Education Tech Research Dev (2013) 61:71–89
DOI 10.1007/s11423-012-9275-0

RESEARCH ARTICLE

The effect of contextualized conversational feedback in a complex open-ended learning environment

James R. Segedy · John S. Kinnebrew · Gautam Biswas

Published online: 6 October 2012

Abstract Betty's Brain is an open-ended learning environment in which students learn about science topics by teaching a virtual agent named Betty through the construction of a visual causal map that represents the relevant science phenomena. The task is complex, and success requires the use of metacognitive strategies that support knowledge acquisition, causal map construction, and progress monitoring. Previous research has established that middle school students struggle at such tasks without proper scaffolding and feedback. In Betty's Brain, this feedback is provided by Betty and Mr. Davis, another virtual agent designed to provide guidance and suggestions as students work. This paper discusses our implementation of contextualized conversational (CC) feedback, and then presents the results of an experimental study exploring the effects of this feedback in two 8th-grade science classrooms. The results illustrate some advantages of the CC feedback in comparison with a baseline dialogue mechanism that presents similar strategies in a non-conversational, non-contextualized form. While both groups showed significant pre-to-post test learning gains, the difference in learning gains between the groups was not statistically significant. However, students who received CC feedback more often performed actions in accordance with the advised strategies, and they created higher quality causal maps.

Keywords Conversational agents · Open-ended learning environment · Metacognition · Student learning behaviors · Mixed-initiative dialogue

---

Electronic supplementary material The online version of this article (doi:10.1007/s11423-012-9275-0) contains supplementary material, which is available to authorized users.

J. R. Segedy (✉) · J. S. Kinnebrew · G. Biswas
Institute for Software Integrated Systems, Vanderbilt University, PMB351829, Nashville, TN 37235, USA
e-mail: james.segedy@vanderbilt.edu

J. S. Kinnebrew
e-mail: john.s.kinnebrew@vanderbilt.edu

G. Biswas
e-mail: gautam.biswas@vanderbilt.edu

Introduction

Betty's Brain (Leelawong and Biswas 2008) is a computer-based learning environment that supports middle school students in learning science. In the system, students teach Betty, a virtual agent, by constructing a visual causal map that represents the relevant science phenomena. As they teach, students can probe Betty's understanding by asking her to answer questions. Her answers, generated by systematically reasoning through a chain of links in the causal model, provide opportunities for students to gain an understanding of how entities and processes in the science domain interact with one another. Consequently, the student's teaching task is targeted toward building and reasoning with causal models, an essential ability underlying much of scientific thinking and reasoning (Jonassen and Ionas 2008). For example, explaining the notions of interdependence and balance in ecological processes relies heavily on the ability to think and reason causally (Biswas et al. 2005).

The student's learning and teaching task in Betty's Brain is *open-ended*, and it requires students to regulate their learning processes. Middle school students need to choose between and sequence their learning activities as they carefully study science phenomena, express them in the causal representation, and monitor their progress toward completing their teaching task. However, several researchers have shown that many students fail to gain an understanding of target domain knowledge in open-ended learning environments (OELEs) (*e.g.*, Land 2000; Mayer 2004). These students may fail to ask the right questions or think meaningfully about how to achieve their learning goals, instead adopting sub-optimal learning behaviors to overcome their lack of understanding and insight (Azevedo 2005; Segedy et al. 2011). To support students through these difficult tasks, the system employs agent-delivered feedback to guide students toward using metacognitive strategies important for success in teaching Betty (Biswas et al. 2010). This support promotes effective learning behaviors and focuses on developing students' abilities to independently regulate their learning processes in preparation for future learning (Bransford and Schwartz 1999).

Feedback is generally defined as information provided to learners in response to their learning decisions (Shute 2008; Hattie and Timperley 2007). It seeks to highlight differences between desired and current learner performance, and when designed well, feedback can motivate learners to improve their approach to accomplishing learning tasks. Much of the research on computer-based feedback design focuses on feedback for tutored, step-by-step problem-solving in domains such as algebra, geometry, and computer programming. This feedback is mainly organized as successive hints that eventually provide the answer to the current problem step (*e.g.*, Koedinger and Aleven 2007; Mendicino et al. 2009; VanLehn 2006). There is little research on feedback design principles for OELEs, which lack the structure of tutored problem-solving. To properly support students, feedback in OELEs must interpret the students' current plan and evaluate its effectiveness.

This paper proposes two guidelines for the design of feedback in OELEs to promote effective learning: feedback should be *contextualized* by the student's task goal (*e.g.*, completing the teaching task), learning artifacts (*e.g.*, the current state of their causal map) and recent activities, and it should be delivered in a mixed-initiative *conversational* format. Students working in OELEs often misunderstand feedback or choose not to meaningfully engage with it (Land 2000). Grounding the feedback in the explicit context of the student's goal, the causal map, and the student's recent activities provides a concrete referent on which to base the feedback. Additionally, delivering the feedback in a mixed initiative conversational manner engages students in a more authentic social interaction with the agents; these conversations invite students to become active participants in their learning by allowing them to influence the direction and depth of the conversation. This control allows students to focus the discussion on topics and information they feel is more helpful or more relevant to their goals.

To explore the effect of contextualized conversational (CC) feedback, this paper presents an experimental study comparing CC feedback to a baseline non-conversational feedback approach in Betty's Brain. We report results on student learning gains from pre- to post-tests and the quality of the causal maps students created during the intervention. To gain further insight into the effects of the enhanced feedback, we also apply data mining methods (Biswas et al. 2010; Kinnebrew et al. in press) to derive student learning behaviors from their activities in the system and compare those behaviors between groups of students. This methodology combines hidden Markov models (HMMs) and a differential sequence mining method to develop more refined interpretations of the students' learning behaviors. The results of this analysis illustrate important differences in learning behaviors between the two groups of students.

