that the learner puts into the task. Adding encouraging feedback telling the learner that she is making progress, making good effort, does not seem to give up easily, etc. should not be an impossible design task for app designers. Overall, learners can use some encouragement when they have made a mistake, but continue working on the task, and not only when they have already finished the task. Only 9 out of 133 apps encouraged the learner in some way to continue. This could be done more often. From my experiences of talking to teachers in schools and preschools, they are well aware of the drawbacks of praise that focuses on the person ("*you are really bright*", "*you are very good in math*", "*oh, you are smart*"). Instead, they praise and encourage with a focus on the task or what has been produced ("*this is very well done*", "*I like how you solved this*", "*this essay is very well written*"). In addition, they all agree on the importance of encouraging effort and providing feedback during the working and learning process. There is a striking mismatch between teachers' views on encouraging feedback and the implementation of encouraging feedback in educational apps.

When it comes to result feedback, 39% of 242 apps presented results that a learner can use to compare herself to others or use as a measure of her own progress. For competitive learners this can be a good way to motivate themselves to continue and try harder, but for learners who do not appreciate competition or have low beliefs in their own ability it can instead be stressful. In most cases this kind of feedback tells the learner nothing about in what respects they need to practice more. They will not know which questions they answered wrong or which topics they did less well at. This type of feedback therefore rarely leads to increased learning, but it can be used by some learners as a measure of when they have to work harder.

Conclusion

The use of digital apps is increasing in schools today, and in order for them to be useful as learning devices, and not only testing devices, they need to provide feedback that is more informative than only telling whether a choice or answer was correct or incorrect. One advantage with technology is that it offers an opportunity to provide all learners with the same or individualized feedback at the same time, and it is up to the designers to make the most out of this, just as it is up to them to reduce the opportunities for trial-&-error, in particular low-cost trial-&-error.

However, reading an introductory text for an app will often not reveal whether the app is indeed a learning – and not a testing – device. It is not forbidden to use the

term ‘supports learning’ in a text that describes an app, even though no learning scientist would approve. While working on this review, I read through all available information texts. Only four out of 99[6] stated that the app in question was designed ‘to test or evaluate skills and knowledge’, whereas many more are suited for testing purposes – but not for learning purposes. The only way to know if a certain educational app matches the purpose you have – whether as a teacher or a parent – is to play it yourself and try to make as many mistakes as you can.

A tentative conclusion on the basis of this review is that many educational app designers view a learner as someone just waiting to be informed whether an answer or a choice was correct or not. This kind of feedback corresponds to a behaviouristic approach comparable to instrumental conditioning by means of reinforcement. In essence, most apps miss the opportunity of treating the learner as an active and constructive being who would benefit from more nuanced feedback.

References

Aleven, V., & Koedinger, K. R. (2001). Investigations into help seeking and learning with a cognitive tutor. In R. Luckin (Ed.), *Papers of the AIED 2001 Workshop on Help Provision and Help Seeking in Interactive Learning Environments* (pp. 47-58). May 19-23, 2001, San Antonio, Texas, TX.

Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*, 1-18. doi: 10.1037/a0021017

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004). Detecting student misuse of intelligent tutoring systems. In J. C. Lester, R. M. Vicari, & F. Paraguaçu (Eds.), *7th International Conference on Intelligent Tutoring Systems*, *ITS 2004* (pp. 531-540). Berlin, Germany: Springer.

Baker, R. S., Corbett, A. T., Koedinger, K. R., Evenson, S., Roll, I., Wagner, A. Z., ... , & Beck, J. E. (2006). Adapting to when learners game an intelligent tutoring system. In M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), *LNCS, vol. 4053: Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, *ITS 2006* (pp. 392-401). Berlin, Germany: Springer.

Baker, R. S., Roll, I., Corbett, A. T., & Koedinger, K. R. (2005, May). Do performance goals lead learners to game the system? In C.-K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Frontiers in Artificial Intelligence and Applications, vol. 125: Proceedings of AIED 2005* (pp. 57-64). Amsterdam, The Netherlands: IOS Press.

