

Title: Preparing Students for Future Learning with Teachable Agents

Author(s):

Doris B. Chin¹, Ilsa M. Dohmen¹, Britte H. Cheng³, Marily A. Oppezzo², Catherine C. Chase², and Daniel L. Schwartz²

¹ [*Stanford Center for Innovations in Learning*](#), Stanford University, 450 Serra Mall, Building 160, Stanford, CA, 94305-2055

² [*School of Education*](#), Stanford University, 485 Lasuen Mall, Stanford, CA, 94305-3096

³ [*Center for Technology in Learning*](#), SRI International, 333 Ravenswood Avenue, Menlo Park, CA, 94025-3493

Background/context:

Over the past several years, we have been developing an instructional technology, called Teachable Agents (TA), which draws on the social metaphor of teaching to help students learn. Students teach a computer character, their “agent,” by creating a concept map of nodes connected by qualitative causal links (e.g., *A increases B*) (insert Figure 1 here). Figure 1 shows the main interface for the TA software, in which the concept map symbolizes the interior of the agent’s brain. The student has taught her agent, “Dee,” about global warming, including the individual propositions that methane is a type of greenhouse gas, greenhouse gases are a type of insulation, and insulation prevents heat radiation. Artificial intelligence techniques for qualitative reasoning (Biswas, Leelawong, Schwartz, Vye, & TAG-V, 2005) enable the agent to use the concept map to answer questions, thereby providing a model of thinking with high interactivity. In Figure 1, Dee uses her map to answer the query, “What happens to ‘heat radiation’ if ‘methane’ increases?” She highlights successive links in her concept map to show the entire chain of inference. To complement the graphical reasoning, Dee also unfolds her reasoning in text (lower panel). An agent always reasons “logically,” even if the propositions it has been taught are incorrect. Students can trace their agent’s reasoning, and then reorganize or amend its knowledge (and their own) if necessary. To further support students in their learning and “teaching,” a variety of affiliated TA technologies have also been developed that provide students with extensive feedback on their agent’s progress (insert Figure 2 here).

We hypothesize that the metaphor of teaching allows students to use a well-known, and potentially productive, schema for organizing their interactions and for interpreting feedback (for reviews of the benefits of learning-by-teaching, see Annis, 1983; Biswas et al., 2005; Renkl, 1995; Roscoe & Chi, 2008). The teaching metaphor can also enlist fruitful social attitudes including a sense of responsibility towards one’s pupil. In a pair of recent studies with TA, students in one condition were told the character was an agent they were teaching, and in the other condition, students thought the character represented themselves. Students who thought they were teaching the agent exhibited more affect, were more attentive to feedback, and spent more time revising their errors by engaging in learning-relevant behaviors. Consequently, they learned more (Chase, Chin, Oppezzo, & Schwartz, 2009).

We have also conducted studies to examine the nature of the learning benefits from TA. The TA’s knowledge organization and visualizations emphasize reasoning through causal chains. We hypothesize that by making the agent’s thinking visible (Collins, Brown, & Holum, 1991), TA can help scaffold students’ own causal reasoning. In a pilot study with 58 sixth-graders learning about global warming, two classes were randomly assigned to use either the TA software ($n = 28$), or the widely-used concept mapping software, Inspiration[®] ($n = 30$). Logistical constraints necessitated the use of intact classes; however, the school matches classes on ability, and a pre-test on global warming showed no significant differences between treatment groups; $p > .25$. Over the course of three weeks, students completed a three-unit course on the mechanisms, causes, and effects of global warming. Both classes completed matched instructional activities that included readings, videos, hands-on experiments, and classroom discussion. After each basic unit, students constructed concept maps with the productivity-focused Inspiration[®] or the interactive, feedback-focused TA system. We assessed the students with three unit tests, each with eight, short-answer, paper-and-pencil questions about causal relations. All tests had questions that required short, medium, and long chains of causal inference.