Betty's Brain

The Betty's Brain learning environment, shown in Fig. 1, tasks students to teach a virtual agent, named Betty, about science topics by constructing a causal map that represents relevant science phenomena as a set of entities connected by directed links which represent causal relations. Once taught, Betty can use the map to answer causal questions and explain those answers by reasoning through chains of links (Leelawong and Biswas 2008). The

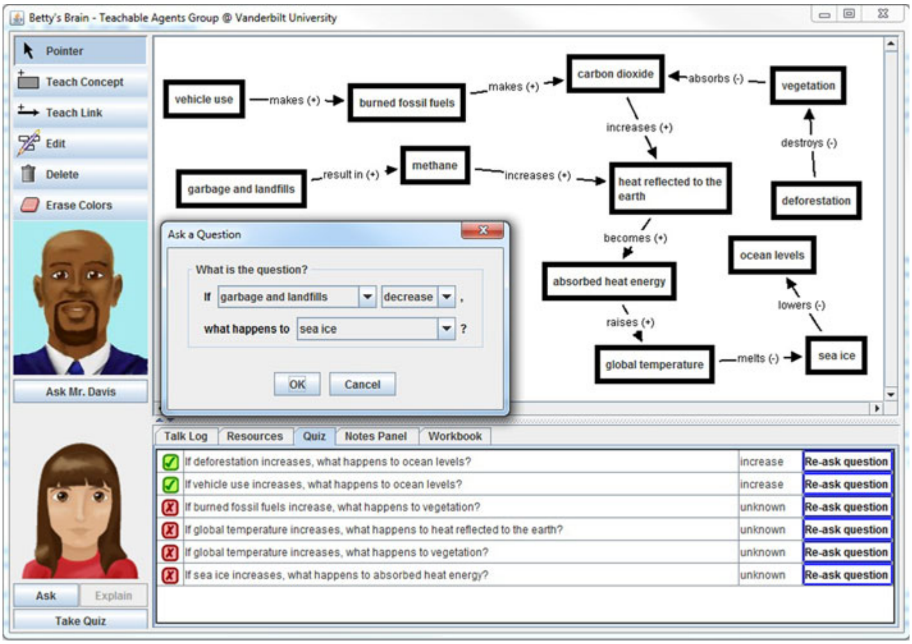

Fig. 1 Betty's Brain system with query window

goal for students using Betty's Brain is to teach Betty a causal map that matches a hidden, expert model of the domain (which is also represented as a causal map).

The students' learning and teaching tasks are organized around five activities: reading, editing the causal map, querying Betty, listening to Betty's explanation, and asking Betty to take a quiz. Students can learn the domain material they need to teach Betty by *reading* the resources, which include both high-level descriptions of scientific processes (*e.g.*, the greenhouse effect) and information pertaining to each concept that appears in the expert map (*e.g.*, solar energy and carbon dioxide). As students read, they need to identify causal relationships, such as "*deforestation* removes *vegetation* from an area," and then explicitly teach the information to Betty by adding the two entities to the causal map and creating the causal link between them (deforestation *decreases* vegetation). In Betty's Brain, link definitions are limited to the qualitative options of "increase" or "decrease." Students can also add textual descriptions to each link. For example, the link in Fig. 1 from deforestation to vegetation is annotated with the word "destroys."

Students can explore Betty's knowledge by *querying* her using the pop-up window displayed in Fig. 1 (*e.g.*, *if garbage and landfills decrease, what effect does it have on polar sea ice?*). To answer questions, Betty uses qualitative reasoning methods that operate through chains of links from the source concept to the target concept (Forbus 1984; Leelawong and Biswas 2008). The learner can further probe Betty's understanding by asking her to *explain* her answer. Betty illustrates her reasoning through text, speech, and animation; she simultaneously explains her thinking (*e.g.*, *the question said that car emissions increase. This causes carbon dioxide to increase. The increase in carbon dioxide causes* …) and animates her explanation by highlighting concepts and links on the map as she mentions them. By asking questions and getting explanations, learners can reflect on their current understanding of the science material and gain a deeper understanding of the processes under study.

Students can check Betty's progress by having her take a *quiz*. Quiz questions are selected dynamically by Mr. Davis using an algorithm that compares Betty's current causal map to the expert map. If Betty's answer and explanation match the expert model (i.e., in answering the question, both maps would utilize the same causal relations between entities), then Betty's answer is marked as being correct. Note that a link's textual description is not considered during this comparison; the algorithm only focuses on the effect of the link (increase or decrease). Since the quiz is designed to reflect the current state of the student's map, a set of questions is chosen (in proportion to the completeness of the map) for which Betty will generate correct answers. The rest of the quiz questions produce incorrect answers, and they are chosen to direct the student's attention to parts of the map with missing or incorrect links. Therefore, the quiz serves as an important source of performance and progress feedback to the learners. When Betty is unable to answer quiz questions correctly, the students can use that information to discover Betty's misunderstandings and correct them by adding to or modifying the causal map. An example quiz from a lesson on climate change is included in Fig. 1. Each row of the quiz contains the quiz question, the grade, Betty's answer, and a button that allows the learner to ask Betty for her current answer to the question.

Contextualized conversational feedback

OELEs like Betty's Brain present significant challenges to novice learners. To successfully complete their learning tasks, students must develop systematic approaches to learning, constructing the causal model, and monitoring their progress toward completion. Overall, this requires the use of several metacognitive strategies. For example, students must choose how to decompose the resources in order to plan their studying. They also need to read and understand the resources in order to first convert relevant material into causal relations between entities and then teach these relations to Betty. As they teach, they must continually monitor Betty's understanding by asking her to answer questions, explain her answers, and take quizzes. These assessments allow students to gauge their progress and identify deficiencies in their causal models. Additionally, they can serve as a focal point for guiding subsequent reading and teaching activities.

However, research on OELEs and other constructivist approaches to learning has shown that many students make ineffective, suboptimal learning choices when they work by themselves in the absence of adequate scaffolding (Mayer 2004; Land 2000; Roll et al. 2011). Further, prior experience with Betty's Brain shows that these students often misinterpret or ignore information provided by the learning environment, thus missing important opportunities to engage in reflective thinking (Segedy et al. 2011). To help students overcome these challenges, Betty and Mr. Davis provide feedback to guide students toward using metacognitive strategies as they plan their studying, construct their causal model, and monitor Betty's progress.

As the student's tutee, Betty provides feedback that is inquisitive rather than explicitly instructive. She focuses mainly on the interactions between herself and the student, pointing out inconsistent or ineffective behaviors and making corrective suggestions. This leads to feedback centered mainly on editing, querying, and requesting explanations. For example, if a student has not been asking Betty to explain many answers, where "many" is defined on a per-experiment basis, Betty might say "Could you listen to my explanations and make sure that they match what the resources say?"

As the student's mentor, Mr. Davis provides feedback related to both the task and the science domain content. He explains aspects of the teaching task by providing advice such as: *Betty's answer to a question won't change unless you make changes to her causal map*. He grades Betty's quizzes, providing both correctness information and clues about the reasons for her incorrect answers (*e.g.*, *Betty's answer is wrong because she is missing some important links*). He also suggests specific activities linked to effective learning strategies (*e.g., When Betty gets a quiz question wrong, compare the links she uses in her explanation to the text in the resources. You may find that one of the links is not correct*).

