Young Children Can Learn Scientific Reasoning with Teachable Agents

Doris B. Chin, Ilsa M. Dohmen, and Daniel L. Schwartz

Abstract—A teachable agent (TA) is an instructional technology that capitalizes on the organizing metaphor of teaching another, in this case, a computer agent. As one instance, students teach their agents by creating concept maps that connect nodes with relational links. TAs use simple artificial intelligence to draw conclusions based on what they have been taught and to trace the path of their reasoning visually. TAs literally make thinking visible, with the goal of helping children learn to reason. TAs also provide interactive feedback and engender in students a sense of responsibility toward improving their agents' knowledge. We describe, in detail, a TA designed to teach hierarchical reasoning in science, and then present a 2-year research study using this TA with 153 fourth-grade children learning about organisms, taxonomies, and ecosystems. We show that the TA improved learning from the standard curriculum as measured by the curriculum's accompanying tests. The TA also helped children learn hierarchical reasoning, as measured by researcher-designed tests. The research indicates that, contrary to some theoretical positions, it is possible to help younger children learn scientific modes of reasoning, specifically by means of TA software.

Index Terms—Computer-assisted instruction, teachable agents, instructional design, science curriculum, science education

1 Introduction

Visual computer simulations, or animations, show changes over time. When used in science, they are often designed to make otherwise invisible processes visible (e.g., [1]). For example, they can portray changes that are too big, too small, too slow, or too fast for the naked eye. Simulations can further augment physical information by incorporating the representations that experts use to think about those processes. A physics simulation, for example, can include vectors to indicate momentum, magnetic fields, and so forth. Here, we go a step further by making the expert *reasoning processes* visible. We use teachable agents (TAs), a type of educational software, to animate the thoughts an expert might use to reason about a topic [2]. For example, using the same well-structured representations as experts, a TA can visually model how to reason through the hierarchies of scientific taxonomies. This is worthwhile for novices, because learning to emulate an expert's reasoning may be as important as learning the bare facts themselves. With TAs, students learn by teaching a computer agent, and they observe its visible thought processes while it uses what it was taught to answer new questions.

In this paper, we present two aspects of the TAs we design. One aspect is the set of three design principles that we use to guide their development: a TA makes its thought processes visible; it exhibits independent performance to produce recursive feedback; and, it engenders social responsibility in the student. Each design principle has some support from the learning literature, but the main value of TAs is that they bring these positive learning mechanisms together into a single learning technology. We provide a description of a TA called "Betty's Brain" to show an instantiation of these design principles.

The second aspect of our TAs is that they are intended to improve students' abilities to reason in specific ways, for example, causally [3] or inductively [4]. As such, we use TAs as a complement to regular classroom instruction, so they can help students learn a reasoning schema suited to the prescribed subject matter. We do not try to displace classroom curriculum or create standalone instruction by producing texts and science kits, but rather, we use TAs to augment the curricula adopted by school districts. To show how this can be done and to provide promissory evidence that it is effective for relatively young children, we use the second half of this paper to describe a research study using Betty's Brain with 153 fourth-graders (9-10 year-old children) in their regular science classrooms. We focused on helping the children reason about scientific hierarchies, such as those found in biological taxonomies. A simple example is that flies are a type of insect; all insects have six legs; therefore, flies have six legs.

A perusal of many elementary-grade science standards, for example, [5], and curricula, for example, www.lhsfoss.org, indicates that they do not emphasize how to draw valid inferences from known facts. While they do include many facts, demonstrations, and advocate careful observations, they do not explicitly teach children how to reason about relations in science. This may be a result of influential theoretical claims that young children cannot understand these types of reasoning, for example, [6], [7]. Research on children's hierarchical reasoning, which is in focus here, typically concentrates on the age at which children can solve class inclusion problems of varying difficulty,

- *D.B. Chin is with the H-STAR Institute, Stanford University, 450 Serra Mall, Building 160, Stanford, CA 94305-2055.*
- *I.M. Dohmen is with Hillbrook School, 300 Marchmont Drive, Los Gatos, CA 95032.*
- *D.L. Schwartz is with the Stanford Graduate School of Education, 485 Lasuen Mall, Stanford, CA 94305-3096.*

Manuscript received 12 Apr. 2012; revised 21 Nov. 2012; accepted 17 June 2013; published online 24 June 2013.
For information on obtaining reprints of this article, please send e-mail to: lt@computer.org, and reference IEEECS Log Number TLTSI-2012-04-0057.
Digital Object Identifier no. 10.1109/TLT.2013.24.

1939-1382/13/$31.00 © 2013 IEEE Published by the IEEE CS & ES

for example, [6], [8], [9], [10], [11], [12]. For instance, studies on children's abilities to reason about inheritance have typically found that roughly half of 9-10-year olds can successfully reason using hierarchical relations [13], [14]. Such claims about developmental maturation, however, often neglect the types of learning experiences available to the children. They may underestimate children's abilities to learn to reason given appropriate instruction [15].

Here, we show that with standard instruction, 9-10-year-old children do *not* learn to reason about inheritance in taxonomies, even though it is implicit in the scientific content they learn. Such a finding would be consistent with the prevailing hypothesis that children of this age are too young to learn how to draw disciplined inferences. However, we also show that when children have an opportunity to use a TA while learning the exact same content, the students improve their abilities to reason about hierarchical relations and they also learn the basic science content more deeply. Ideally, demonstrations like this, done in regular classroom settings with their high variability, can motivate the improvement of elementary science education so it incorporates scientific ways of thinking. And, the design principles plus the example of a classroom deployment can provide guidance on how to do so effectively.

2 Three Design Principles for Teachable Agents

A TA is a type of intelligent *pedagogical agent* [16] or *learning companion* [17], where students teach a computer character, rather than the computer character teaching the students. The teaching narrative is familiar to students, and it quickly helps to organize student-agent interactions into a teach-test-remediate cycle. The adoption of the teaching narrative also leverages positive results found in the learning-by-teaching and peer-tutoring literature [18], [19], [20], [21]. For example, people learn better when they prepare to teach someone who is about to take a test, compared to when they prepare to take the test themselves [22], [23]. They try harder to organize their understanding for the task of teaching another person than they do for themselves [24]. TAs have been used as instructional technologies to teach qualitative causal relations in science [2], [3], hypothetico-deductive reasoning [25], and a variety of mathematical topics [26], [27], [28].

In creating our TAs, we rely on three main design principles to encourage learning. We describe these principles below and instantiate them with the example of a TA designed to support the learning of hierarchical reasoning. The design principles are not intended to exclude other approaches to TA design, and there are many creative touches that one might add that go beyond any specific principle. These three simply reflect the features that have asserted themselves most strongly over the past 15 years of work with many TA variations.