[6] Texts were not found for all 103 apps.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why learners engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185.

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213-238.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7-74.

Blair, K. P. (2009). *The neglected importance of feedback perception in learning: An analysis of children and adults' uptake of quantitative feedback in a mathematics simulation environment* (Doctoral Dissertation). Stanford University, Stanford, CA. Retrieved from: https://eric.ed.gov/?id=ED532821

Blair, K. P. (2013). Learning in critter corral: evaluating three kinds of feedback in a preschool math app. In *Proceedings of the 12th International Conference on Interaction Design and Children*, *IDC'13* (pp. 372-375). New York, NY: ACM.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn*. Washington, DC: National Academy Press.

Brockbank, A., & McGill, I. (1998). *Facilitating reflective learning in higher education*. Bristol, PA: Taylor & Francis.

Burnett, P. C. (2002). Teacher praise and feedback and learners' perceptions of the classroom environment. *Educational Psychology*, *22*(1), 5-16.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology*, *79*(4), 474-482.

Cherner, T., Dix, J., & Lee, C. (2014). Cleaning up that mess: A framework for classifying educational apps. *Contemporary Issues in Technology and Teacher Education*, *14*(2), 158-193.

Chi, M. T. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1*(1), 73-105.

Clariana, R. B. (1990). A comparison of answer until correct feedback and knowledge of correct response feedback under two conditions of contextualization. *Journal of Computer-Based Instruction*, *17*(4), 125-29.

Craven, R. G., Marsh, H. W., & Debus, R. L. (1991). Effects of internally focused feedback and attributional feedback on enhancement of academic self-concept. *Journal of educational psychology*, *83*(1), 17-27.

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

Elwell, W. C., & Tiberio, J. (1994). Teacher praise: What learners want. *Journal of Instructional Psychology*, *21*(4), 322-328.

Finn, B., & Metcalfe, J. (2010). Scaffolding feedback to maximize long-term error correction. *Memory & Cognition*, *38*(7), 951-961.

Fox, B. A. (1991). Cognitive and interactional aspects of correction in tutoring. In P. Goodyear (Ed.), *Teaching knowledge and intelligent tutoring* (pp. 149-172). Norwood, NJ: Ablex.

Gibbons, P. (2002). *Scaffolding language, scaffolding learning: Teaching second language learners in the mainstream classroom* (pp. 110-119). Portsmouth, NH: Heinemann.

Gunderson, E. A., Gripshover, S. J., Romero, C., Dweck, C. S., Goldin-Meadow, S., & Levine, S. C. (2013). Parent praise to 1-to 3-year-olds predicts children's motivational frameworks 5 years later. *Child development*, *84*(5), 1526-1541.

Haake, M. (2018). *No child left behind, nor singled out – reasons for combining adaptive instruction and inclusive pedagogy in early math software.* Manuscript submitted for publication.

Handal, B., El-Khoury, J., Campbell, C., & Cavanagh, M. (2013). A framework for categorising mobile applications in mathematics education. In *Proceedings of the Australian Conference on Science and Mathematics Education* (pp. 142-147), Canberra, Australia, Sept 19-21, 2013.

Hanna, G. S. (1976). Effects of total and partial feedback in multiple-choice testing upon learning. *The Journal of Educational Research*, *69*(5), 202-205.

Harlen, W., & Deakin Crick, R. (2003). Testing and motivation for learning. *Assessment in Education: Principles, Policy & Practice*, *10*(2), 169-207.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Higgins, R., Hartley, P., & Skelton, A. (2001). Getting the message across: the problem of communicating assessment feedback. *Teaching in Higher Education*, *6*(2), 269-274.

Highfield, K., & Goodwin, K. (2013). Apps for mathematics learning: A review of 'educational'apps from the iTunes App Store. In V. Steinle, L. Ball, & C. Bardini (Eds.), *Proceedings of the 36th Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 378-385). Melbourne, Australia: MERGA.

Hirsh-Pasek, K., Zosh, J. M., Golinkoff, R. M., Gray, J. H., Robb, M. B., & Kaufman, J. (2015). Putting education in "educational" apps: Lessons from the science of learning. *Psychological Science in the Public Interest*, *16*(1), 3-34.