Feedback in Betty's Brain is *contextualized* by the student's task goal (teach Betty the correct map), the current causal map, and the student's recent activities on the system; it explicitly references specific concepts, links, quizzes, and questions that are related to the student's recent activities (*e.g.*, adding a link or asking Betty to take a quiz), and it provides explicit information about how these activities contribute to the student completing their teaching task. Additionally, the feedback is delivered through *conversations*: mixed-initiative, back-and-forth dialogues between the student and the agent implemented as conversation trees (Adams 2010). The nodes of a conversation tree represent a computer character's dialogue and the branches represent conversational choices available to the user. Such a structure captures the possible directions that a single conversation might take once it has been initiated. Thus, students can control the depth and direction of the conversation within the space of possible conversations provided by the dialogue and response choices.

Figure 2 shows an excerpt from a conversation tree in Betty's Brain. This conversation is initiated by Betty after she answers a question, and its design encourages students to engage with her reasoning process. When Betty answers a question from the student, she asks if her answer makes sense. After reflecting on Betty's answer, the student may suggest

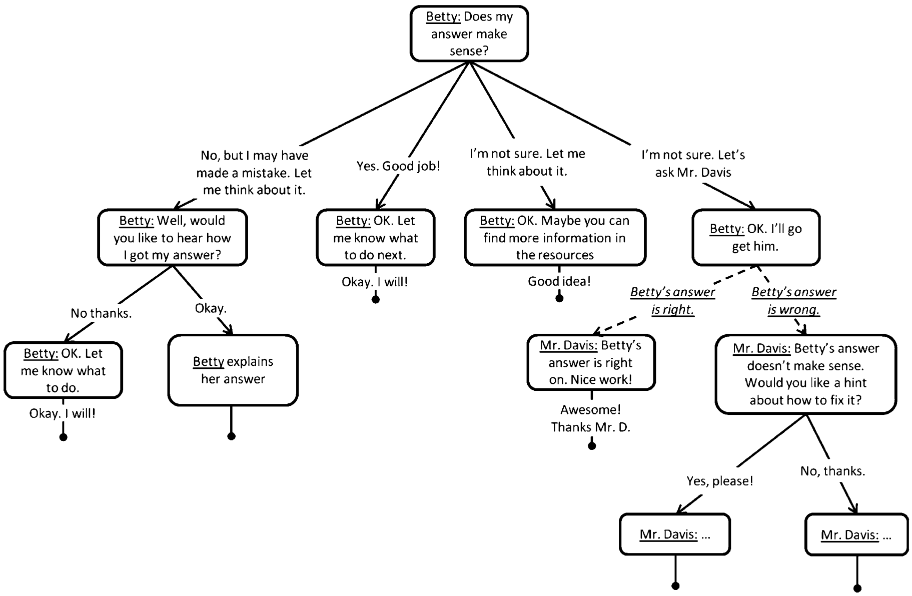

Fig. 2 Partial conversation tree from Betty's Brain

asking Mr. Davis by choosing the response "I'm not sure. Let's ask Mr. Davis." Betty will then agree, saying "OK. I'll go get him," and this is followed by Mr. Davis providing feedback on Betty's answer. The conversation continues in this back-and-forth manner until it reaches a leaf node, which indicates the end of the conversation.

This example illustrates an important advantage of using conversations: agents and learners can together negotiate goals and plans. This effectively enriches agent-learner interactions by incorporating a mix of both dialogue and action. Additionally, dialogue options can be designed such that a student's choices are diagnostic of their current understanding, priming the agents to deliver more targeted feedback in future conversations (Segedy et al. 2012).

Method

The present experimental study tests the effect of CC feedback in Betty's Brain by comparing it to a baseline feedback approach called prompt-based action-oriented (PA) feedback. PA feedback is characterized by two main attributes. First, it is organized into *prompts*: short statements delivered as one-way communication. After an agent speaks, the learner has no opportunity to respond to the agent. Second, PA feedback is *action-oriented*; when students take an action in the system, agents delivering PA-feedback suggest potentially useful behaviors and strategies that are linked to that action. For example, every three-to-five times the student asks Betty a question, Mr. Davis encourages students to "ask her to explain her reasoning … to see if her explanations make sense." Note that this feedback statement does not reference specific causal links; nor does it reference the question that the student just asked or provide information about how requesting Betty's explanation relates to the student completing their teaching task. An equivalent CC statement would ground the same advice in the context of the student's goals and the causal map: "Betty's answer to the question you just asked her, 'if car emissions increase, what happens to vegetation?' is not right. You should have Betty explain her answer to this question and try to find her mistake."

Tables 1 and 2 display several examples of PA and CC feedback statements and the student behaviors that triggered them. Both sets of feedback encourage students to integrate their reading, map-building, and assessment activities; students are told to read the

Table 1 PA feedback statements and their triggers

Targeted skill	Triggering student behavior	Agent	Feedback
Reading	Adding 5–7 concepts	Mr. Davis	When you read about the concepts in the resources, try to explain the information in your own words before teaching Betty.
Reading/monitoring with quizzes	Adding 5–7 links		To help Betty pass her quizzes, make sure that every link is correct. Compare the links you have taught to what is in the resources.
Reading/monitoring with questions/explanations	Making Betty take 3–4 quizzes		When Betty answers a question incorrectly, use the resources to double-check all of the links in her explanation of the answer.
Reading/monitoring with explanations	Getting 3–5 explanations		You should check if Betty's explanations are correct by comparing them to the information in the resources.
Monitoring with questions	Deleting 3–5 concepts and/or links		If you need to remove things from the map, only delete one thing at a time. Ask questions to see how each change affects Betty's answers.
Monitoring with explanations	Asking 3–5 questions		After Betty answers your questions, ask her to explain her reasoning. Follow each step carefully to see if her explanations make sense.
Monitoring with quizzes	Making Betty take 4 quizzes within 4 min		Betty's scores on the quizzes will not change much if you do not teach more between quizzes.
	Asking 2–4 questions that Betty can't answer		Betty cannot answer a question if she cannot follow links from the first concept in the question to the last concept in the question.
Reading/monitoring with explanations	Asking 3–4 questions without getting explanations	Betty	Can we go over my explanation step by step and check it with the resources?
Monitoring with questions	Adding 4–6 links		Maybe you should ask me some practice questions to make sure I understand.
	Asking a question		Does my answer make sense?
Monitoring with quizzes	Making Betty take 2 quizzes within 4 min		Do you really think I'm ready for a quiz? It has only been a few minutes since the last one.
Monitoring with quizzes/explanations	Making Betty take 2–3 quizzes		These quizzes can be tough. Can we go over my explanations to see how well I understand?
	Adding 4–6 concepts		Could you explain how this concept affects the other concepts?