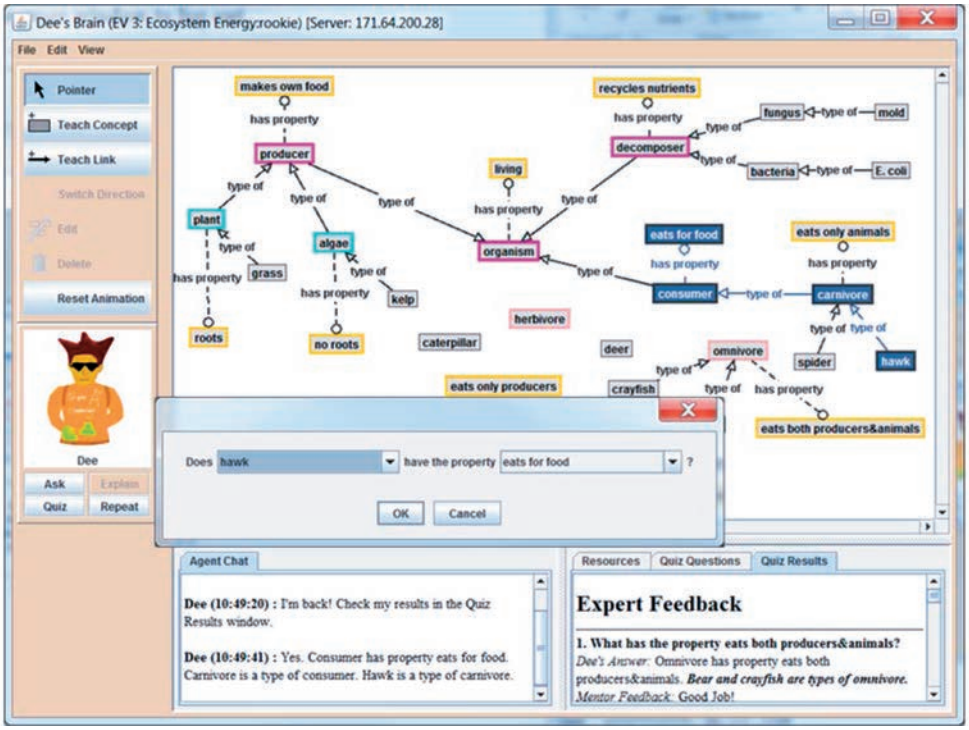
2.1 Make Thinking Visible

The first design principle is to make thinking visible. TAs help students literally see the reasoning process that their agents use to draw conclusions. In primary school science, methods of reasoning, such as causal inference or hierarchical inheritance, are rarely taught explicitly. Science curricula provide instances or demonstrations of phenomena that have causal or hierarchical relations, but the methods for reasoning about these relations are implicit. This means that students can learn the relevant facts and pairwise relations, but they may still not be able to reason with them very well. This problem is exacerbated by the fact that reasoning is largely invisible and it is difficult to induce reasoning processes through the observation of the teacher's behaviors [29]. Collins et al. [30] have proposed that sustained cognitive apprenticeships can help reveal patterns of thinking to novices. This is a heavy solution that requires displacing extant curricula with a cognitive apprenticeship curriculum. We take a lighter approach. We use TAs to make the thinking visible. So, rather than showing the scientific phenomenon, for example, through a science simulation, a TA simulates thoughts about the phenomena by tracing its reasoning visually.

Fig. 1 shows an instantiation of this design principle in the interface of the TA used to teach hierarchical reasoning. This TA is a variant of Betty's Brain [2]. The original version of Betty's Brain focused on helping students learn to construct and reason through causal chains. In response to curricular needs for younger children, we have more recently created Taxonomy Betty, which focuses on hierarchical reasoning and the learning of properties, classes, and class inclusion rules.

To teach an agent, students build the agent's brain using a concept map formalism [31], adding nodes and relational links. In Fig. 1, a student is teaching the agent about a food web and the properties of various classes of organisms. To add a concept, the student clicks on "Teach Concept," which produces a textbox in which the name of the node is entered. To create a link, the student clicks on "Teach Link" and draws a line connecting two nodes. Next, a palette appears that asks the student to specify the kind of relationship desired, which for Taxonomy Betty can be a "type-of" or "has-property" link.

Once the student has taught the TA enough nodes and links, the student can ask the agent a question, and it can answer. Using generic artificial intelligence techniques, the TA can chain through the map it has been taught (see [2]), tracing its reasoning by progressively highlighting the links and nodes it traverses. This explicitly renders the thinking visible. In Fig. 1, the student has asked the TA, "Does a *hawk* have the property of *eats for food*?" The TA highlights successive links in its concept map (shown in blue) to show the entire hierarchical chain from "hawk," up to "carnivore," up to "consumer," and finally linking to the property "eats for food." To complement its graphical reasoning, the TA also unfolds its reasoning in text (lower left panel). The graphical rendering of relational reasoning differentiates Betty's Brain from other TAs, for example, [27], [28], [32], that show the results of their reasoning, but do not graphically show how the TA reached its conclusions. Our TAs always reason correctly, which also differentiates them from other TAs that may exhibit faulty reasoning that students need to "debug," for example, [27]. Despite a TA's valid reasoning, it can still reach wrong answers if the student has taught it incorrect links and nodes. As students track down their mistakes, they trace their agent's correct reasoning (but with the wrong facts), which in turn helps them learn to reason. For example, in



Does hawk have the property eats for food ?

Dee (10:49:20) : I'm back! Check my results in the Quiz Results window.

Dee (10:49:41) : Yes. Consumer has property eats for food. Carnivore is a type of consumer. Hawk is a type of carnivore.

Expert Feedback

1. What has the property eats both producers&animals?
Dee's Answer: Omnivore has property eats both producers&animals. ***Bear and crayfish are types of omnivore.***
Mentor Feedback: Good Job!

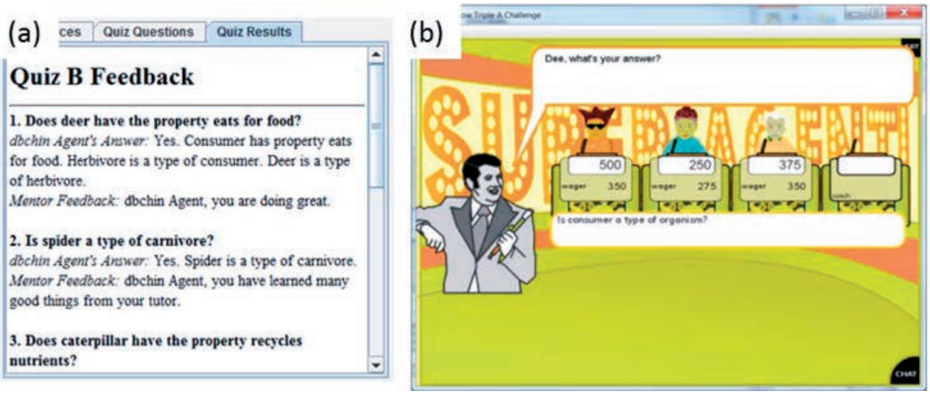
Fig. 1. The main interface for making a TA's thinking visible.

one study, 11-12-year-old students worked with a causal reasoning version of Betty's Brain to map out river ecosystems [33]. In one condition, students simply created their concept maps. In a second condition, students could quiz their agents and see the reasoning. Over time, students in the latter condition included more causal links in their maps. Similarly, in another study [3], we compared the value of TAs versus a popular concept mapping program, Inspiration (www.inspiration.com). Students in both conditions were taught to make causal links of the form "an increase/decrease in X causes an increase/decrease in Y." Inspiration students made a map like the TA students, but the Inspiration software could not reason over the map. Students in the TA condition better learned to reason through causal chains than the students in the Inspiration condition. These two studies did not parse out the relative contributions of feedback versus dynamic visualizations of reasoning, but rather, the studies indicated that together these two features work well for learning. In the study described here, we show that Taxonomy Betty has a similar learning benefit for reasoning through hierarchical chains.