Figure 3 shows the average score broken out by treatment, lesson unit, and the length of inferential chain needed to answer the question (insert Figure 3 here). After the first unit, the two groups overlap, with the TA students showing a very modest advantage for longer inferences. After the second unit, the TA students show a strong advantage for the medium-length inferences. By the final unit, the TA students show an advantage for short, medium, and long inferences. One interpretation of this pattern is that the TA students were getting progressively better at reasoning about longer and longer chains of inference in the context of global warming.

To further examine if students had internalized ways to integrate and track causal chains, the two groups of students were given an opportunity to learn new content at the end of the study, but without support from the technologies. This is a so-called Preparation for Future Learning (PFL) assessment, because students have an opportunity to learn during the assessment (Bransford & Schwartz, 1999). Students in both conditions made paper-and-pencil concept maps summarizing a new text passage on what they could do to help prevent global warming. Students received four starter nodes to get them started. Students in both conditions added approximately four concepts, with no differences between conditions; $p > .4$. The TA students, however, showed twice as many well-integrated nodes (2.5) compared to the Inspiration[®] students; $p < .01$.

Purpose / objective / research question / focus of study:

If asked, many parents and educators would agree that incorporating technology into the curriculum is a good idea for our schools. However, there are concerns that computer technologies may fail to bring “added-value” to student learning, or worse, they may displace curricula that once provided “basic-value” learning (Clarke & Dede, 2009). Another critique is that technologies may over-scaffold student learning, leaving students overly dependent on their technological scaffolds such that they cannot perform basic procedures on their own. Consider, for instance, the debates over whether students should be allowed to use hand-held calculators in school (Ellington, 2003), or whether word-processing programs and spell-checkers have degraded our nation’s writing skills (Galletta, Durcikova, Everard, & Jones, 2005).

John Dewey (1916) stated, “the aim of education is to enable individuals to continue their education . . . the object and reward of learning is continued capacity for growth” (p. 117). Given this tenet of constructivism, the gold standard for a good instructional technology is one that will not only help students learn the content-at-hand, which is an important outcome, but will also prepare students for future learning (Bransford & Schwartz, 1999).

Our studies on the effectiveness on TA have been, in the past, of relatively short duration, used specially designed content, and were taught under the strict edicts of the research designs. We wanted to see how the technology would fare in the more complex ecology of regular instruction, when teachers could integrate TA as they chose, into the flow of their normal curriculum over a sustained period of time. Of particular interest were 1) whether we could replicate our earlier PFL results showing that learning benefits persisted for students, *even when no longer supported by the technology*, 2) whether gains would emerge on the standard, basic-value assessments, in addition to our own added-value assessments that focused on the TA’s strength of promoting causal reasoning, and 3) whether the learning benefits would be associated with particular student behaviors in the TA system.

Setting:

A small, local school district agreed to use the TA technology as added-value instruction

to complement their regular science curriculum. The district had adopted 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.

Population / Participants / Subjects:

The study involved six teachers and 134 5th-grade students (104 with permission to analyze their data). The six teachers and their classes had been split into two teaching teams by the school for scheduling purposes. An analysis of students' math and reading STAR scores (Standardized Testing and Reporting) from the previous year indicated no pre-existing achievement differences between the teams; $p > .98$. (STAR does not include a science component in the 4th-grade). We also administered the pretest for the first FOSS unit and found no pre-existing differences; $p > .68$.

Intervention / Program / Practice:

Teachers were trained on TA during a full-day, in-service workshop. They were then allowed to integrate the technology as they wanted into their own lesson plans. Teachers tended to implement the TA software differently with their students, e.g. they used TA at different points in their lesson plans, or preferred one feedback tool over another. We provided each teacher with as much technical and curricular support as she wanted.

The timing of state testing plus end-of-year school events yielded different durations for the two FOSS kits, the biology-focused Living Systems (LS) and the earth-science-focused Water Planet (WP). The teachers had approximately 10 weeks for LS, and about 5 weeks for WP. Four expert maps were used for the LS kit (totaling 41 nodes with 42 links), and two maps for the WP kit (21 nodes with 21 links). Overall, teachers averaged two TA sessions per map. These differences had implications for the amount of data we could collect for each unit, as described next.

