Behav Res (2013) 45:623–636
DOI 10.3758/s13428-013-0352-z

The development and analysis of tutorial dialogues in AutoTutor Lite

Christopher R. Wolfe · Colin L. Widmer · Valerie F. Reyna · Xiangen Hu · Elizabeth M. Cedillos · Christopher R. Fisher · Priscilla G. Brust-Renck · Triana C. Williams · Isabella Damas Vannucchi · Audrey M. Weil

Published online: 25 May 2013
© Psychonomic Society, Inc. 2013

Abstract The goal of intelligent tutoring systems (ITS) that interact in natural language is to emulate the benefits that a well-trained human tutor provides to students, by interpreting student answers and appropriately responding in order to encourage elaboration. BRCA Gist is an ITS developed using AutoTutor Lite, a Web-based version of AutoTutor. Fuzzy-trace theory theoretically motivated the development of BRCA Gist, which engages people in tutorial dialogues to teach them about genetic breast cancer risk. We describe an empirical method to create tutorial dialogues and fine-tune the calibration of BRCA Gist's semantic processing engine without a team of computer scientists. We created five interactive dialogues centered on pedagogic questions such as "What should someone do if she receives a positive result for genetic risk of breast cancer?" This method involved an iterative refinement process of repeated testing with different texts and successively making adjustments to the tutor's expectations and settings in order to improve performance. The goal of this method was to enable BRCA Gist to interpret and respond to answers in a manner that best facilitated learning. We developed a method to analyze the efficacy of the tutor's dialogues. We found that BRCA Gist's assessment of participants' answers was highly correlated with the quality of the answers found by trained human judges using a reliable rubric. The dialogue quality between users and BRCA Gist predicted performance on a breast cancer risk knowledge test completed after exposure to the tutor. The appropriateness of BRCA Gist's feedback also predicted the quality of answers and breast cancer risk knowledge test scores.

Keywords Intelligent tutoring system · Discourse technology · Auto tutor lite · Dialogue · Fuzzy-trace theory

Human one-on-one tutoring is arguably the "gold standard" for teaching complex conceptual material, with trained human tutors reportedly producing gains as high as two standard deviations over standard classroom practice, sometimes labeled the "2 sigma effect" (Bloom, 1984; Chi, Siler, Jeong, Yamauchi, & Hausmann, 2005), though a recent review of the literature has suggested that more modest effect sizes of about 0.79 are typical (VanLehn, 2011). Tutors have the ability to engage a student's attention, ask students questions, and give them immediate feedback on their progress (Graesser & McNamara, 2010). However, perhaps the greatest benefit of one-on-one tutoring is that tutors typically encourage their students to elaborate on their answers to knowledge questions (Chi et al., 2005).

Research suggests that actively generating and elaborating explanations of material is more beneficial to learning than is passively spending time with the material by reading or listening to lectures (Graesser, McNamara, & VanLehn, 2005). A current challenge is to create advanced learning technologies that routinely achieve the strong learning gains achieved by the best of well-trained human tutors. One promising tack is to develop learning systems that feature some of the same processes that lead to effective learning in

---

Electronic supplementary material The online version of this article (doi:10.3758/s13428-013-0352-z) contains supplementary material, which is available to authorized users.

C. R. Wolfe · C. L. Widmer · E. M. Cedillos · C. R. Fisher · T. C. Williams · I. Damas Vannucchi · A. M. Weil
Miami University, Oxford, OH, USA

V. F. Reyna · P. G. Brust-Renck
Cornell University, Ithaca, NY, USA

X. Hu
University of Memphis, Memphis, TN, USA

C. R. Wolfe (✉)
Department of Psychology, Miami University, Oxford, OH 45056, USA
e-mail: WolfeCR@MiamiOH.edu

one-on-one tutoring, which is the goal of most intelligent tutoring systems (ITS). These ITS facilitate human-computer interactions, which is meant to simulate the experience of a student talking with a human tutor (Graesser et al., 2004). One of the most promising methods for ITS to elicit self-explanation from students is to communicate with them using natural language.

AutoTutor Lite is a Web-based ITS that uses semantic decomposition to interact with people in natural languages such as English (Hu et al., 2009; Hu, Han, & Cai, 2008). AutoTutor Lite "stands on the shoulders" of AutoTutor, an ITS that has benefited from over two decades of systematic research and development. AutoTutor has been successfully applied to tutoring students in many knowledge domains, including computer science (Craig, Sullins, Witherspoon, & Gholson, 2006; Graesser et al., 2004), physics (Jackson, Ventura, Chewle, Graesser, & the Tutoring Research Group, 2004; VanLehn et al., 2007), and behavioral research methods (Arnott, Hastings, & Allbritton, 2008; Malatesta, Wiemer-Hastings, & Robertson, 2002). In AutoTutor, a talking agent facilitates communication with facial expressions and simulated facial movements, voice inflection, and conversational phrasing (Graesser, VanLehn, Rose, Jordan, & Harter, 2001). Graphical displays include animation or video with sound. At the heart of AutoTutor is the insight that when people actively generate explanations and justify their answers, learning is more effective and deeper than when learners are simply given information (Arnott et al., 2008). The explanations are pedagogically deep because the user must learn to express causal and functional relationships rather than mechanically applying procedures (VanLehn, Jones, & Chi, 1992).