2.2 Enable Independent Performance and Recursive Feedback

The second design principle is to ensure a TA can perform independently. Rather than designing an instructional system where students answer all the questions, we create a system in which the computer agent also has to answer questions on its own. The ostensibly independent performance of the TA helps sustain the narrative of teaching. It also generates a unique kind of feedback that we call *recursive feedback*. The term is based on Premack's proposal [34] that only humans teach, because only humans have the unique capacity of recursive thought—the capacity to constitute a thought as a function of itself ("I think that you think that I think..."). In the case of TAs, students constitute the thoughts of their TAs as a function of their own thinking. By observing how a TA answers and reasons with its map, a learner receives recursive feedback based on the TA's performance. This form of feedback has proven valuable as an inducement to reflect over the agent's thoughts, and by recursion, one's own thoughts [32].

Okita and Schwartz [35] tested the value of TAs' independent performance and recursive feedback compared to direct feedback. They used a TA designed for learning inductive and hypothetico-deductive reasoning. High-school students played a game with multiple levels, and on each level they had to induce a rule from available evidence and then express the rule. The rules became increasingly complex across the levels, so that the top levels had rules such as, "The presence of rain and the absence of fire is necessary but not sufficient for flowers to grow." To pass a level, students further played a prediction game in which they used the rule they had induced. The experimental difference was that students in one condition played the prediction game themselves and received direct feedback. Students in a second condition saw their agent play the game using the rule they had expressed earlier. On a post-test of inductive and hypothetico-deductive reasoning, students who received the recursive feedback of watching their agents play did substantially better than students who received direct feedback from their own game play. Presumably, one source of the effect is that students could reflect on their TA's thoughts as it made decisions, which should lead to deeper cognitive processing than simply noting that one's own answer was right or wrong. Thus, to maximize the value of a TA's independence, it is useful to ensure that the students see the feedback generated by their TA's performance.



Quiz B Feedback

1. Does deer have the property eats for food?
dbchin Agent's Answer: Yes. Consumer has property eats for food. Herbivore is a type of consumer. Deer is a type of herbivore.
Mentor Feedback: dbchin Agent, you are doing great.

2. Is spider a type of carnivore?
dbchin Agent's Answer: Yes. Spider is a type of carnivore.
Mentor Feedback: dbchin Agent, you have learned many good things from your tutor.

3. Does caterpillar have the property recycles nutrients?

Dee, what's your answer?

Is consumer a type of organism?

Fig. 2. Elements for enabling independent performance and recursive feedback.

For the current TA, Fig. 2 shows two separate mini-environments designed to enhance opportunities for recursive feedback. Panel (a) shows a quiz feature. Students can submit their TAs to take a quiz, and the system uses a hidden "expert map" or answer-key to check the agents' answers and give the students feedback on how well their agents performed. Panel (b) shows a Jeopardy-like game called Triple-A-Challenge. Students can enter their agents in the game and see how well their agents perform. A retrostyle host uses the hidden expert map to both generate questions for the agents and evaluate answers. Two additional game elements are intended to further encourage student reflection on what their agents understand. One element allows students to choose the amount of their wager, depending on whether they think their agents will answer correctly. A second element allows students, as they progress through the game, to pit their agents against smarter house agents that are more difficult to outperform or choose harder questions that involve longer chains of reasoning.

2.3 Engender Student Responsibility for the Agent

The third design principle is to engender students' responsibility toward their TAs. Enlisting social schemas can enhance people's investment in their interactions with computers [36]. Determining how to do this well is a major component of learning companion research [37]. With respect to TAs, a sense of responsibility to one's agent can lead to improved learning for students. Chase et al. [24] demonstrated the protégé effect—students worked harder to learn more on behalf of their agents than they would for themselves. Students in a Teach condition were told the agents were their pupils, and to use the software to teach their agents. Students in the Self condition were told that the characters represented themselves, and they were just using new educational software. Thus, the only manipulation was the narrative of teaching. Students in the Teach condition spent nearly twice as long reading about the topic they were teaching compared to the Self students. One speculation for why TAs helped was that when students' agents gave the wrong answer, students had an "ego-protective buffer"—their TAs got it wrong, not them. This was especially true for the lowest achieving students in the classes. Despite common beliefs that feedback is always good, for lower achieving students, negative feedback can trigger ego-involvement such that students make negative attributions about themselves, rather than using the feedback to improve task performance [38]. Evidence for an ego-protective buffer was found in the students' statements following correct and incorrect answers during game play. This ego-protection allowed Teach students to persist longer, compared to the Self students, who had to take full blame for getting wrong answers.

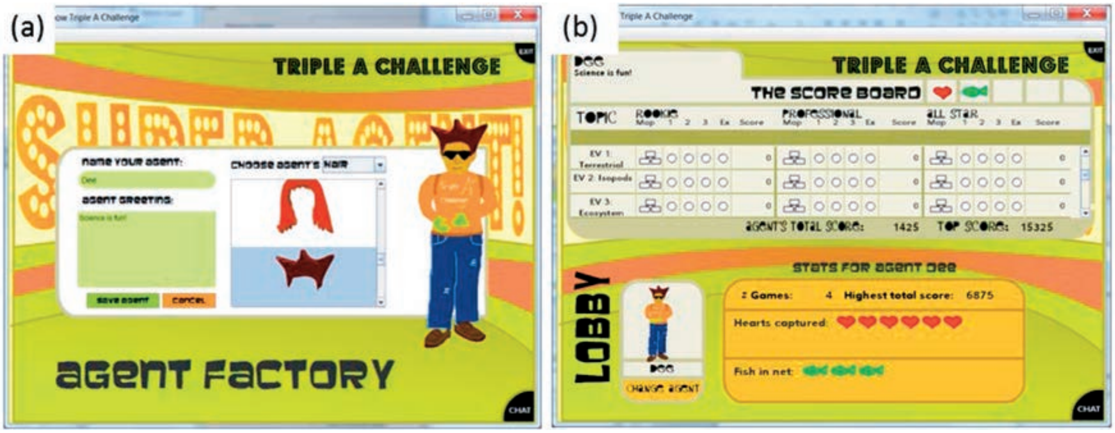
The overarching narrative of TAs enlists the familiar social schema of responsibility toward one's pupils. To further enhance responsibility, the TA system requires that students name and customize the look of their agents, as shown in panel (a) of Fig. 3. The Triple-A-Challenge game show is also embedded in a larger environment shown in panel (b). Students can earn hearts and fish when their TAs win in the game show. They can then redeem the hearts and fish to release new customization options, allowing them to further individuate their agents. There are, of course, other ways to increase the sociability of agents, including more sophisticated dialog systems, more advanced graphics, and placing the TAs in a social environment with other agents (e.g., [37], [39], [40]).

3 Demonstration Experiment

Given the three design principles and the description of the TA environment, we now turn to the second component of the paper—a demonstration of using TA to improve students' hierarchical reasoning. We describe a classroom study that used Taxonomy Betty with fourth-graders (9-10-years old), replicated over 2 years.