Research Design:

The study was designed as a simple cross-over. In the winter, one teaching team integrated TA with the first kit, LS. We will refer to these three classes as Cohort 1. The other team served as the control, using the FOSS materials as they normally would. We will refer to this second set of classes as Cohort 2. In the spring, the cohorts crossed over. Cohort 2 used TA for the WP kit, while the Cohort 1 teachers taught WP without TA. This permitted us to examine whether the TA benefits, if any, would continue forward when Cohort 1 students stopped using the technology for the WP kit.

Data Collection and Analysis:

For the LS kit, the teachers covered three sub-units: Human Body, Vascular Plants, and Photosynthesis & Cellular Respiration. For the WP kit the teachers covered the extensive Water Vapor sub-unit. The FOSS kits come with summative assessments for each sub-unit, called I-Checks. They contain multiple-choice, fill-in, and short-answer questions. We sorted the FOSS items into 4 content categories based on their "prompt" word: *Why* questions asked about causal inferences; *How* questions probed internal mechanisms; *What* questions tested declarative factual recall; and, *Data* questions asked students to interpret charts or tables. These items served as the measure of "basic-value" to determine whether TA displaced or augmented the intended goals of

the original curriculum. To each of the four I-Checks, we appended four “added-value” assessment items that tapped the types of causal reasoning modeled by TA.

All learning assessment items were scored on a scale of 0 to 1. Answers received 0 points (incorrect or no answer), ½ point (partially correct answer), or 1 point (correct answer). Inter-coder reliability, using a random subset of at least 20% of the answers for each item, had correlations of greater than .92 for all tests. In addition, the student scores from the study indicated that the reliability of the 16 added-value items ($\alpha = 0.83$) matched the reliability of the 47 FOSS basic-value items ($\alpha = 0.86$).

We also collected complete log data profiles to see when and how each student used the system and to examine whether working with TA more, or in specific ways, correlated with learning outcomes.

Findings / Results:

The added-value results for the first kit, LS, are shown on the left side of Figure 4. Cohort 1 (using TA) significantly outperformed Cohort 2 (not using TA); $F_{(1,101)} = 5.2, p < .05$. Moreover, the class mean for each of the three Cohort 1 teachers was higher than the class mean for each of the three Cohort 2 teachers. Thus, TA provided a significant added-value for the 5th-graders (insert Figure 4 here).

The next question addressed what changes would appear when the cohorts crossed-over in the use of TA for the WP kit. A repeated-measures analysis compared learning for the two kits crossed by condition. The interaction was significant; $F_{(1,96)} = 4.7, p < .05$. The interaction addresses two questions. The first was whether the learning of the Cohort 2 students would improve on the added-value questions once they used the software. Figure 4 shows that they did. The second question was whether the Cohort 1 students would continue to perform at the same level once they stopped using TA. Figure 4 shows they did.

There was a concern that the TA lessons might detract from the basic-value of the FOSS kits. To examine this issue, we analyzed the I-Check results from the LS unit (several of the teachers did not give the I-Check for the WP unit, though they did give our added-value questions as requested and analyzed above).

Figure 5 shows that the cohorts did similarly on three question types, but Cohort 1 students did better on the *Why* questions. A repeated-measures analysis compared question types by cohort. The interaction driven by the *Why* questions was significant; $F_{(3, 84)} = 7.0, p < .001$. Thus, the TA system did not reduce students’ learning of basic FOSS material, and improved it for the *Why* questions. The benefit for the *Why* questions fits our general story about TA, because these questions asked students about cause and effect relationships (insert Figure 5 here).

The preceding analyses compared two treatments experimentally to determine the effectiveness of TA. A complementary approach is to look at effects within the TA treatment. If the technology is responsible for improved learning, then we should expect to see “dosing effects” – students who more frequently use productive elements of the software should learn more. The following analyses are exploratory, because the effects are only correlations, and because we were not sure which aspects of the TA system were especially useful for learning if used more frequently.