Table 2 CC feedback statements and their triggers

Targeted skill	Triggering student behavior	Agent	Feedback
Reading	Adding 4–5 incorrect links	Mr. Davis	The link you just added from solar energy to global temperature is not right. You should read about these concepts carefully before you teach Betty a link.
	Asks for help on a question and Betty's explanation has an incorrect link		Betty's answer is wrong because she does not have the right understanding of how vegetation affects global temperatures. You should read about these concepts and re-teach her this information.
	Asks for help on a question and Betty's explanation is missing a link		Betty's answer is wrong because she hasn't learned some important links. You should look through the resources to try to find the links you are missing.
	Accepts "no progress" help		The link on your map that starts at vegetation and goes to solar energy is confusing Betty. You should look through the resources to see if this link should be changed or deleted.
Monitoring with explanations	Asks what to do next		If I were to ask Betty the question *if car emissions increase, what happens to vegetation*, her answer would be wrong. You should have Betty explain her answer to this question and try to find her mistake.
Monitoring with quizzes	Making Betty take 3–4 quizzes		Would you like help fixing one of Betty's quiz answers?
Help seeking	Makes no progress on map in 10 min		You seem to be having trouble. Would you like some help?
Reading	States intent to read	Betty	Remember, Mr. Davis says to look for words like "causes" and "produces." These usually mean that the sentence is describing a relationship.
	Asks to think about Betty's answer		OK. Maybe you can find more information in the resources.
Monitoring with questions	Asking a question		Is that the answer you expected?
Monitoring with explanations	Says Betty's answer is wrong		Well, would you like to hear how I got my answer? (If student selects yes, Betty explains her answer)
Monitoring with questions/ quizzes	Makes Betty take a quiz without asking her one of the questions she got wrong on the last quiz		Can we please hold off on this quiz? I'd really like to do better, but we didn't practice any of the quiz questions that I got wrong last time.
	Says Betty's answer is right		OK. Let me know what I should do next.

resources carefully and think deeply about the information; continually assess Betty's understanding by asking her to answer questions, explain her answers, and take quizzes; and use the results of these assessments to motivate additional reading activities. However, the CC feedback interactions are grounded in the context of the student's causal map and task goals, and they are presented in a mixed-initiative conversational format.

The research hypothesis was that the conversational and contextual nature of the CC feedback helps students better understand the relevance and context of the feedback. Thus, students receiving CC feedback would:

(1) Build causal maps that more closely match the expert map;
(2) Gain a better understanding of the scientific information presented in the resources;
(3) More often take actions in accordance with the feedback, when compared to the PA group. Thus, students receiving CC feedback should better integrate reading, causal map editing, and assessing activities.

Participants

Forty-four eighth-grade students from 2 intact middle Tennessee science classrooms, taught by the same teacher, were divided by classroom into two treatment groups: PA and CC. The two groups differed only by the agent interactions that occurred while they used the system: students in the PA condition received PA feedback, and students in the CC group received CC feedback.

Because use of the system relies on students' ability to independently read and understand the resources, the system is not suited to students with limited English proficiency or cognitive-behavioral problems. Therefore, data from ESL and special education students were not analyzed. Similarly, we excluded the data of students who missed more than two class periods of work on the system. The final sample was 16 PA students and 21 CC students.

Topic unit and text resources

Students used the Betty's Brain system to learn about climate change. The expert map (Fig. 3) contained 15 concepts and 18 links representing three themes: the greenhouse effect (solar energy, absorbed light energy, heat energy, global temperature, greenhouse effect, heat radiated to space), human activity (deforestation, vegetation, car emissions, carbon dioxide), and effect on climate (sea ice, carrying capacity, condensation, water vapor, and precipitation). The resources were organized into one section per theme, and each concept was discussed on its own page. The text was 25 pages (319 sentences and 4,296 words), with a Flesch-Kincaid reading grade level of 8.1.

Learning and performance assessments

To test predictions one and two, we employed two measures to assess students' task performance and learning gains: (i) a calculated score for the accuracy of the causal map that students created while teaching Betty and (ii) gains in pre- to post-test scores. The map score was computed as the number of correct links (the links in the student map that appeared in the expert map) minus the number of incorrect links in the student's final map. The maximum possible map score was 18. The pre- and post-tests included three kinds of questions that were scored separately. Seven multiple-choice definition questions, each

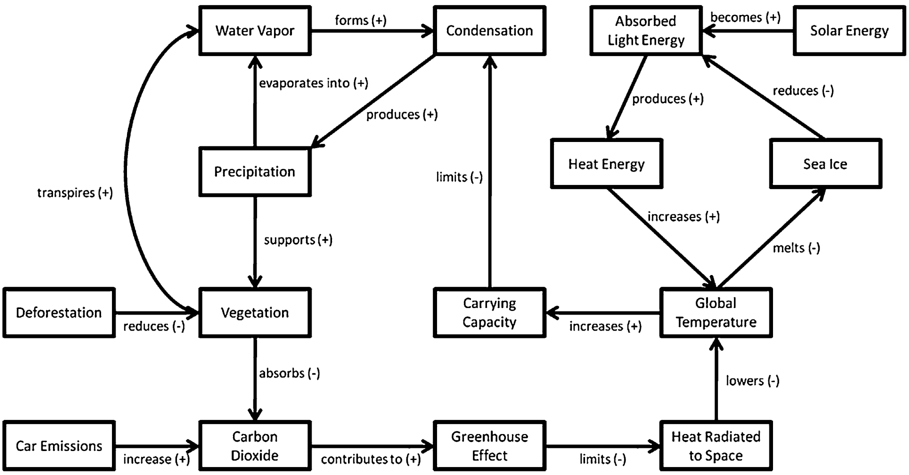

Fig. 3 The global climate change expert map

with four choices, tested students' understanding of primary concepts and processes (*e.g., what is the greenhouse effect*?) and simple relations among concepts (*e.g., which of the following best describes the relationship between sea ice and global temperature*?). Three short-answer questions asked students to employ their understanding of the causes and effects of both climate change phenomena and the greenhouse effect in combination with causal reasoning to explain how certain events might affect global warming (*e.g., How would switching from gas cars to electric cars affect global temperature*?). Finally, 20 causal-reasoning questions presented an abstract causal map (i.e., concepts were named A, B, etc.) and asked students to reason with the map and answer questions such as "if concept A increased what would happen to concept B?" Causal reasoning questions presented students with three possible choices (B would increase, decrease, or not be affected).