At this age, school science includes a number of topics that involve classification. However, typical instruction rarely models the appropriate kind of reasoning for class inclusion relations and does not provide tests that directly evaluate this type of reasoning. We thought that having students tutor Taxonomy Betty would help them focus on thinking *per se.*

Fig. 4 may help clarify how the three design principles could be especially helpful for students learning hierarchical reasoning. The figure provides a simple schematized example of a good and bad map on the left, and how the agent would indicate its reasoning on the right. First, take the case of a fictitious student who does not have this visualization, but instead, answers a question from a flashcard-like computer program that presents questions



TOPIC	ROOKIE Map	1	2	3	Ex	Score	PROFESSIONAL Map	1	2	3	Ex	Score	ALL STAR Map	1	2	3	Ex	Score
EV 1: Terrestrial	品	○	○	○	○	0	品	○	○	○	○	0	品	○	○	○	○	0
EV 2: Isopods	品	○	○	○	○	0	品	○	○	○	○	0	品	○	○	○	○	0
EV 3: Ecosystem	品	○	○	○	○	0	品	○	○	○	○	0	品	○	○	○	○	0

AGENT'S TOTAL SCORE: 1425 TOP SCORE: 15325

STATS FOR AGENT DEE

Games: 4 Highest total score: 6875

Hearts captured: ❤❤❤❤❤❤

Fish in net: 🐟🐟🐟

Fig. 3. Elements that foster a sense of responsibility for the TA.

and offers right/wrong feedback. Imagine the student enters a correct answer to the question of whether a ladybug has six legs. She could have achieved this correct answer based on direct memory of the fact that ladybugs have six legs, or she might have guessed the right answer, both of which are undesirable when the goal is learning to reason about taxonomies. In contrast, consider a student who has been working with Taxonomy Betty and produced the good map in the left panel of Fig. 4a. The right panel shows how the student can visually follow along as the TA figured out the answer by connecting ladybug to insects, which have six legs. TA students see how they are supposed to figure out the answer and how to organize their knowledge. This is a simple example of the first design principle to make thinking visible. Hierarchical reasoning is particularly easy to make visible, so it should be an ideal application of TA. One could also imagine a TA version that uses Venn or Euler diagrams to show nested relations, instead of nodes and labeled links, and these alternative visualization schemes could also work quite well.

Next, take the opposite case where the flashcard student gives an incorrect answer to the question and receives negative feedback that she is wrong. It would be hard for her to figure out the mistake, because she is trapped by her own thoughts—after all, she thought her answer was right to start with. In contrast, consider a TA student who had produced the bad map in Fig. 4b. The right panel shows how the student can see the implications of how he organized his agent's knowledge and how the map interacts with the agent's proper reasoning behavior. This enables TA students to reflect explicitly on the reasoning process and notice how their knowledge organization led to the wrong conclusions. This is an example of the second principle to provide recursive feedback through independent performance. Students see how someone else reasons about their knowledge, in this case a TA, which presumably helps them clarify their own knowledge and reasoning.

Finally, consider again the case where the flashcard student gets the wrong answer and receives negative feedback. It is an important metacognitive strategy to use negative feedback as a chance to learn. Nevertheless, negative feedback often causes students to avoid the situation that generates the feedback [41]. To mitigate this problem, TA provides students with an "ego-protective" buffer—the wrong answer can be attributed to the agent, so the student does not shy away from the negative feedback. To motivate students to use the feedback and help their agent, it further helps to catalyze the student's sense of responsibility. Hence, the third principle: engender students' responsibility to the agent.

Despite our thought experiment and rationale for the design principles, it is also possible that TAs do not help at

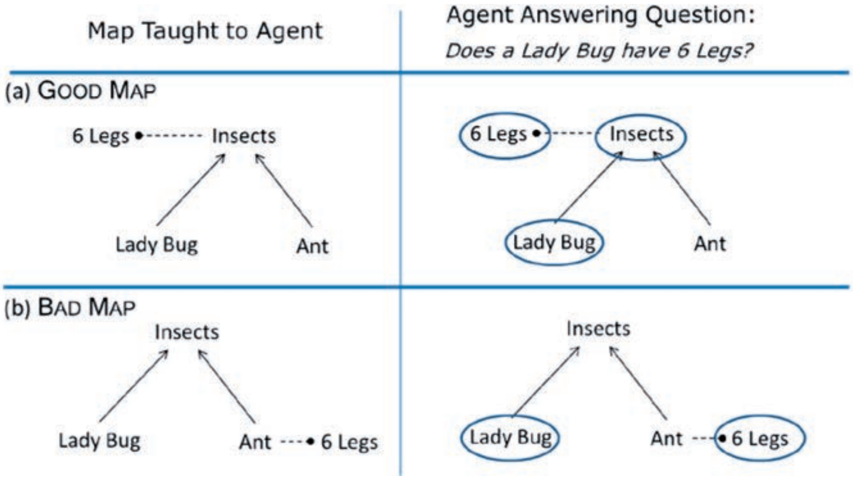


Fig. 4. Schematized examples of hierarchical organization and reasoning in TA.

all. The visual representation may just bring more complexity, the recursive feedback may be obscure because it was not the student's own answer, and the students may feel less responsible to their agents than they do to their own understanding. Moreover, there is the question of whether simply adding TAs to an extant curriculum will be so incongruous and distracting that it interferes with the intended trajectory of the curriculum. We have reasons to believe this is not the case. Earlier versions of TAs, based on the same design principles, have been successful in helping older science students learn to think in terms of causal relations. For example, in one series of studies, Chin et al. [3] found that TAs helped late primary and middle school students use causal reasoning better on subsequent lessons involving a different science topic, even when the students were no longer using the TA system. These studies involved children in the 10-12-year-old range, and the TA modeled causal chaining. Here, we moved to a younger age, and we reprogrammed the TA so it showed how to reason about taxonomies. Thus, a major goal of the research was to determine whether the TA platform would generalize to help younger children learn, and whether it could generalize to teaching another type of reasoning, in this case, hierarchical reasoning.

The research was not an attempt to isolate any single factor of the TA system with strict controls on implementation, but rather to see how the system as a whole would fare when used in regular classrooms with teachers who could choose to use it as they wanted. Our hope is that the TA design principles work despite reasonable variation in teacher implementation.

In the current study, control students completed the standard practice of the school district, whereas the TA students completed the standard practice plus the TA. The total amount of instructional time was the same for both treatments. We included two types of learning measures. One measure was the test that came with the curriculum. We will label this the "basic-value" measure, because it measures the value of the extant curriculum on its own terms. We will label the second measure "added-value," because this was researcher-designed to see if TAs would improve children's abilities to reason through taxonomies. Our leading hypothesis was that an emphasis on hierarchical reasoning would not only improve performance on our added-value measures, but it would also improve student understanding of the scientific content *per se*, as measured by the curriculum's own basic-value measures. This would demonstrate that teaching scientific reasoning at this age is useful, and that TAs are one recipe for success. An alternative hypothesis is that the time spent with TAs would unproductively displace time spent learning the basic science, and therefore the students using TAs would do worse on the basic-value test, regardless of their performance on the added-value measures.

3.1 Methods

3.1.1 Participants

A small, public school district agreed to use the TA technology to complement their regular, kit-based science curriculum in the fourth-grade. The district is high SES, with a reported 4 percent socioeconomically disadvantaged population, 7 percent English learners, and ethnicity of 71 percent white, 7 percent Asian, 6 percent Hispanic, 2 percent African American, and 14 percent other/not reported. A total of seven classes participated in our study over 2 years. District officials called for voluntary participation. In the first year, two teachers (Teachers A and B) agreed to use TAs with their students ($n = 49$). They integrated TAs into their regular science lesson plans (Kit+TA condition). Two other teachers (Teachers C and D) volunteered as control classrooms ($n = 34$) and conducted their science lessons as they normally would during this same time period (Kit-Only condition). The study was replicated the following year using the same two Kit+TA teachers ($n = 46$), while a new teacher volunteered as a Kit-Only teacher (Teacher E, $n = 24$). All told, there were two treatment teachers who were sampled twice, and three control teachers each sampled once. We recognize the unavoidable confound of allowing teachers to self-select into treatment, and discuss this in the Results.

