

# Transferring Teaching to Testing – an Unexplored Aspect of Teachable Agents

Björn Sjödén<sup>1</sup>, Betty Tärning<sup>1</sup>, Lena Pareto<sup>2</sup>, Agneta Gulz<sup>1</sup>

<sup>1</sup> Lund University Cognitive Science, Sweden

{Bjorn.Sjoden, Betty.Tarning, Agneta.Gulz}@lucs.lu.se

<sup>2</sup> Media Production and Informatics Departments, University West, Sweden  
Lena.Pareto@hv.se

**Abstract.** The present study examined whether social motivational influence from working with a Teachable Agent (TA) might transfer from the formative learning phase to a summative test situation. Forty-nine students (9-10 years old) performed a digital pretest of math skills, then played a TA-based educational math game in school over a period of seven weeks. Thereafter, the students were divided into two groups matched according to their pretest scores, and randomly assigned one of two posttest conditions: either with the TA present, or without the TA. Results showed that low-performers on the pretest improved dramatically on the posttest when tested with the TA, whereas high-performers did not. We reason that low-performers might be more susceptible to a supportive social context – as provided by their TA – for performing well in a test situation. Ratings of fun, effort and confidence in taking the test pointed to a complexity of effects for future research.

**Keywords:** Learning-by-teaching, teachable agent, assessment, transfer

## 1 Introduction

Teachable Agents, *TAs*, is a form of educational technology based on the idea that a good way to learn is to teach someone else. In brief, a TA is a computer agent that is taught by a student, where AI techniques guide the agent's behavior based on what it is taught. Students can revise their TA's knowledge (and their own) based on the agent's behavior [1, 2].

Numerous studies have shown that TA-based software can be powerful in terms of learning outcomes. For example, students working with a TA exhibited deeper causal understanding than students using the same software without a TA [1], and they produced more accurate concept maps [10]. In a comparison to traditional, “pen and paper”-methods, Chin and colleagues [4] demonstrated that an equivalent system using a TA provided “added value” in terms of students learning more complex ways of reasoning and being more successful in taking on new learning material.

Lately, there has been an increased focus on the motivational aspects of TA software. In particular, students' feelings of responsibility and engagement from developing a social relation to their TAs, has been proposed as an explanatory mechanism as to why students seem to make greater efforts and spend more time on

learning material when using a TA. Chase and colleagues [3] reported two studies to this effect, noting that students acted as though their TAs were sentient, semi-independent beings, which engage in mental activity and were given partial credit for the outcomes. Students instructed to learn for their TA, rather than for themselves, were more inclined to approach, discuss and attempt to revise errors and misunderstandings. The authors suggest that the TA may provide an “ego-protective buffer” by offering a means for students to distribute the responsibility for errors and mistakes, thereby decreasing their fear of failure. Besides, such effects would seem beneficial for counteracting test anxiety in general.

In sum, TA studies suggest that the sense of social relationship between students and TAs can have positive effects on learning through an impact on motivation and engagement. But what happens when turning from a learning situation to a test situation? Can the sense of meaningfulness, engagement and responsibility developed in relation to the TA be reestablished when performing a formal test, and improve performance? We set out to explore these questions by having the TA from a prolonged learning session reappear in a digital summative assessment form, removed from its learning context, but present as a social and visual entity.<sup>1</sup>

## 1.2 Present research aims and relation to previous studies

To our knowledge, no previous study has targeted the reconstruction of social-motivational factors particularly associated to a TA outside its learning context. Many TA systems are based on making explicit concept maps, which illustrate how the TA “learns” in the form of causal chains specified by the student [e.g. 1, 3, 4, 9]. Students can get feedback on how well they have “taught” their TAs by testing them under various forms, for example how well they make inferences from questions or perform in quizzes. Notably, these learning and testing elements all take place as part of the same software. We wanted to examine how students would perform in the presence of their TAs, but in a situation that was presented as a formal test to the student (rather than the TA), clearly separated from the learning phase and the primary TA environment.

In the present study, we used a TA system (a learning game in mathematics, see below) which is not based on making concept maps, but centers on the identification and effective use of game rules [8]. However, rather than focusing on how well students taught their TAs, we focused on the motivation and engagement aspects related to the TA’s social role as a “protégé” [cf. 3] and learning companion.