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

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75-86.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254-284.

Lepper, M. R., Aspinwall, L., Mumme, D., & Chabay, R. W. (1990). Self-perception and social perception processes in tutoring: Subtle social control strategies of expert tutors. In J. M. Olson, M. P. Zanna, & C. P. Herman (Eds.), *Self inference processes: The Ontario Symposium*, *Vol. 6* (pp. 242-257). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, *1*(4), 279-308.

Larkin, K. (2015). "An app! An app! My kingdom for an app": An 18-month quest to determine whether apps support mathematical knowledge building. In T. Lowrie & R. Jorgensen (Eds.), *Digital games and mathematics learning* (pp. 251-276). Dordrecht, The Netherlands: Springer.

Lea, M. R., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education*, *23*(2), 157-172.

Marsh, E. J., Lozito, J. P., Umanath, S., Bjork, E. L., & Bjork, R. A. (2012). Using verification feedback to correct errors made on a multiple-choice test. *Memory*, *20*(6), 645-653.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, *59*(1), 14.

McKendree, J. (1990). Effective feedback content for tutoring complex skills. *Human-computer interaction*, *5*(4), 381-413.

Moreno, R. (2004). Decreasing cognitive load for novice learners: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional science*, *32*(1-2), 99-113.

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational psychology review*, *19*(3), 309-326.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, *31*(2), 199-218.

Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using learner derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education*, *27*(4), 309-323.

Orsmond, P., Merry, S., & Reiling, K. (2005). Biology students' utilization of tutors' formative feedback: A qualitative interview study. *Assessment & Evaluation in Higher Education*, *30*(4), 369-386.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 3.

Phye, G. D., & Sanders, C. E. (1994). Advice and feedback: Elements of practice for problem solving. *Contemporary Educational Psychology*, *19*(3), 286-301.

Pridemore, D. R., & Klein, J. D. (1995). Control of practice and level of feedback in computer-based instruction. *Contemporary Educational Psychology*, *20*(4), 444-450.

Schwarz, D. L., Tsang, J. M., & Blair, K. P. (2016). *ABC's of how we learn – 26 scientifically proven approaches, how they work, and when to use them*. New York, NY: W. W. Norton & Company.

Segedy, J. R., Kinnebrew, J. S., & Biswas, G. (2013). The effect of contextualized conversational feedback in a complex open-ended learning environment. *Educational Technology Research and Development*, *61*(1), 71-89.

Sharp, P. (1985). Behaviour modification in the secondary school: A survey of students' attitudes to rewards and praise. *Behavioral Approaches with Children*, *9*, 109-112.

Shute, V. J (2008). Focus on formative feedback. *Review of educational research, 78*(1), 153-189.

Sjödén, B. (2017). What teachers should ask of educational software: Identifying the integral digital values. In *Proceedings of 10th annual International Conference of Education, Research and Innovation*, *ICERI2017* (pp. 6491-6500), Seville, Spain, November 16-18, 2017.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, *4*(4), 295-312.

Walonoski, J., & Heffernan, N. (2006). Detection and analysis of off-task gaming behavior in intelligent tutoring systems. In M. Ikeda, K. D. Ashley, & T.-W. Chan (Eds.), *LNCS, vol. 4053: 8th International Conference of Intelligent Tutoring Systems*, *ITS 2006* (pp. 382-391). Berlin, Germany: Springer.

Van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher-learner interaction: A decade of research. *Educational Psychology Review*, *22*(3), 271-296.

Wiliam, D. (2007). Keeping learning on track: Classroom assessment and the regulation of learning. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1053-1098). Charlotte, NC: Information Age Publishing.

Wilkinson, S.S. (1980). *The relationship of teacher praise and learner achievement: A meta-analysis of selected research* (Part of Doctoral Dissertation). University of Florida, Gainesville, FL. Retrieved from: https://archive.org/stream/relationshipofte00wilk/relationshipofte00wilk_djvu.txt

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, *17*(2), 89-100.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard university press.