3.1.2 Materials

The district-adopted science curriculum is the Full Option Science System (FOSS), developed by the Lawrence Hall of Science (www.lhsfoss.org). FOSS kits come complete with teacher guides, textbooks, videos, hands-on activities, worksheets, and assessments. The FOSS curriculum used in this study was the *Environments* kit, which focuses on organisms and ecosystems, and is rich in class inclusion and property relations.

To tailor TA for an extant curriculum, the only requirement is that the teacher, curriculum specialist, or researcher produces the relevant hidden answer-key or "expert" maps. The hidden expert maps allow the TA system to generate questions and feedback for the agents. In this case, the FOSS *Environments* kit contains five subunits, and we produced four hierarchical expert maps for the curriculum. Figs. 5a and 5b show the expert maps for two subunits: the relatively simple, introductory map on terrestrial environments (13 nodes and 12 links), and then a much more complex map for the third subunit on food energy webs (30 nodes and 29 links). The Kit+TA students did not see these maps, but instead, for each map, they received the nodes, unconnected at the bottom of the screen. Their task was to teach their agents by connecting the nodes. Figs. 5c and 5d show sample student maps matched for the same subunits. (The Kit-Only students never explicitly received a list with the nodes, however, most of the nodes were the concepts bolded in the FOSS textbook and highlighted in the chapter summaries and glossary.)

Learning gains were measured with identical pre- and post-tests. The basic-value test is packaged with the FOSS kit and includes 10 multiple-choice, fill-in, and short-answer questions. (Actual questions may be purchased from FOSS.) The two added-value measures were specifically designed by us to probe for hierarchical reasoning. Examples of two basic- and all the added-value items are in Table 1.

The added-value measures were designed to evaluate whether students would, without specific prompting, use hierarchical relations to organize their answers, just as their TAs had visually modeled for them. The first added-value item ("What is a ladybug?") was deliberately open-ended. The expectation was that Kit+TA students would spontaneously provide more hierarchical class and property

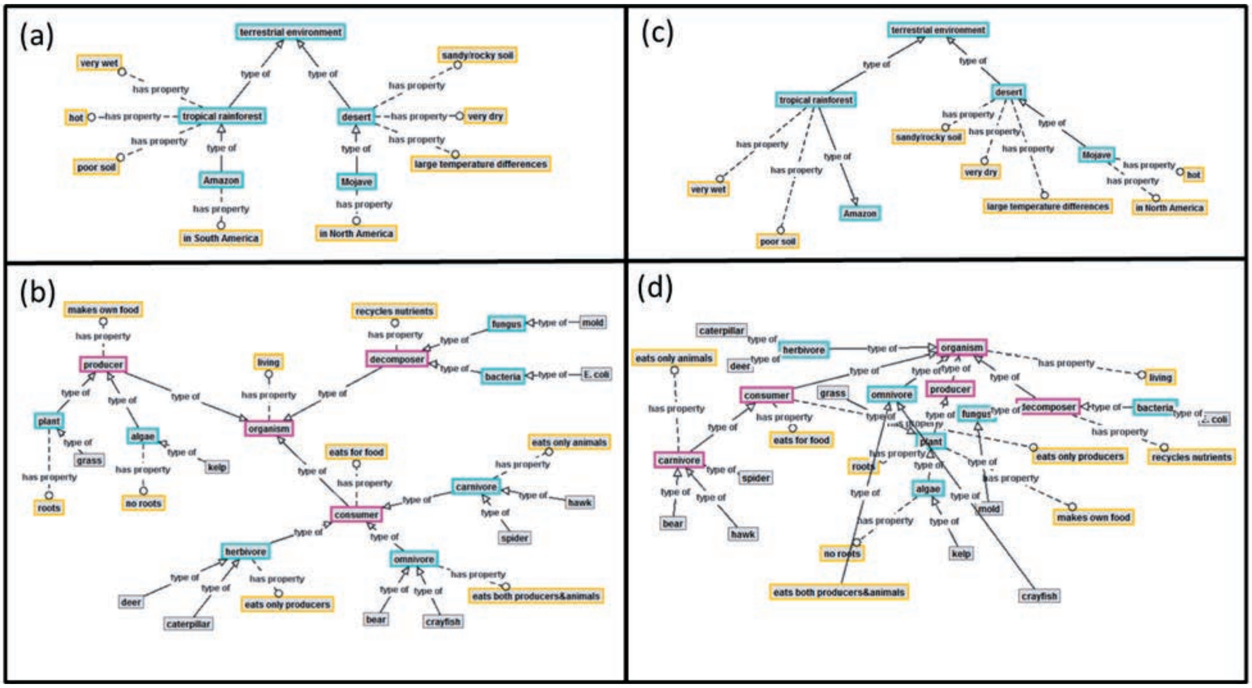


Fig. 5. Examples of expert and student maps for FOSS *Environments* kit.

information (this item was coded separately for both types of information). The second added-value item ("Write these words in the boxes below...") was intended to elicit the conceptual organization students had developed.

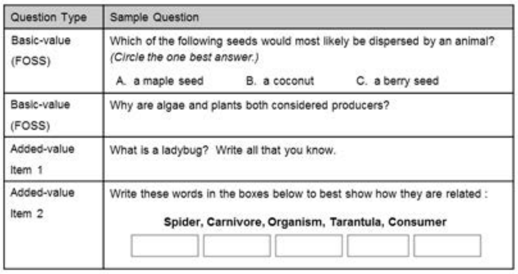
All measures were coded with a rubric and used the scale: 0 points (incorrect/no answer), 1/2 point (partially correct), or 1 point (correct answer). Inter-rater reliability between two coders (one blind), was $r > 0.93$ for both basic and added-value tests using a random sample of 25 percent of all responses. Cronbach's alpha for reliability across all measures was 0.81.

Additional measures collected include students' standardized achievement scores (STAR assessments, www.cde.ca.gov/ta/tg/sr/) for English language arts (ELA) and math, taken from both the year before and the end of the year during which students were in the study.

3.1.3 Design and Procedures

The research design had the between-subjects factor of instructional treatment (Kit+TA versus Kit-Only). It also had the within-subjects factor of time (pre- versus post-test) that was measured on two test types (basic- and added-value).

TABLE 1
Examples of Assessment Items



Question Type	Sample Question
Basic-value (FOSS)	Which of the following seeds would most likely be dispersed by an animal? *(Circle the one best answer.)* A. a maple seed B. a coconut C. a berry seed
Basic-value (FOSS)	Why are algae and plants both considered producers?
Added-value Item 1	What is a ladybug? Write all that you know.
Added-value Item 2	Write these words in the boxes below to best show how they are related : **Spider, Carnivore, Organism, Tarantula, Consumer** □ □ □ □ □

Overall, the instruction lasted approximately eight weeks for each teacher, fitting the timeline of the school district. Kit-Only teachers used the kit as they normally would in this period, without interference from the researchers other than the addition of added-value questions for the pre-post assessments. Researchers trained Kit+TA teachers on the TA software in one-on-one sessions prior to the first lesson. On the first lesson day, to ensure that all students received the same social narrative of teaching their agent, researchers introduced Kit+TA classes to the software and showed students how to "take responsibility" for their agents by first customizing them (see Fig. 3a), then teaching and quizzing them (see Figs. 1 and 2a). Subsequently, researchers only provided technical and classroom support to teachers as requested, in particular, the introduction of the game element (see Fig. 2b). Kit+TA teachers were asked to use the TA system with their students at least one time for each of the four concept maps. They did not add more days to the overall instructional time to accommodate the TA activities. Otherwise, we did not prescribe how or when to use TAs for each concept map, because we wanted to see if TAs would be effective within the variations of the teachers' self-chosen integration into the regular instruction.