It is relevant to the present study that we have conducted two recent studies of the same TA system. We found empirical support for (a) that the learning game was effective for learning and increased performance on subsequent standardized math tests more than for students not playing the game [5], and (b), that students became emotionally involved in their TAs and related to it socially while playing the game [6]. We posed two main questions for the present study:

---

<sup>1</sup> This does not imply that we take sides with traditional, summative evaluation before formative evaluation, or assessments more closely integrated with learning (in which TAs can be productively involved). Rather, we opted for a summative assessment as the subject of study because it is (still) the kind of test most commonly used in school education.

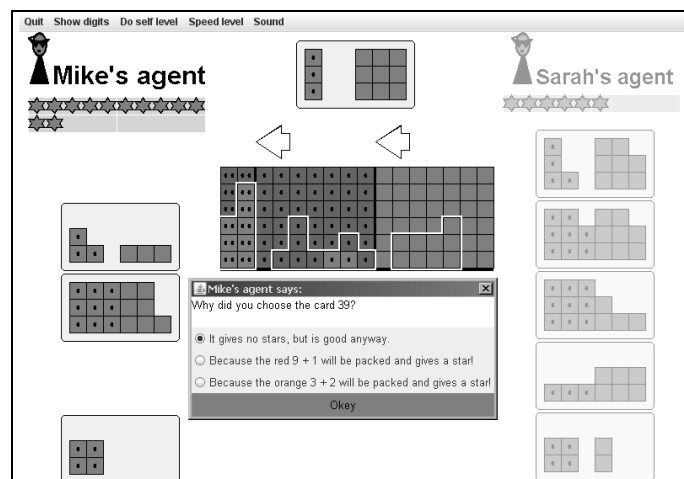
1. Following an extended period of learning with the TA, how would test performance compare for those students performing a summative test in the presence of their TA, to that of students performing the same test without the TA?

2. How would the TA affect students' experiences, such as their feelings of engagement, effort, difficulty and confidence for taking the summative test, and how would this relate to their test performance?

Next, we describe the TA environment and the method we used for measuring how the social-motivational effects of the TA might transfer from the learning phase to a test situation. We report our primary analysis of the results and how these relate to relevant subgroups of students. We close by discussing how the complexity of effects point to some unexplored aspects of using TAs, for future research.

## 2 The TA environment: an educational math game

The TA learning environment used in the present study is an educational game in elementary mathematics called “The Squares Family” [7, 8], specifically aimed at training the base-10 system (such as carry-overs and borrowings). The game employs a board-game design, including playing cards and a common game board, with several game modes and levels of difficulty. All arithmetic operations are visualized, using the graphical metaphor of squares and boxes that can be “packed” or “unpacked” in numbers of 10. Students typically play the game in pairs, either in their own name or with a TA. A game move consists of picking a card depicting a certain constellation of squares and boxes, which is then played and adds or subtracts to the present (previously played) squares and boxes on the game board. The goal is to consistently pick the cards that, in combination with what is represented on the game board, maximize the number of carry-overs (in the addition games) or borrowings (in the subtraction games). A screenshot is shown in Fig. 1.



**Fig. 1.** Screenshot of the math game, which depicts two competing TAs in an addition game. Here, “Mike’s agent” poses a question as to why Mike picked a particular card.

The TA can be set in one of three different modes. In “Watch and learn”-mode, the TA successively learns the game rules, by “watching” the student’s game moves and how the student responds to occasional multiple-choice questions. A typical question from the TA would be “Why did you pick this card?”. The student chooses between a list of potential explanations (but only one correct answer) and a “don’t know” option. In “Try and play”-mode, the TA suggests game cards, which the student can confirm by clicking “Ok” or deny by suggesting another card. In “Play Self”-mode, students can watch their TA’s performance, as it plays a session of the game automatically, against the computer or another TA or human player. If a card yields a carry-over or borrowing, a star is rewarded. The number of stars as well as whether the TA wins the game, can be seen as a form of feedback on how well the student taught the TA.

### **3 Method**

#### **3.1 Participants**

Forty-nine 4<sup>th</sup>-graders (9-10 years old) from two school classes participated in the study, 24 from one class and 25 from the other. There were 19 girls and 30 boys. The two classes followed the same educational curriculum in the same school. The students were experienced in using laptops and were familiar with the question and answering formats (e.g. Likert scales) used in this study. Due to student absence and some computer mishaps when saving test data, only the results of 43 students could be used from the pretest, and of 47 students from the posttest.