System-use metrics (e.g., number of mapping sessions, map edits, resource accesses, agent queries, quizzes, etc.) were used to predict outcome performance on the added-value measures (LS for Cohort 1, and WP for Cohort 2). We forced the STAR scores into the

regression equation to control for prior achievement, and then conducted a stepwise regression with the metrics (the sample size reduces because not all students had STAR data). For Cohort 1, the stepwise regression found that number of map edits was most predictive; $F_{(3, 37)} = 10.3$, $p = .001$, $R^2 = .45$. For Cohort 2, number of quizzes proved to be most predictive; $F_{(3, 38)} = 6.1$, $p = .002$, $R^2 = .33$). Although the specific, predictive actions differed between the cohorts, quizzing and editing are highly correlated ($r = .70$) and indicative of productive behaviors in the system.

Conclusions:

Teachable Agents weathered its first test in the complexity of the real world, where teachers chose how to use the software for several months as an added-value to their normal instruction. Students exhibited a deeper causal understanding of the FOSS material, as measured by the added-value tests and the *Why* questions in FOSS's own basic-value assessments. The TA activities did not displace basic learning from the FOSS kit. Moreover, the degree to which students used the map editing and feedback features correlated with learning, even after controlling for prior achievement. And finally, perhaps the most exciting result is that experience with TAs supported students' future learning of new content, even when they no longer used the software.

In conclusion, there has been a good deal of recent discussion about the importance of 21st-century skills and competencies (Banta, 2009; Rotherham, 2008; Silva, 2008). The assumption of these discussions appears to be that times have changed, and they will continue to do so. The latter assumption – that times will continue to change – suggests that no set of static skills or competencies will do; people will need to continue to learn. Therefore, it seems worthwhile to prepare students to continue learning so they can adapt (and contribute) to dynamic times where new technologies come and go. In the current study, we demonstrated that one approach is to provide students with ways of thinking about specific content. We did not teach children how to learn in general, for example, by taking notes or explicitly self-explaining. Instead, we provided them with the powerful, integrative, and visible idea of causal chains in science. We further added interactive feedback, so they could model and refine this type of reasoning. Interactive technologies can provide added-value to many current curricula by targeting critical organizing principles with a range of feedback and visualizations, and in the process, indirectly increase the value of future curricula.

Appendices

Appendix A. References

- Annis, L. (1983). The processes and effects of peer tutoring. *Human Learning*, 2, 39-47.
- Banta, M. (2009, February 24). The value of teaching 21-st century skills. *The Boston Globe*. Retrieved from <http://www.boston.com/bostonglobe/>
- Biswas, G., Leelawong, K., Schwartz, D. L., Vye, N., & TAG-V (2005). Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence*, 19, 363-392.
- Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education*, 24, (pp. 61-101). Washington DC: American Educational Research Association.
- Chase, C., Chin, D. B., Opezzo, M., & 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.
- Clarke, J., & Dede, C. (2009). Robust designs for scalability. In L. Moller, J. B. Huett, & D. M. Harvey (Eds.), *Learning and instructional technologies for the 21st century: Visions of the future*, (pp. 27-48). New York: Springer.
- Collins, A., Brown, J. S., & Holum, A. (1991). Cognitive apprenticeship: Making thinking visible. *American Educator*, 15(3), 6-11, 38-46.
- Dewey, J. D. (1916). *Democracy and education*. New York: Macmillan.
- Ellington, A. J. (2003). A meta-analysis of the effects of calculators on students' achievement and attitude levels in precollege mathematics classes. *Journal for Research in Mathematics Education*, 34(5), 433-463.
- Galletta, D.F., Durcikova, A., Everard, A., and Jones, B. (2005) "Does Spell-Checking Software Need a Warning Label?" *Communications of the ACM*, 48(7), 82-85.
- Renkl, A. (1995). Learning for later teaching: An exploration of mediational links between teaching expectancy and learning results. *Learning and Instruction*, 5, 21-36.
- Roscoe, R. D. & Chi, M. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science*, 36, 321-350.
- Rotherham, A. J. (2008, December 15). 21st-century skills are not a new education trend but could be a fad. *U. S. News & World Report*. Retrieved from <http://www.usnews.com>
- Silva, E. (2008). Measuring skills for the 21st century. *Education Sector Reports*. Retrieved from http://www.educationsector.org/usr_doc/MeasuringSkills.pdf