Over the course of the unit, TA students logged into the system an average of 9.6 times ($\text{SD} = 4.7$), made 167.4 map edits ($\text{SD} = 56.8$), had their agents take 34.1 quizzes ($\text{SD} = 19.3$), accessed the online reading resources 35.4 times ($\text{SD} = 25.1$), and played 21.7 games ($\text{SD} = 19.7$). Observational records indicated that the two Kit+TA teachers had different patterns of usage. Teacher A tended to control system use, limiting TA sessions to near the end of the subunits as checkpoints, and emphasizing the quizzes to her students as a valuable form of feedback on the quality of

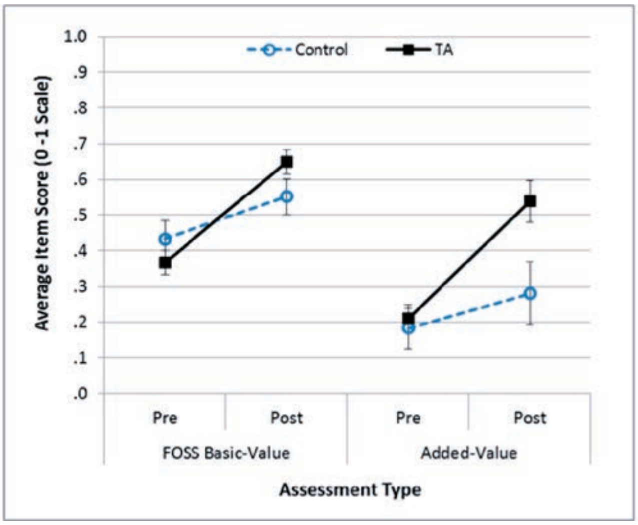


Fig. 6. Average performance on assessment items—six classes represented, 95 percent C.I.'s on means.

their agent and their teaching. Teacher B was more free-form, interspersing TA sessions more frequently throughout the unit, and allowed her students more unconstrained access to the software, valuing the more social elements of the software, for example, the game show, as ways to engage her students. Univariate tests of the log data files supported these observations. Teacher A's students quizzed their agents more, $F(1, 93) = 14.6, p < 0.001$, and teacher B's students logged on more, $F(1, 93) = 63.2$, $p < 0.001$, and played more games $F(1, 93) = 52.9, p < 0.001$.

3.2 Results

3.2.1 Learning Gains

Prior to the instructional experiment, the students in the two conditions had statistically indistinguishable scores for ELA, math achievement, and added-value measures, all F's < 0.8. The Kit-Only (control) students, however, started significantly higher on the basic-value measures, $F(1, 140) = 5.2, p = 0.024$, as shown in Fig. 6. Even so, after instruction, the Kit+TA students outperformed the Kit-Only students for both basic- and added-value measures.

For the first statistical analyses, we only use those students for whom we have complete data and we exclude students from one of the control classes (year 1) that did not take the basic-value post-test. The basic- and added-value scores were entered as doubly repeated measures in a treatment (Kit+TA versus Kit-Only) by time (pre- versus post-test) repeated measures analysis. There was an overall effect of time; $F(2, 130) = 113.3, p <$ 0.001. Students improved on both the basic-value measures $F(1, 131) = 197.3$, $p < 0.001$, and the added-value measures $F(1, 131) = 64.9$, $p < 0.001$. There was also an overall effect of treatment; $F(2, 130) = 9.7, p < 0.001$. More importantly, there was a significant time x treatment interaction; $F(2, 130) = 21.9$, $p < 0.001$. The Kit+TA students improved more on both measures; basic-value $F(1, 131) = 32.8, p < 0.001$; added-value $F(1, 131) = 19.0$, $p < 0.001$. The effect sizes for the learning gains of the Kit+TA condition over the Kit-Only condition were $d =$ 1.1, and $d = 0.93$ for the basic- and added-value measures, respectively.

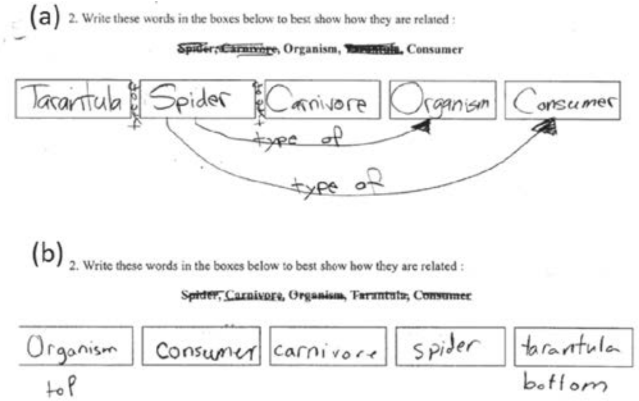

(a)
2. Write these words in the boxes below to best show how they are related :
~~Spider~~, ~~Carnivore~~, Organism, ~~Tarantula~~, Consumer
Tarantula
Spider
Carnivore
Organism
Consumer
type of
type of
(b)
2. Write these words in the boxes below to best show how they are related :
~~Spider~~, ~~Carnivore~~, ~~Organism~~, ~~Tarantula~~, ~~Consumer~~
Organism
Consumer
carnivore
spider
tarantula
top
bottom

Fig. 7. Examples of student post-test answers, added-value item 2.

Teacher C did not administer the basic-value post-test, and was not included in the preceding analysis. A separate analysis reincorporated the added-value data for this control class with the other classes and yielded similar results. There was a main effect of time, $F(1{,}146) = 76.1$, $p < 0.001$, and treatment, $F(1{,}146) = 13.6, p < 0.001$. The key time x treatment interaction was also significant, $F(1{,}146) = 29.1, p < 0.001$. (When including the third control class into the added-value means (not shown in Fig. 6), the values did not change appreciably, $\mathrm{M}_{\mathrm{pre}} = 0.21, \mathrm{SE} = 0.02$, and $\mathrm{M}_{\mathrm{post}} = 0.29, \mathrm{SE} = 0.03$.)

The preceding analyses combined the two cohorts to increase the overall sample size. One may also interpret cohort 2 as a second study (albeit with the same Kit+TA teachers). To determine if the results replicated from cohort 1 to cohort 2, we added cohort (year 1 versus year 2) as a between-subjects variable. Treatment was the second between-subjects variable, with time and test-type as within-subject measures. The cohort factor did not interact at any level with the time factor, all F's < 1.5, and the other findings held up. This indicates that the key time × treatment interaction replicated over both years and for both measures.

3.2.2 Transfer of TA's Concept Map Formalism

Our assumption (and intended design) is that students entrain on the graphical way that the TA organizes information and reasons. To examine this assumption, we coded for the spontaneous inclusion of hierarchy-relevant spatial information on added-value item 2 (see Table 1). Fig. 7 provides two positive examples.