Multiple choice questions were scored as 1 (correct) and 0 (incorrect), with a maximum total of 7 for the definition questions and 20 for the causal reasoning questions. Short answer questions were coded by the chain of causal relationships learners used to explain their answers to the questions, which were then compared to the chain of causal links that would be used to derive the answer from the expert map. Note that students did not have access to their causal maps or any other resources when they took the pre- and post-tests. One point was awarded for each causal relationship in the student's answer that came from or was closely related to the expert causal links. The maximum combined score for the three questions was 11. Two coders independently scored a subset of the pre- and post-tests with over 95 % agreement, at which point the coders split the remaining tests and individually coded the answers and computed the scores.

Log file analysis

Prediction 3 was tested by analyzing trace logs of all of the actions that students performed using Betty's Brain, and these logs were automatically converted into a set of sequences of actions, one per student (Kinnebrew et al. in press). For example, if a student asked Betty a question, accessed a page in the resources, and then deleted a link from their causal map, the actions would be coded as QUERY → READ → EDIT. Similarly, if a student took a quiz, asked Betty a question, and then asked her to explain her answer, the actions would be coded as QUIZ → QUERY → EXPLAIN. The generated sequences combined student actions across days such that each student's sequence included all of the actions they took using Betty's Brain over the five-day period.

Once these action sequences were generated, each action was annotated with a measure of *relevance* (Biswas et al. 2010). An action was considered relevant to recent actions if it was related to, or operated on, one of the same map concepts or links. For example, if a student edits a link that Betty recently used to answer a question, the link-editing action is considered "relevant to" the query action. Higher relevance scores for actions suggest that they are more informed by previous actions and, in general, indicate a more systematic or focused approach to the learning task. In this analysis, actions were marked as either having high or low relevance to recent actions, and the relevance was represented with a -L or -H following the action. For example, a highly relevant query would be coded as QUERY-H.

Students' action sequences were then used to derive HMMs (Rabiner 1989), which provide a concise, aggregated representation of student learning behaviors and strategies over the entire time students used the system (Jeong and Biswas 2008). HMMs include a set of states and probabilistic transitions between those states; each state represents a particular set of learning strategies employed by students, and the transitions show how students moved between their use of these strategies. For example, one state may refer to "researching" and another might refer to "problem solving." Further information about the action sequence generation, the structure of HMMs, HMM generation, and HMM interpretation are given in Online Resource 1. Two HMMs were generated, one for each condition, in order to compare the strategies adopted by the two groups of students as they used the system.

Students' action sequences were also analyzed using differential sequence mining (DSM) (Kinnebrew et al. in press). This technique was used to compare two groups of action sequences, one group for each condition, and it automatically discovered subsequences of actions that were more often employed by one group, when compared to the other. For example, if students in one group performed the subsequence QUIZ → QUERY → READ more often than the other group, DSM would identify this difference. More information on differential sequence mining is provided in Online Resource 1.

Procedure

The study proceeded as follows: during the first 45 min class period, students in both treatment groups were introduced to the science topic (climate change) by the classroom teacher. During the next class period (the following day), they completed a pre-test that included questions on both climate change and causal reasoning. During the next two classes, they were introduced to the causal reasoning method used in the system and provided with hands-on system training by the researchers. Students then spent five class periods using their respective versions of Betty's Brain with minimal intervention by the teachers and the researchers. Once students in both groups finished teaching, they took a post-test identical to the pre-test.

Results

On average, there were no differences between students from both conditions in total TCAP scores ($F = 0.815$, $p =$ n.s.) and pre-test scores on definition questions ($F = 0.221$,

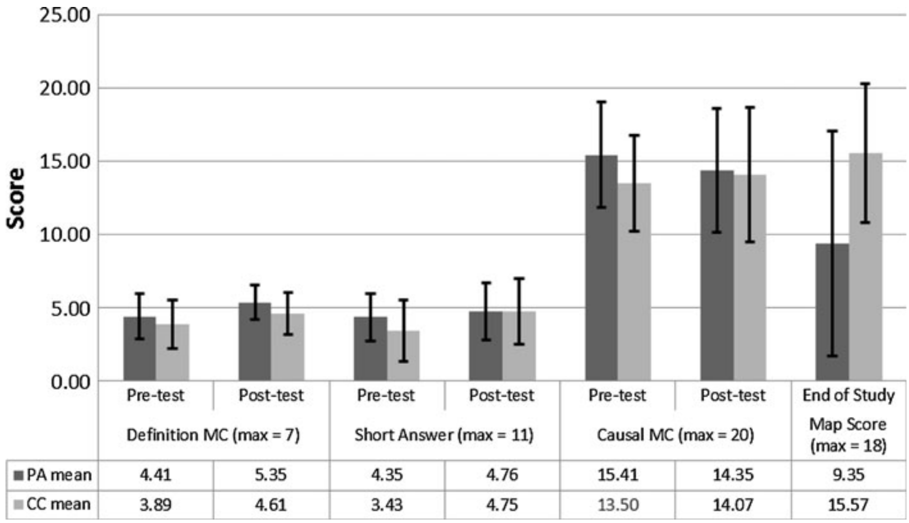

Fig. 4 Means and standard deviations of test scores and map scores

p = n.s.), short answer questions ($F = 0.989$, p = n.s.), and causal questions ($F = 1.714$, p = n.s.). Figure 4 presents means (and standard deviations) of both the pre- and post-test measures (multiple choice, short answer, and causal reasoning questions) and the causal map scores collected in the study.

Repeated-measures ANOVAs were used to analyze each of the three pre-post test scores. For definition question scores, the analysis showed a main effect of time ($F = 11.472$, $p = 0.002$, Cohen's $f^2 = 0.267$), but there was no statistically significant interaction effect of time and treatment ($F = 0.215$, p = n.s., Cohen's $f^2 = 0.005$). For short-answer question scores, the analysis showed a main effect of time ($F = 7.183$, $p = 0.010$, Cohen's $f^2 = 0.167$), but it revealed no statistically significant interaction effect of time and treatment ($F = 1.979$, p = n.s., Cohen's $f^2 = 0.046$). For causal reasoning question scores, the analysis showed no significant main effect of time ($F = 0.304$, p = n.s., Cohen's $f^2 = 0.007$). It revealed only a small trend for an interaction effect of time and treatment ($F = 3.398$, $p = 0.072$, Cohen's $f^2 = 0.079$). An ANOVA conducted on the map scores showed a significant difference between the two groups ($F = 11.386$, $p = 0.002$, Cohen's $f^2 = 0.264$).