Appendix B. Tables and Figures

The screenshot displays the 'Dee's Brain (Global Warming 2)' interface. On the left is a toolbar with options: Pointer, Teach Concept, Teach Link, Switch Direction, Edit, Delete, and Erase Colors. The main workspace contains a concept map with nodes like Sun, light, absorbed light, heat, average Earth temperature, reflective surface, snow/ice, heat radiation, insulation, greenhouse gases, water vapor, carbon dioxide, methane, deforestation, burning fossil fuels, factories, cars, electricity, and garbage. A dialog box asks: 'What is the question? If methane increase, what happens to heat radiation?'. The bottom panel shows a chat log with the agent 'Dee' and a custom avatar.

Dee (13:40:23) : You asked me: If methane increases, what happens to heat radiation?
 I think that if methane increases, heat radiation decreases. This is the path I followed to get my answer. If you want me to explain my answer, click Explain.

Dee (13:40:38) : I'll explain how methane affects heat radiation. An increase in methane causes greenhouse gases to increase, which causes insulation to increase, which causes heat radiation to decrease. Click Explain to hear more.

Figure 1. The Teachable Agent Interface. The student has a) named her agent “Dee,” b) customized Dee’s look, c) taught Dee about global warming, and d) asked her, “What happens to ‘heat radiation’ if ‘methane’ increases?”