None of the Kit-Only students incorporated spatial information in their answers, whereas 13 percent of Kit-TA students did, $\chi^2(1, \mathrm{N} = 150) = 8.2, \mathrm{p} = 0.004$. While 13 percent is modest, the question did not explicitly require this information. At the least, this result may be interpreted as an indication that the TA system helped some students incorporate visual or spatial organizations of hierarchies into their repertoire of cognitive skills.

3.2.3 Teacher Effects

Due to the wishes of the school district, we could not require teachers to participate nor randomly assign them to treatment. Teachers who self-selected to the Kit-TA condition may have caused the treatment differences rather than Taxonomy Betty. While this cannot be definitively refuted by the current research design, there are three lines of evidence that mitigate the concern.

First, all four of the Kit+TA classes showed greater learning gains than all three of the Kit-Only classes. We computed percent of possible gain (100 × (post-pre mean item score)/(1 − pre mean item score)). For added-value measures, the TA classrooms had gain scores of 40.5 percent, 45.8 percent (year 1), and 43.1 percent, 36.4 percent (year 2). The control classrooms had added-value gain scores of 1.3, -10.1, and 16.5 percent. For basic-value measures, the TA classrooms had gain scores of 43.5 percent, 44.5 percent (year 1) and 49.8 percent, 40.1 percent (year 2). The control classrooms exhibited basic-value gain scores of 18.7 and 20.4 percent (one teacher did not give the basic-value posttest). By these nonoverlapping distributions, the effect of treatment is substantially larger than the effect of teacher (or year) within each treatment.

Second, the Kit+TA teachers were not better teachers based on California's standardized tests. These tests are administered yearly to evaluate how much teachers are helping students gain over the prior year, among other things. A doubly repeated measures analysis crossed time of test (incoming versus outgoing) by treatment (Kit-Only versus Kit+TA) using both ELA and math achievement scores. There was no treatment effect, $F(1{,}141) = 0.81$, $p = 0.37$, and no treatment x time interaction (all F'$s < 1.94$). By these measures, there is no evidence that TA teachers were just better teachers.

Third, the two Kit+TA teachers had very different approaches to how they used the system, as summarized in Section 3.1.3. For example, subsequent analyses of the log files showed that number of map edits correlated with learning gains for Teacher B, who allowed children more choice in how to use the system, but not for Teacher A, who exerted more control over student activity. Yet, their students exhibited similar learning gains. This means that the teachers who self-selected to the TA condition were not uniform in implementation, yet all their classes did better than the control classes.

4 Conclusion

TAs provide the overarching metaphor and core mechanic of teaching as a means for learning. We provided three design principles for creating TA systems: making thinking visible, enabling independent performance and recursive feedback, and finally, engendering a sense of social responsibility toward the agent. Each of these principles is based in research from the learning literature, but the main value of TAs is that they can bring a harmonious confluence of positive learning mechanisms together into a single learning technology.

We also described the results of a demonstration study that implemented TAs in regular 4th-grade classrooms as a complement, rather than replacement, for current curriculum. We allowed teachers to use TAs as they wished, so we could gather some initial evidence on whether TAs could handle variation in "real world" classroom implementations. When integrated into the school's kit-based science curriculum, all TA classes exhibited greater learning gains by both the kit's own measures and by our measures of hierarchical reasoning. So rather than displacing the value of the original curriculum, the results suggest that TAs provided a way for students to organize the facts they learned from the science kit, which had a leveraging effect on how much they learned overall from the kit lessons.

Teachers' self-selection into the experiment and their specific treatment created an unavoidable confound to our study. We did not find evidence to support the alternative hypothesis that the results of the study were due to teacher effects rather than treatment effects, although it is still possible. A second class of alternative hypotheses involves novelty effects. Perhaps it was simply the general presence of technology in the classroom, rather than TAs specifically, that enhanced children's motivations or classroom experience and led to their learning gains. This seems unlikely to be a full account of the differences because specific aspects of system use (number of map edits) correlated with learning gains. Additionally, the spontaneous transfer of hierarchical thinking on the post-test is selective to the precise relations that the TAs modeled. A nonspecific "halo" effect should not show these patterns.

Another possibility is that the Kit+TA teachers may have been more enthusiastic about using technology in their classrooms, which is why they self-selected to the TA condition. If this caused the gains, then a reasonable conclusion is that TAs are successful when implemented by teachers who are enthusiastic about their use.

At a minimum, our results indicate that it is possible to improve younger students' hierarchical reasoning, which also leads to better learning of the science content itself. This provides an argument for including scientific reasoning in the primary-school curriculum. At a maximum, our results suggest the value of creating additional TAs that visually model other forms of age-appropriate reasoning, ranging from legal to statistical to metacognitive. The three design principles offer suggestions for going forward.

Acknowledgments

This material is based upon work supported by the Institute of Education Sciences within the US Department of Education under Award no. R305H060089, and the US National Science Foundation under grant nos. 0634044 and 0904324. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the granting agencies.

References

[1] W.K. Adams, S. Reid, R. LeMaster, S.B. McKagan, K.K. Perkins, M. Dubson, and C.E. Wieman, "A Study of Educational Simulations Part I - Engagement and Learning," *J. Interactive Learning Research,* vol. 19, no. 3, pp. 397-419, July 2008.

[2] G. Biswas, K. Leelawong, D.L. Schwartz, N. Vye, and TAG-V, "Learning by Teaching: A New Agent Paradigm for Educational Software," *Applied Artificial Intelligence,* vol. 19, pp. 363-392, 2005.

[3] D.B. Chin, I.M. Dohmen, B.H. Cheng, M.A. Oppezzo, C.C. Chase, and D.L. Schwartz, "Preparing Students for Future Learning with Teachable Agents," *Educational Technology Research and Development,* vol. 58, no. 6, pp. 649-669, 2010.

[4] S.Y. Okita and D.L. Schwartz, "Learning by Teaching Humans and Computer Agents," *J. Learning Sciences,* to be published.

[5] Nat'l Research Council, *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas.* Nat'l Academies, 2012.

[6] J. Piaget, *The Equilibration of Cognitive Structures.* Univ. of Chicago, 1985.

[7] K.E. Metz, "Reassessment of Developmental Constraints on Children's Science Instruction," *Rev. of Educational Research,* vol. 65, no. 2, pp. 93-127, 1995.

[8] P. Cormier and Y. Dagenais, "Class-Inclusion Developmental Levels and Logical Necessity," *Int'l J. Behavioral Development,* vol. 6, no. 1, pp. 1-14, 1983.

[9] V.J. Haars and E.J. Mason, "Children's Understanding of Class Inclusion and Their Ability to Reason with Implication," *Int'l J. Behavioral Development,* vol. 9, no. 1, pp. 45-63, 1986.

[10] U. Müller, B. Sokol, and W.F. Overton, "Developmental Sequences in Class Reasoning and Propositional Reasoning," *J. Experimental Child Psychology,* vol. 74, no. 2, pp. 69-106, 1999.

[11] G.S. Halford, G. Andrews, and I. Jensen, "Integration of Category Induction and Hierarchical Classification: One Paradigm at Two Levels of Complexity," *J. Cognition and Development,* vol. 3, no. 2, pp. 143-177, 2002.