#### **3.2 Design and instruments**

Because there were no established instruments for the kind of manipulations we wanted to make, we needed to develop new tentative test materials. These included a digital pretest and a digital posttest, each of which appeared in a “TA version” and a “standard version” (that is, one test including the TA and the same test without the TA). There was one main test, with questions targeting base-10 transformations (e.g., telling what sums end in “00” from six alternatives, which sum is bigger of “28+64” and “52+35”). In total, the main test comprised 41 items, each scored zero if incorrect and one if correct (theoretical score range 0–41). One item was a control question, which addressed multiplication instead of base-10. In order for the pretest and posttests not to be completely identical, two forms of the main test were created (form A and form B). These forms had only superficial differences (e.g. the item “27+13” in form A was replaced by “13+37” in form B). Students were randomly assigned A or B as their main pretest, and the other as their main posttest.

In addition, we constructed a digital “Quick test” which consisted of 60 items, to be answered as quickly as possible. This test was added in order to have a secondary performance measure, which required less reflection and where time was a factor. The results from the Quick test will be reported elsewhere.

Exclusively for the TA versions of the posttests, the graphical TA was copied from the math game was placed in the margin of the screen. The TA's role was restricted to its visual, non-animated presence, including some introductory phrases (displaying e.g. "Hi, it's me – your agent – can you help me answering this questionnaire? I learn from you."). In the main posttest, there were two opportunities for clicking on the TA, which displayed a short sequence in which the TA answered similar items. The TA did not respond to the same items as the student, and did not provide any feedback as to whether answers were right or wrong. The TA was programmed so that it displayed as many correct answers as the student had done on the immediate previous tasks.

*Attitudes and experiences questionnaires.* Two pen-and-paper questionnaires were administered, one in connection with the pretest, which related to self-efficacy and motivation for math, and one after the posttest, relating to the student's experience of answering the posttest. This study focused on the posttest ratings. Four questions applied to all students (e.g., "How fun was the posttest?", "How much effort did you put into it?"). Students doing the TA version of the posttest were given two additional questions, relating to their sense of being helped by their TA, and wanting to teach their TA, respectively. All answers were rated on a 0-10 Likert scale, where 0 represented the negative end ("not at all fun", "no effort", "very disturbing", etc) and 10 the positive end ("very much fun", "very much effort", "very helpful", etc).

### 3.3 Procedures

In the *pretest phase*, all 49 students did the pretest on one day, on individual laptops. Two experimenters led each testing session. Participants were informed that all data were reported anonymously. The students filled out the pretest attitude questionnaire on paper and were then presented with a practice main test. This short practice test was a simplified version of the main test, for students to familiarize themselves with the test format (e.g., how to click and scroll through questions). After five minutes' practice, students did the proper main pretest (either form A or B). The students were not timed on the main test, but had about 25 minutes; almost all finished within this limit. Then they performed the timed Quick test, which took 3-4 minutes, including some practice items. The students were not given any feedback on their performance, nor informed that they were going to perform a similar test (the posttest) later.

In the *learning phase*, students participated in one session of 30 minutes per week over a period of eight weeks (including one week's intermission due to holidays). On average, students played the math game eight at a time, seated in pairs before four laptops. The sessions were semi-structured, such that new elements (the addition game, the TA and the subtraction game) were introduced by the experimenters in the beginning of each session, but the students were largely free to "practice what they felt needed", with the instruction to teach their TA what they found most difficult.

In the *posttest phase*, students followed a similar procedure to the pretest, but now doing the posttests (and without initial practice). Half the students were given the TA version of the posttest, and half were given the standard version of the posttest, by random assignment. The two groups were matched on basis of their pretest scores, so there were as many students scoring above the median as below the median in each group. Finally, each student filled out an attitudes and experiences questionnaire.