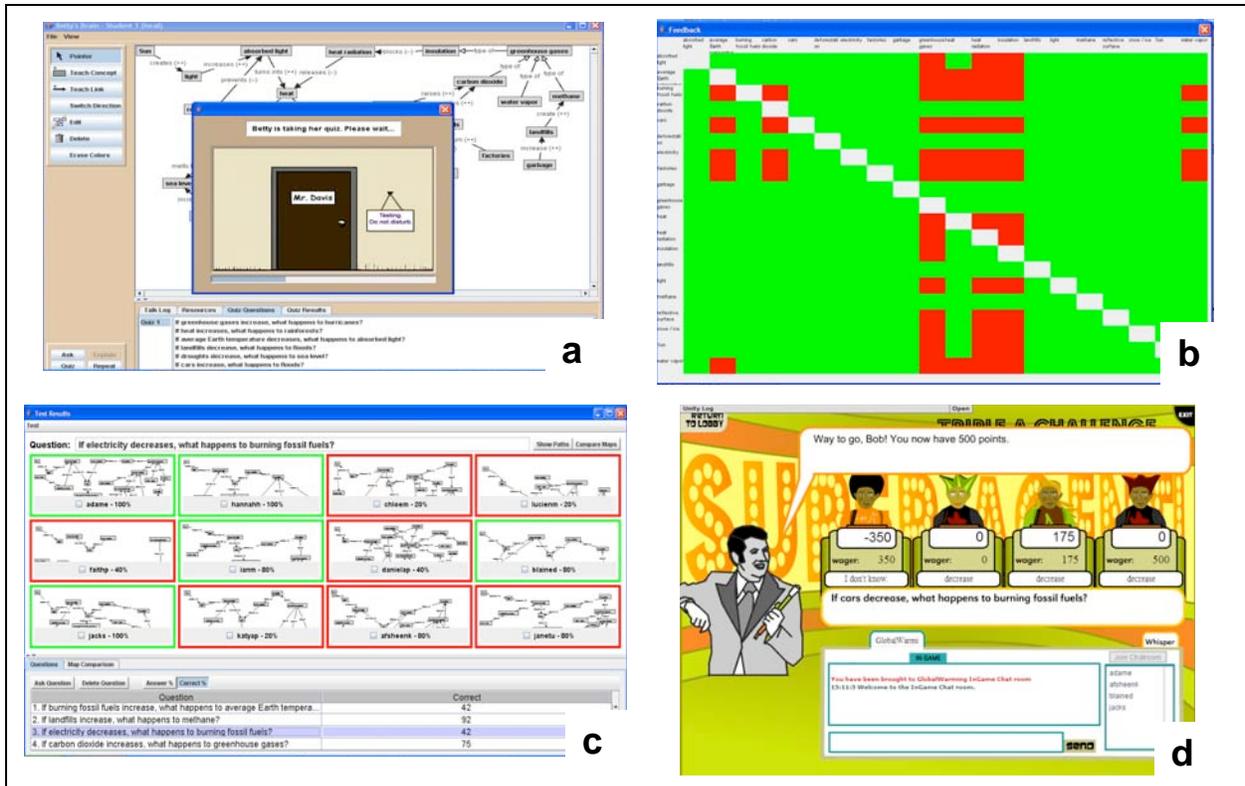


Figure 2. TA-Affiliated Feedback Technologies. a) Quiz Feature: students can have their agents take a quiz to test their knowledge and determine if revision is needed. b) All-Possible-Questions (APQ) matrix: automated scoring indicates TA accuracy for all possible questions [Green = correct; Red = incorrect; Yellow = correct but reasoning path is wrong]. c) Front-of-Class (FOC) display: teachers can project and quiz multiple agents simultaneously to provide a visual anchor for classroom discussion. d) Triple-A Game Show: students can chat and have their agents compete in an on-line game for homework.

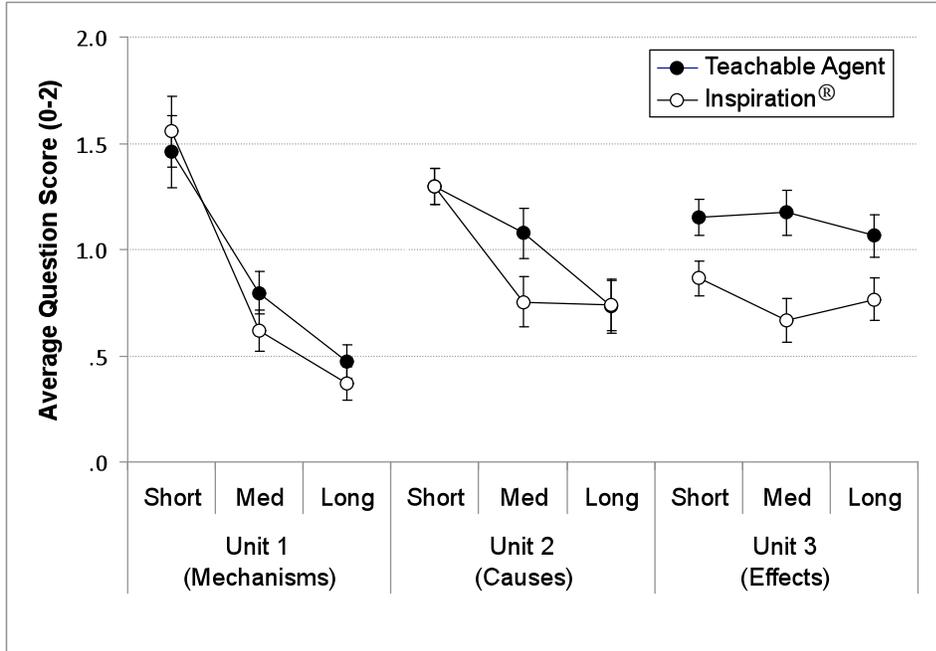


Figure 3. Average Item Scores for Global Warming Assessments. Scores are broken out by unit test, inference length, and treatment. One interpretation of this pattern is that the TA students were getting progressively better at reasoning about longer and longer chains of inference, relative to the Inspiration® students.