Table 3 shows correlations between post-test measures and map scores. Map scores correlated moderately and significantly with causal reasoning scores ($r = 0.36$, $p \leq 0.05$). Causal reasoning scores were also moderately and significantly correlated with definition

Table 3 Correlations of post-test measures and map scores

	Definition MC	Short answer	Causal MC	Map score
Definition MC	–	0.303[a]	0.381[b]	0.148
Short answer	0.303[a]	–	0.533[c]	0.187
Causal MC	0.381[b]	0.533[c]	–	0.357[b]
Map score	0.148	0.187	0.357[b]	–

[a] $p \leq 0.07$; [b] $p \leq 0.05$; [c] $p \leq 0.001$

multiple-choice scores ($r = 0.38$, $p \leq 0.05$), and they were strongly and significantly correlated with short answer scores ($r = 0.53$, $p = 0.001$). Additionally, short answer scores correlated moderately, but not significantly, with definition multiple choice scores ($r = 0.30$, $p \leq 0.07$).

Since the definition and short-answer questions test students on the scientific information in the resources, these results illustrate that the intervention led to statistically significant increases in the students' understanding of that material. However, these results also show that students did not gain additional understanding of the causal reasoning mechanism used in the system. This may be partially explained by the particularly high pretest scores (15.41 and 13.50, out of a possible 20). In considering differences between the two treatment groups, the data present mixed results: while students that received CC feedback created causal maps that better match the expert model (prediction 1), they also suggest that whereas students as a whole showed statistically significant improvements in definition and short answer questions from pre- to post-test, CC students did not gain more than PA students (prediction 2).

To identify differences in learning behaviors between the two conditions (to investigate prediction 3), we analyzed learning activity traces for students in the PA and CC groups by employing the HMM and DSM methods described previously. To ensure that behavioral differences could not be attributed to one group of students receiving more or less feedback from the agents, we counted the number of behavioral suggestions made by the agents (*e.g.*, "you should ask Betty to explain her answer") for each student. An ANOVA conducted on these data showed no main effect for condition on the number of suggestions received by PA students (M = 24.15, SD = 12.65) and CC students (M = 26.95, SD = 11.47), $F = 0.421$, $p =$ n.s.

Details of the HMM generation process and interpretation methods can be found in Online Resource 1. The resulting HMMs, one for each condition, contain five distinct states were interpreted as representing:

- *Reading*: students are primarily reading the resources.
- *Informed editing*: students are primarily making high-relevance edits, suggesting a more focused map-building effort.
- *Uninformed editing*: students are primarily making low-relevance edits, possibly indicating the use of guessing behaviors.
- *Uninformed editing and checking*: students are performing assessment behaviors like querying and quizzing to check the correctness of their causal maps, but are also making a significant number of low-relevance edits to their maps. This implies that students are not reflecting on the results of their assessments, and they may have resorted to trial-and-error methods to correct their maps.
- *Informed editing and checking*: students are making high-relevance changes to their map while also using queries and quizzes to assess the correctness of their causal maps. This state likely corresponds to more effective attempts at employing monitoring strategies to identify and correct erroneous (or missing) information in the map.

The HMM results (Fig. 5) illustrate a similar set of behaviors employed by both the PA and CC groups, although all of the uninformed editing actions in the PA group are combined in the uninformed editing and checking state rather than being split between it and a separate uninformed editing state in the CC group. Additionally, the proportions of expected state occurrences are also relatively similar between the two groups. However, two interesting differences in behavior patterns are worth noting: (i) the PA group's model has a smaller likelihood (39 %) of transitioning directly from informed editing to informed (editing and) checking activities compared to the CC group's model (61 % transition probability); and (ii) the CC group's model exhibits a higher likelihood of following these informed editing and checking activities with more reading (29 vs. 14 % for the PA group's model). This indicates that students in the CC group were more likely to (1) intersperse diagnostic assessment activities between their reading and informed editing activities and (2) return to reading after informed editing and checking. These students may have used monitoring strategies to both identify potential problems in their causal map for further exploration with the resources and also confirm that their current causal map produced correct answers. This analysis provides some evidence for prediction 3, as it suggests that while the CC group did not access the resources more often than the PA group, they may have accessed them as part of more effective strategies.

Table 4 presents the top five differentially frequent subsequences for both conditions, as calculated using DSM. To control for outliers, subsequences were only included if a majority of the students in the frequent group performed the subsequence. In this analysis, series of repeated actions are condensed into one action with the "-MULT" identifier to

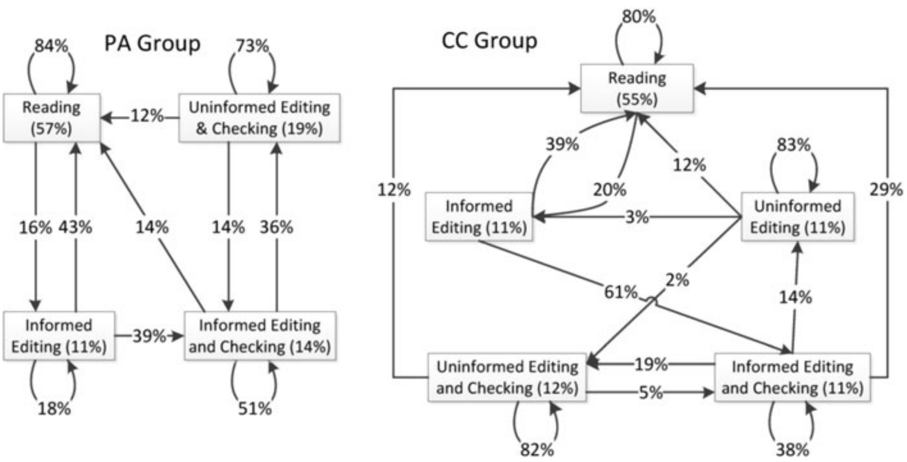

Fig. 5 HMMs for the PA and CC groups

Table 4 Top differentially frequent patterns between CC and PA groups

#	Learning activity subsequence	Frequent group
1.	READ-MULT → EDIT-H-MULT → READ	CC
2.	READ-MULT → READ-MULT → EDIT-H-MULT → READ	
3.	EDIT-H → READ	
4.	READ-MULT → EDIT-H → READ-MULT → EDIT-H → READ	
5.	READ → EDIT-H → READ	
6.	QUERY-L → EDIT-L	PA
7.	EDIT-L → EDIT-L	
8.	QUERY-H → EXPLAIN-H	
9.	QUERY-H → QUERY-H	
10.	EDIT-H → EDIT-L-MULT	

more effectively identify differences in learning behaviors (Kinnebrew et al. in press). For more details on the generation and analysis of the DSM results, please consult Online Resource 1.