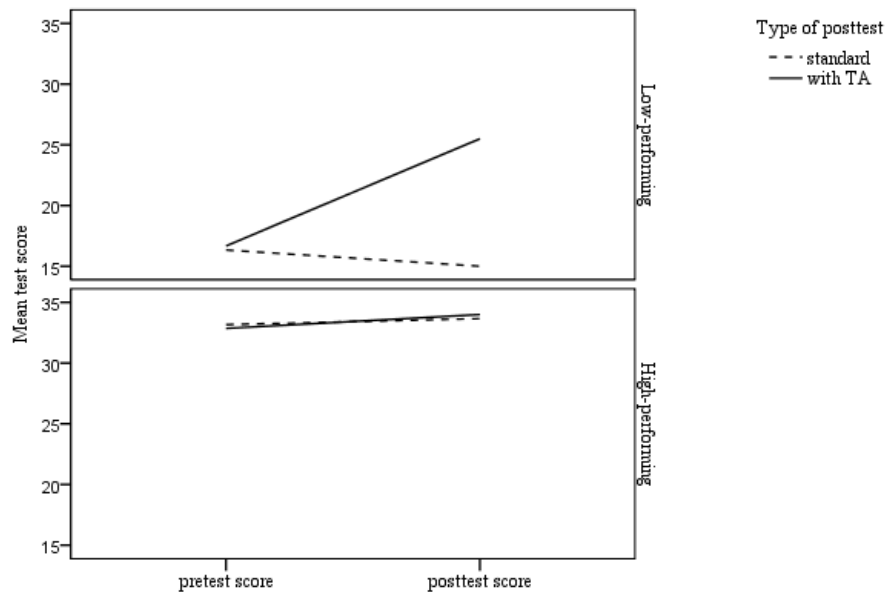
## 4 Results

### 4.1 Summative test performance in relation to the presence of the TA

Our first research question was how the performance of students doing a summative test in the presence of their TA would relate to the performance of students doing the same test without the TA. Our primary analysis was concerned with results on the main posttest. On average, students ( $n = 23$ ) who performed the TA version of the main posttest scored higher ( $M = 30.0$ ,  $SD = 5.5$ ) than students ( $n = 24$ ) performing the standard version of the same test ( $M = 26.5$ ,  $SD = 8.9$ ). However, an independent samples t-test comparing the two means was not significant;  $t(45) = 1.572$ ,  $p = .12$ .

Upon closer analysis, we were interested in how students' posttest scores related to their baseline in terms of their pretest scores. A linear regression analysis showed that pretest score was a significant predictor;  $t(40) = 6.06$ ,  $p < .05$ , which improved the overall model fit from .039 to .486 (adjusted R square). When plotting the results, we discovered that the non-significant results of posttest version seemed to be very likely due to low scorers improving dramatically with the TA, whereas high scorers did not.

We therefore decided to compare the subgroups of "high-performing students" ( $n = 13$ ), represented by the top quartile of pretest scorers ( $M = 33.0$ ,  $SD = 1.9$ ), to "low-performing students" ( $n = 12$ ), represented by the bottom quartile of pretest scorers ( $M = 16.5$ ,  $SD = 4.5$ ). As seen in Fig. 2, high-performers did not seem to differ neither between test versions ( $M = 34.0$ ,  $SD = 3.9$  in the TA version, and  $M = 33.7$ ,  $SD = 6.7$  in the standard version), nor improve much from their pretest. Low-performers, on the other hand, improved considerably on the TA version of the posttest ( $M = 25.5$ ,  $SD = 6.0$ ), but slightly decreased their scores on the standard version ( $M = 15.0$ ,  $SD = 6.4$ ).



**Fig. 2.** Graphs showing the pretest and posttest mean scores, for low-performing students (bottom 25% on pretest) as compared to high-performing students (top 25% on pretest), in relation to posttest version (with TA or standard).

## 4.2 Students' subjective experiences

Our second research question referred to how the TA would affect students' self-rated experiences of taking the posttest. In view of the posttest results, we chose to focus on the experience ratings by low and high performing students. See Table 1. The results showed some striking differences: low-performing students' ratings of fun were nearly twice as high for the TA version ( $M = 8.0$ ,  $SD = 2.5$ ) than for the standard version ( $M = 4.3$ ,  $SD = 4.6$ ). A comparison of the effort and confidence ratings revealed an intriguing pattern: low-performers with the standard version rated their confidence higher than their effort into doing the test, whereas low-performers with the TA version did the reverse. This pattern was not reproduced by high-performers.

**Table 1.** Mean ratings of experience items by low-performers and high-performers for the TA and standard posttest versions (on a 0-10 Likert scale, where 0 = very little, 10 = very much).