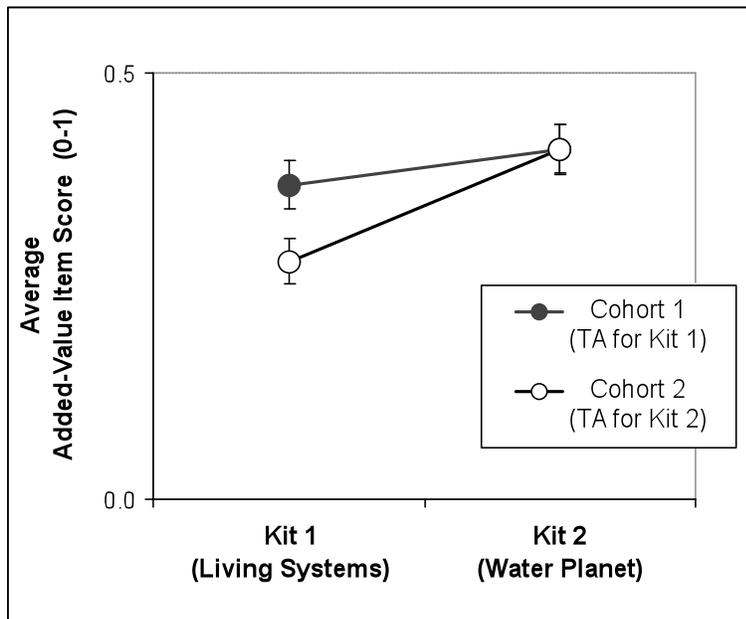


Figure 4. Average Scores on Added-Value Items for FOSS Kits. Scores are broken out by kit and treatment. Kit 1 results indicate that the TA system provided a significant added-value benefit for Cohort 1 students. Kit 2 results seem to show that Cohort 2 students improved once they used the software, and Cohort 1 students continued to perform at the same level once they stopped using the technology.

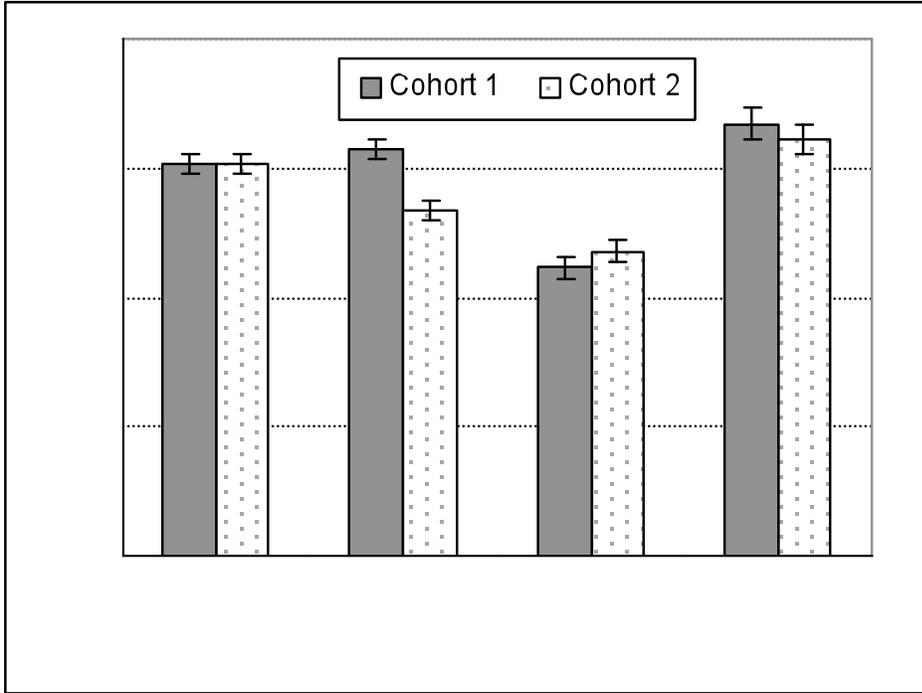


Figure 5. Average Scores on Basic-Value Item Types. Scores are for LS unit only and broken out by item types and treatment. Data indicate that the TA system did not reduce Cohort 1 students' learning of the basic FOSS material, and improved it for the *Why* questions relative to Cohort 2 students.