This analysis illustrates that the subsequences that were more often performed by the CC group (patterns 1–5) were repeated edit-and-read patterns. Pattern 1, for example, indicates that students in the CC group more often performed a behavior in which they first accessed multiple pages in the resources, then made multiple high-relevance edits to the map, and finally accessed a single page in the resources. In contrast, the subsequences more often performed by the PA group included queries and low relevance edits. For example, pattern 6 indicates that PA students were more likely to ask Betty a low-relevance question before making an unrelated change to their causal map. An important observation emerges from these results: while CC group students more frequently edited their map in conjunction with and most likely informed by recent readings, the PA group students more frequently edited their map without taking advantage of information gained by recent actions (as evidenced by the fact that most of the edits in their subsequences were all low-relevance).

Because the CC group's frequent subsequences are dominated by read and edit actions, Table 4 does not provide insight into how CC group students employed assessment activities as they worked. However, the HMM analysis suggested that there may be important differences in how these two groups monitored their maps via these assessment activities. To better compare the groups' use of these activities, Table 5 presents the top five differentially frequent subsequences that included query, quiz, and explain actions for both conditions.

This analysis illustrates that the CC group was more likely to use queries and quizzes before or after reading (patterns 2–4) and informed editing (patterns 1, 3–5). Further, queries and quizzes were sometimes performed in between sequences of both reads and informed edits (patterns 3, 4). In contrast, the PA group more often used queries and quizzes both in succession (pattern 8) and before and after low-relevance edits (patterns 6, 9). Additionally, the PA group was more likely to ask a high-relevance question and follow it up by asking Betty to explain her answer (pattern 7), and they were more likely to ask a low-relevance question and then perform a high-relevance map edit (pattern 10). These results show that CC group students were more likely to intersperse reading, editing, and

Table 5 Top differentially frequent patterns including assessment actions

#	Learning activity subsequence	Frequent group
1.	READ-MULT → EDIT-H → QUIZ-H → EDIT-H	CC
2.	READ-MULT → EDIT-H → READ-MULT → QUERY-H	
3.	READ-MULT → EDIT-H → QUIZ-H → READ-MULT	
4.	EDIT-H → READ-MULT → QUIZ-H → EDIT-H	
5.	READ-MULT → READ-MULT → EDIT-H → QUERY-H	
6.	QUERY-L → EDIT-L	PA
7.	QUERY-H → EXPLAIN-H	
8.	QUERY-H → QUERY-H	
9.	EDIT-H → QUIZ-L → EDIT-L	
10.	QUERY-L → EDIT-H	

assessing activities; and PA group students were more likely to perform assessment activities in conjunction with one another and in conjunction with edits.

Taken together, Tables 3 and 4 provide further support for prediction 3, suggesting that CC group students were more systematic in their teaching activities: they more often interspersed reading, editing, and assessing activities, indicating that their reading behaviors may have been used to inform their map edits and their assessing behaviors may have been used to inform further reading activities.

Discussion and conclusions

This paper presented a comparative study using the Betty's Brain learning environment in two 8th-grade science classrooms. The experiment tested the effectiveness of feedback that is both *contextualized* by the student's task goal (teaching Betty the correct map), learning artifact (the causal map) and recent actions; and presented in a *mixed-initiative conversational* format. The research hypothesis was that feedback characterized by these two features would help students gain a better understanding of the feedback and consider it more deeply. As a consequence, students would more often take action in accordance with the feedback, build causal maps that more closely match the expert map, and gain a better understanding of the science knowledge. The present study, while it does not provide a definitive prescription for designing feedback in OELEs like Betty's Brain, does offer some insight into the characteristics of effective feedback. Results showed that grounding feedback in the student's learning context (the causal map and recent student actions in Betty's Brain) and organizing the feedback into a mixed-initiative conversational format may lead to effective changes in the learning behaviors of students working in an OELE.

However, there are also limitations to these results. The study did not isolate the effects of contextualization and conversation, which limits the interpretability of the results. Further, the CC students did not improve their pre- to post-test scores significantly as compared to the PA students. Consequently, this research study joins with others that have found a behavioral effect of feedback without finding a clear learning effect (see Shute 2008 for an excellent review of research in feedback). One possible reason for this lack of a clear effect on learning may be related to the fact that the feedback, while being contextualized and conversational, remained focused primarily on metacognitive strategies. For example, it suggested reading a section of the resources but didn't provide support for students who don't understand how to identify causal relationships from reading materials. These students may identify the correct concepts, but fail to understand how to translate the verbal relation into a causal link. As one example, some students did not know the meaning of the word "reduce," and they had to guess whether or not the word implied a causal relationship. Future versions of Betty's Brain must explicitly scaffold such cognitive skills in order to support students' understanding of all aspects of the learning activity.

Another important limitation lies in the aggregated form of the behavioral analyses. Both the HMM and differential sequence mining results represent analyses of student behaviors over all of the days they worked on the system. Thus, they cannot be used to associate specific student behaviors with specific behavioral suggestions delivered by Betty and Mr. Davis; nor do they track how individual student behaviors changed as they progressed through the intervention. However, the fact that students received similar amounts of behavior recommendations suggests that the observed differences can be qualitatively attributed to the differences in the nature of the feedback by the two groups. As we move forward with this work, we will refine our data analysis tools such that they can correlate specific student behavior patterns with the particular kinds of feedback they received. Further, it would be useful to track differences in an individual student's behaviors and performance before and after they received feedback on specific strategies.

Many other computer-based concept mapping tools provide feedback to students as they work. For example, (Lukasenko et al. 2010) present a concept-mapping system that provides link-by-link correctness feedback in terms of: (1) whether or not a link should be placed between the two concepts, (2) the positioning of the concepts in the map, (3) the type of link created, and (4) the direction of the link. When a link is incorrect, students receive a breakdown of what is correct and incorrect. Additionally, the system provides information related to the definitions of the concepts in the link in hopes that students will read that information and make appropriate revisions to their links. Concept Connector (Luckie et al. 2011) provides feedback by grading the connecting words placed on links as correct or incorrect. The COMPASS system (Gouli et al. 2004) provides map correctness feedback by pointing students to incorrect and missing links and asking "initiating questions" such as "Do you really think there should be a link between deforestation and solar energy?"