[12] J. Deneault and M. Ricard, "The Effect of Hierarchical Levels of Categories on Children's Deductive Inferences about Inclusion," *Int'l J. Psychology,* vol. 40, no. 2, pp. 65-79, 2005.

[13] G.A. Winer, "Class-Inclusion Reasoning in Children: A Review of the Empirical Literature," *Child Development,* vol. 51, pp. 309-328, 1980.

[14] G.S. Halford, *Children's Understanding: The Development of Mental Models.* Lawrence Erlbaum, 1993.

[15] D.L. Schwartz, T. Martin, and J. Pfaffman, "How Mathematics Propels the Development of Physical Knowledge," *J. Cognition and Development,* vol. 6, pp. 65-88, 2005.

[16] A.L. Baylor, "Pedagogical Agents as a Social Interface," *Educational Technology,* vol. 47, no. 1, pp. 11-14, 2007.

[17] C.Y. Chou, C.J. Lin, and T.W. Chan, "An Approach to Developing Computational Supports for Reciprocal Tutoring," *Knowledge-Based Systems,* vol. 15, no. 7, pp. 407-412, 2002.

[18] L. Annis, "The Processes and Effects of Peer Tutoring," *Human Learning,* vol. 2, pp. 39-47, 1983.

[19] A. Renkl, "Learning for Later Teaching: An Exploration of Mediational Links between Teaching Expectancy and Learning Results," *Learning and Instruction,* vol. 5, pp. 21-36, 1995.

[20] R.D. Roscoe and M. Chi, "Tutor Learning: The Role of Explaining and Responding to Questions," *Instructional Science,* vol. 36, pp. 321-350, 2008.

[21] J.A. Bargh and Y. Schul, "On the Cognitive Benefits of Teaching," J. Educational Psychology, vol. 72, pp. 593-604, 1980.

[22] G. Biswas, D.L. Schwartz, and J.D. Bransford and The Teachable Agents Group at Vanderbilt, "Technology Support for Complex Problem Solving: From SAD Environments to AI," *Smart Machines in Education,* K. Forbus and P. Feltovich, eds., pp. 71-98, AAAI/MIT, 2001.

[23] L. Martin and D.L. Schwartz, "Prospective Adaptation in the Use of Representational Tools," *Cognition and Instruction,* vol. 27, no. 4, pp. 1-31, 2009.

[24] C. Chase, D.B. Chin, M. Oppezzo, and D.L. Schwartz, "Teachable Agents and the Protégé Effect: Increasing the Effort toward Learning," *J. Science Education and Technology,* vol. 18, no. 4. pp. 334-352, 2009.

[25] K.P. Blair and D.L. Schwartz, "Milo and J-Mole: Computers as Constructivist Teachable Agents," *Proc. Sixth Int'l Conf. Learning Sciences,* pp. 588-588, 2004.

[26] L. Pareto, "A Teachable Agent Game for Elementary School Mathematics Promoting Causal Reasoning and Choice," *Proc. First Int'l Workshop Adaptation and Personalization in EB/Learning Using Pedagogic Conversational Agents (APLEC '10),* pp. 13-19, 2010.

[27] L. Pareto, M. Haake, P. Lindström, B. Sjödén, and A. Gulz, "A Teachable-Agent-Based Game Affording Collaboration and Competition: Evaluating Math Comprehension and Motivation," *Educational Technology Research and Development,* vol. 60, pp. 723-751, 2012.

[28] N. Matsuda, W.W. Cohen, K.R. Koedinger, V. Keiser, R. Raizada, E. Yarzebinski, and G. Stylianides, "Studying the Effect of Tutor Learning Using a Teachable Agent that Asks the Student Tutor for Explanations," *Proc. IEEE Fourth Int'l Conf. Digital Game and Intelligent Toy Enhanced Learning (DIGITEL),* pp. 25-32, 2012.

[29] M.T.H. Chi, M. Roy, and R.G.M. Hausmann, "Observing Tutorial Dialogues Collaboratively: Insights about Human Tutoring Effectiveness from Vicarious Learning," *Cognitive Science: A Multidisciplinary J.,* vol. 32, no. 2, pp. 301-341, 2008.

[30] A. Collins, J.S. Brown, and A. Holum, "Cognitive Apprenticeship: Making Thinking Visible," *Am. Educator,* vol. 15, no. 3, pp. 6-11 and 38-46, 1991.

[31] J.D. Novak and D.B. Gowin, *Learning How to Learn.* Cambridge Univ., 1984.

[32] S.Y. Okita and A. Jamalian, "Learning from the Folly of Others: Learning to Self-Correct by Monitoring the Reasoning of Projective Pedagogical Agents," *Proc. 10th Int'l Conf. Learning Sciences (ICLS),* pp. 281-285, 2012.

[33] D.L. Schwartz, K.P. Blair, G. Biswas, K. Leelawong, and J. Davis, "Animations of Thought: Interactivity in the Teachable Agents Paradigm," *Learning with Animation: Research and Implications for Design,* R. Lowe, R. and W. Schnotz, eds., pp. 114-140, Cambridge Univ., 2007.

[34] D. Premack, "Is Language the Key to Human Intelligence," *Science,* vol. 303, no. 5656, pp. 318-320, 2004.

[35] S.Y. Okita and D.L. Schwartz, "When Observation Beats Doing: Learning by Teaching," *Proc. Seventh Int'l Conf. Learning Sciences (ICLS),* 2006.

[36] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television and New Media like Real People and Places.* Stanford Univ. Center for the Study of Language and Information, 1996.

[37] Z.-H. Chen, C. Liao, Q.-C. Chien, and T.-W. Chan, "Animal Companions: Fostering Children's Effort-Making by Nurturing Virtual Pets," *British J. Educational Technology,* vol. 42, no. 1, pp. 166-180, 2011.

[38] A.N. Kluger and A. DeNisi, "Feedback Interventions: Toward the Understanding of a Double-Edged Sword," *Current Directions in Psychological Science,* vol. 7, no. 3, pp. 67-72, 1998.

[39] A. Gulz, M. Haake, and A. Silvervarg, "Extending a Teachable Agent with a Social Conversation Module-Effects on Student Experiences and Learning," *Artificial Intelligence in Education,* pp. 106-114, Springer Berlin/Heidelberg, 2011.

[40] J.R. Segedy, J.S. Kinnebrew, and G. Biswas, "The Effect of Contextualized Conversational Feedback in a Complex Open-Ended Learning Environment," *Educational Technology Research and Development,* rapid post, 2012, doi:10.1007/s11423-012-9275-0.

[41] J. Hattie and H. Timperley, "The Power of Feedback," *Rev. of Educational Research,* vol. 77, no. 1, pp. 81-112, 2007.

Doris B. Chin received the PhD degree in genetics from the University of California, Davis. She is a researcher at Stanford University's Graduate School of Education. She also spent many years as a science museum educator, an experience which provides her a unique, perhaps tilted, perspective on academic research.

Ilsa M. Dohmen received the BA degree in biological anthropology and english from Tufts University and the MA degree in anthropological sciences from Stanford University. She currently teaches middle school science. She served as a researcher at Stanford University's Center for Innovations in Learning.

Daniel L. Schwartz received the PhD degree in human learning and cognition from Columbia University. Before that, he taught middle school for many years. He is a professor of Education at Stanford University and the director of the AAALab. His niche is the ability to bridge basic research on human cognition with creative designs and experiments to improve STEM learning, often using computer technology.