Experience item	Low-performers' rating ( <i>SD</i> )		High-performers' rating ( <i>SD</i> )	
	Standard	TA version	Standard	TA version
Fun	4.3 (4.6)	8.0 (2.5)	7.7 (1.5)	6.8 (2.6)
Ease	6.2 (3.3)	6.5 (1.8)	8.0 (1.8)	8.2 (1.8)
Effort	5.3 (3.9)	7.6 (2.9)	7.1 (1.9)	7.0 (1.6)
Confidence	6.7 (3.0)	6.1 (1.3)	8.2 (1.8)	7.4 (1.8)

## 5 Discussion

This study set out to examine how a TA from a math learning game may be used to affect students' social motivations and performance in a summative test. The results indicated that a TA might indeed be helpful in this respect, but pointed to a complexity of effects. First of all, we observed that not all students were affected equally by the presence of their TA when tested themselves. A closer analysis confirmed this observation with respect to students' previous performance level. The effects of the TA were most clearly pronounced for the low-performing students, who obtained 70% higher scores when tested with their TA compared to those tested without it. For high-performing students, the TA seemed to have little effect. Overall, experience ratings varied more for low-performers than for high-performers, and there appeared an inconsistent pattern as to how the subgroups experienced different aspects of the test.

We suggest two lines of reasoning for explaining the results. First, one may consider how the TA affects students' "mindset of teaching" [3] (in contrast to, in the present study, the mindset of taking a test). An important function of a TA is that it provides also low-performing students with a teaching opportunity, which may strengthen their sense of being able to achieve results by their own efforts. Bringing

this mindset to the test situation would likely have a positive effect on performance. For high-performing students, the situation is quite different. They are more likely to have experience from aiding and instructing other students and to have more positive associations to the situation of doing a test. This idea fits well with the results that low-performers with the TA rated their efforts higher than any other group and also rated the test itself as much more fun. Second, low-performing students may be more susceptible to a supportive social context for demonstrating what they have learned, when being under the constraints of a conventional test. The presence of a TA changes both the test form and the perception of being tested. We hold as future research questions how assessment forms can be more effectively designed to benefit from factors of social interaction in educational technologies, such as TA systems.

**Acknowledgments.** This research was financed by the Wallenberg Foundation.

## References

1. Biswas, G., Katzlberger, T., Brandford, J., Schwartz D., TAG-V.: Extending intelligent learning environments with teachable agents to enhance learning. In: Moore, J. D., Redfield, C. L., Johnson, W.L. (eds.) *Artificial Intelligence in Education*, pp. 389--397. IOS Press, Amsterdam (2001)
2. Blair, K., Schwartz, D.L., Biswas, G., Leelawong, K.: Pedagogical agents for learning by teaching: Teachable Agents. *Educational Technology Special Issue on Pedagogical Agents*. 47, 56--61 (2007)
3. Chase, C., Chin, D., Oppezzo, M., Schwartz, D.: Teachable Agents and the Protégé Effect: Increasing the Effort Towards Learning. *J. Sci. Educ. Technol.* 18, 334--352 (2009)
4. Chin, D.B., Dohmen, I.D., Cheng, B.H., Oppezzo, M.A., Chase, C.C., Schwartz, D.L.: Preparing students for future learning with Teachable Agents. *Education Tech Research Dev* 58, 649--669 (2010)
5. Gulz, A., Lindström, P., Haake, M., Pareto, L., Sjöden, B.: A Teachable Agent Based Game Affording Collaboration and Competition – Evaluating Math Comprehension and Motivation. Submitted
6. Lindström, P., Gulz, A., Haake, M., Sjöden, B.: Matching and mismatching between the pedagogical design principles of a math game and the actual practices of play. *Journal of Computer Assisted Learning* 27, 90--102 (2011)
7. Pareto, L.: The Squares Family: A Game and Story based Microworld for Understanding Arithmetic Concepts designed to attract girls. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications 1*, pp. 1567--1574 (2004)
8. Pareto, L., Schwartz, D., Svensson, L.: Learning by guiding a teachable agent to play an educational game. In: *Proceeding of the International Conference on Artificial Intelligence in Education*, pp. 662--664. IOS Press, Amsterdam (2009)
9. Schwartz, D.L., Blair, K.P., Biswas, G., Leelawong, K., Davis J.: Animations of thought: interactivity in the teachable agents paradigm. In: Lowe, R., Schnotz, W. (eds.) *Learning with animation: research and implications for design*, pp. 114--140. Cambridge University Press, Cambridge (2007)
10. Wagster J., Tan J., Wu Y., Biswas G, Schwartz D. L.: Do learning by teaching environments with metacognitive support help students develop better learning behaviors? In: *The Proceedings of the 29th Meeting of the Cognitive Science Society*, pp. 695--700. August, Nashville, USA (2007)