The feedback in these systems all directly relate to the correctness of the student's concept map. They explicitly differ from Betty's Brain, where the feedback focuses both on map correctness (via the quiz and hints from Mr. Davis) and on supporting students' use of effective knowledge construction and monitoring strategies (by asking students to read the resources carefully and engage with Betty's understanding through questions and explanations). In other words, while these systems focus solely on the *product* of students' learning activities, Betty's Brain focuses on both the *product* of learning and the *process* students employ during learning. The system that most closely resembles Betty's Brain may be COMPASS, which provides correctness feedback in the form of reflective prompts to encourage a deeper engagement with the material. Like Betty's Brain, however, these reflective prompts remain at a metacognitive level, and COMPASS does not provide support for students who lack the pre-requisite understanding of reading and modeling.

Future work with Betty's Brain will expand upon this research through a variety of enhancements to the agent feedback and the data mining techniques. In addition to providing feedback to support students' practice of cognitive skills, future versions of the system will detect and respond to the ineffective learning behaviors identified through the data mining analysis presented in this paper. Along this line, we are currently creating a library of interaction trace segments that are representative of identified learning behaviors and strategies. As more recurring behavior patterns are collected and characterized, the system will have more tools to evaluate and support students using targeted scaffolding and feedback.

Acknowledgments This work has been supported by the National Science Foundation's Information and Intelligent Systems Award #0904387.

References

Adams, E. (2010). *Fundamentals of game design* (2nd ed.). Berkeley, CA: New Riders Pub.

Azevedo, R. (2005). Using hypermedia as a metacognitive tool for enhancing student learning? The role of self-regulated learning. *Educational Psychologist, 40*(4), 199–209. doi:10.1207/s15326985ep4004_2.

Biswas, G., Leelawong, K., Schwartz, D. L., & Vye, N. (2005). Learning by teaching: Anew agent paradigm for educational software. *Applied Artificial Intelligence, 19*(3–4), 363–392. doi:10.1080/08839510590910200.

Biswas, G., Jeong, H., Kinnebrew, J. S., Sulcer, B., & Roscoe, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology Enhanced Learning, 5*(2), 123–152. doi:10.1142/S1793206810000839.

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*, 61–101. doi: 10.3102/0091732X024001061.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence, 24*(1–3), 85–168. doi:10.1016/0004-3702(84)90038-9.

Gouli, E., Gogoulou, A., Papanikolaou, K., & Grigoriadou, M. (2004). Compass: An adaptive web-based concept map assessment tool. In A. J. Cañas, J. D. Novak, F. M. González (Eds.), *Proceedings of the first international conference on concept mapping*, Pamplona, Spain.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. doi:10.3102/003465430298487.

Jeong, H., & Biswas, G. (2008). Mining student behavior models in learning-by-teaching environments. In R. S. J. d. Baker, T. Barnes, & J. E. Beck (Eds.), *Educational data mining 2008: 1st international conference on educational data mining, proceedings* (pp. 127–136). Montreal, Quebec, Canada.

Jonassen, D. H., & Ionas, I. G. (2008). Designing effective supports for causal reasoning. *Educational Technology Research and Development, 56*(3), 287–308. doi:10.1007/s11423-006-9021-6.

Kinnebrew, J. S., Loretz, K. M., & Biswas, G. (in press). A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining*.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review, 19*(3), 239–264. doi:10.1007/s10648-007-9049-0.

Land, S. (2000). Cognitive requirements for learning with open-ended learning environments. *Educational Technology Research and Development, 48*(3), 61–78. doi:10.1007/BF02319858.

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

Luckie, D., Harrison, S. H., & Ebert-May, D. (2011). Model-based reasoning: Using visual tools to reveal student learning. *Advances in Physiology Education, 35*, 59–67. doi:10.1152/advan.00016.2010.

Lukasenko, R., Anohina-Naumeca, A., Vilkelis, M., & Grundspenkis, J. (2010). Feedback in the concept map based intelligent knowledge assessment system. *Scientific Journal of Riga Technical University, 41*, 17–26. doi:10.2478/v10143-010-0020-z.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59*(1), 14–19. doi:10.1037/0003-066X.59.1.14.

Mendicino, M., Razzaq, L., & Hefferman, N. T. (2009). A comparison of traditional homework to computer-supported homework. *Journal of Research on Technology in Education, 41*(3), 331–359.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286. IEEE Computer Society.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Metacognitive practice makes perfect: Improving students' self-assessment skills with an intelligent tutoring system. In G. Biswas, S. Bull. J. Kay, & T. Mitrovic (Eds.), *Proceedings of the 15th international conference on artificial intelligence in education* (pp. 288–295), Berlin: Springer.

Segedy, J. R., Kinnebrew, J. S., & Biswas, G. (2011). Modeling learner's cognitive and metacognitive strategies in an open-ended learning environment. *Advances in cognitive systems: Papers from the AAAI Fall Symposium* (pp. 297–304). Menlo Park, CA: AAAI Press.

Segedy, J. R., Kinnebrew, J. S., & Biswas, G. (2012). Supporting student learning using conversational agents in a teachable agent environment. In J. van Aalst, K. Thompson, M. J. Jacobson, & P. Reimann (Eds.), *The future of learning: Proceedings of the 10th international conference of the learning sciences (ICLS 2012): Vol. 2. Short Papers, Symposia, and Abstracts* (pp. 251–255). International Society of the Learning Sciences: Sydney, NSW, Australia.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. doi: 10.3102/0034654307313795.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*(3), 227–265.

James R. Segedy is a doctoral student in Educational Technology at the Institute for Software Integrated Systems at Vanderbilt University. He has an M.S. in Computer Science, and has experience teaching a college preparatory course to high school students. His research focuses on supporting students in independent and open-ended learning in computer-based learning environments.

John S. Kinnebrew is a researcher at the Institute for Software Integrated Systems at Vanderbilt University and received his Ph.D. in Computer Science from Vanderbilt University. His research focuses on the use of machine learning and data mining techniques to assess learning behaviors from activity traces of student interaction in computer-based learning environments.

Gautam Biswas is a Professor of Electrical Engineering and Computer Science at Vanderbilt University and has a Ph.D. in Computer Science from Michigan State University. He conducts research in Intelligent Systems with primary interests in hybrid modeling, simulation, and development of open-ended learning environments for math and science education, which includes the Teachable Agents project, where students learn science by building causal models of natural processes.