Engineering an ITS such as AutoTutor to engage in a natural-language dialogue is complex and typically requires a team of highly experienced computer scientists, in addition to cognitive psychologists and content experts. AutoTutor's pattern of interaction is called *expectation and misconception tailored dialogue* (Graesser, Person, & Magliano, 1995; Graesser et al., 2001). This is accomplished through the development of curriculum scripts including each of the following elements: the ideal answer; a set of expectations; a set of likely misconceptions; responses for each misconception; a set of hints, prompts, and statements associated with each expectation; a set of key words; a set of synonyms; a canned summary to conclude the lesson; and a markup language to guide the actions of speech and gesture generators (Graesser et al., 2004). AutoTutor has a list of anticipated good answers (called *expectations*) and a list of *misconceptions* associated with each question. One goal is to encourage the user to cover the list of expectations. A second goal is to correct misconceptions exhibited in a person's responses and questions. Finally, a third goal is to give good feedback. The expectations associated with a question are stored in the curriculum script (Graesser, Chipman, Haynes, & Olney, 2005). AutoTutor answers questions; provides positive, neutral, and negative feedback; asks for more information; gives hints; prompts the user for specific missing words; fixes incorrect answers; and summarizes responses. AutoTutor's conversational agent provides pedagogic scaffolding (Graesser, McNamara, & VanLehn, 2005) to help people construct explanations. Controlled experiments (Arnott et al., 2008; Jackson et al., 2004) consistently demonstrate that AutoTutor is effective in helping people learn. In ten controlled experiments with over 1,000 participants, AutoTutor produced statistically significant gains of 0.2 to 1.5 standard deviations, with a mean of 0.81 (Graesser, Chipman, et al., 2005).

The research on AutoTutor has extended to a Web-based version called AutoTutor Lite (Hu et al., 2008, 2009; Hu & Martindale, 2008; Wolfe, Fisher, Reyna, & Hu, 2012). Perhaps the most important contribution of AutoTutor Lite is that it has the potential to allow developers to create effective tutorial dialogues without the team of highly experienced computer scientists needed to develop dialogues in other ITS.

Both AutoTutor and AutoTutor Lite have a talking animated agent interface, converse with users on the basis of expectations using hints and elaboration, use a speech act classifier (SAC) for speech act analysis, and present users with images, sounds, text, and video. They both compare the text entered by a student to a set of expectation texts using latent semantic analysis (LSA; Graesser et al., 2000; Landauer, Foltz, & Laham, 1998). LSA is a computational technique that mathematically measures the semantic similarity of sets of texts (Hu, Cai, Wiemer-Hastings, Graesser, & McNamara, 2007). It accomplishes this by creating a semantic space from a large corpus of text. The semantic space is a representation of the semantic relations of words on the basis of their co-occurrences in the corpus (Landauer & Dumais, 1997). In the context of an intelligent tutoring system, LSA is used to compare sentences entered by students to a specially prepared text that embodies good answers. The tutor can then give appropriate feedback to the student in order to encourage elaboration and other verbal responses on the basis of this comparison (Kopp, Britt, Millis, & Graesser, 2012).

Like human tutors and other ITS, AutoTutor Lite elicits verbal responses from learners and encourages them to further elaborate their understanding. AutoTutor Lite can thus be used to encourage self-explanation (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Chi, de Leeuw, Chiu, & LaVancher, 1994). Through a natural-language dialogue with the learner, AutoTutor Lite guides the learner toward a set of target expectations. With AutoTutor Lite, tutorials are built from units called sharable knowledge objects (SKOs). Each SKO presents materials to the learner didactically and then solicits a verbal response from the learner.

The didactic presentation is made by an animated talking agent with the ability to present text, still images, movie clips, and sounds.

We developed an ITS called Breast Cancer Genetics Intelligent Semantic Tutoring (BRCA Gist) using AutoTutor Lite. BRCA Gist is a Web-based ITS (Wolfe, Fisher, et al., 2012) that teaches women about genetic risk for breast cancer. Our goal was to create an ITS to engage women in a dialogue about the myriad difficult issues associated with genetic testing for breast cancer risk (Armstrong, Eisen, & Weber, 2000; Berliner, Fay, & the Practice Issues Subcommittee of the National Society of Genetic Counselors' Familial Cancer Risk Counseling Special Interest Group, 2007; Chao et al., 2003; Stefanek, Hartmann, & Nelson, 2001). Azevedo and Lajoie (1998) developed a prototype tutor to train radiology residents in diagnosing breast disease with mammograms. However, to the best of our knowledge, this is the first use of any ITS in the domain of patients' medical decision making. The content taught by the tutor was adapted from information on the National Cancer Institute's website, with input from medical experts.

We developed BRCA Gist guided by fuzzy-trace theory (FTT; Reyna, 2008a, b, 2012; Reyna & Brainerd, 1995). FTT is a dual-process theory (Reyna & Brainerd, 2011; Sloman, 1996) that holds that when information is encoded, people form multiple representations of information along a continuum: from verbatim representations that include a high amount of superficial detail, to gist representations that are fuzzier representations capturing the bottom-line meaning of information. An important difference between FTT and other dual-process theories is that when people make decisions, it is often *more* helpful to rely on these fuzzy gist representations (Reyna, 2008a). Thus, the manner in which BRCA Gist tutors is to encourage people to form useful gist representations rather than to drill them on specific verbatim facts. This is accomplished by presenting the concepts clearly with multiple explanations and figures that convey the bottom-line gist meaning of core concepts. Additionally, a medical expert reviewed the tutor to ensure accuracy.

In BRCA Gist, a speaking avatar delivers the content, and can present information as text, images, and videos in the provided space. In addition, the avatar can communicate with gestures such as head nodding and facial expressions. AutoTutor Lite provides 32 such commands, including "shake head," "make eyes wider," and "look confused."

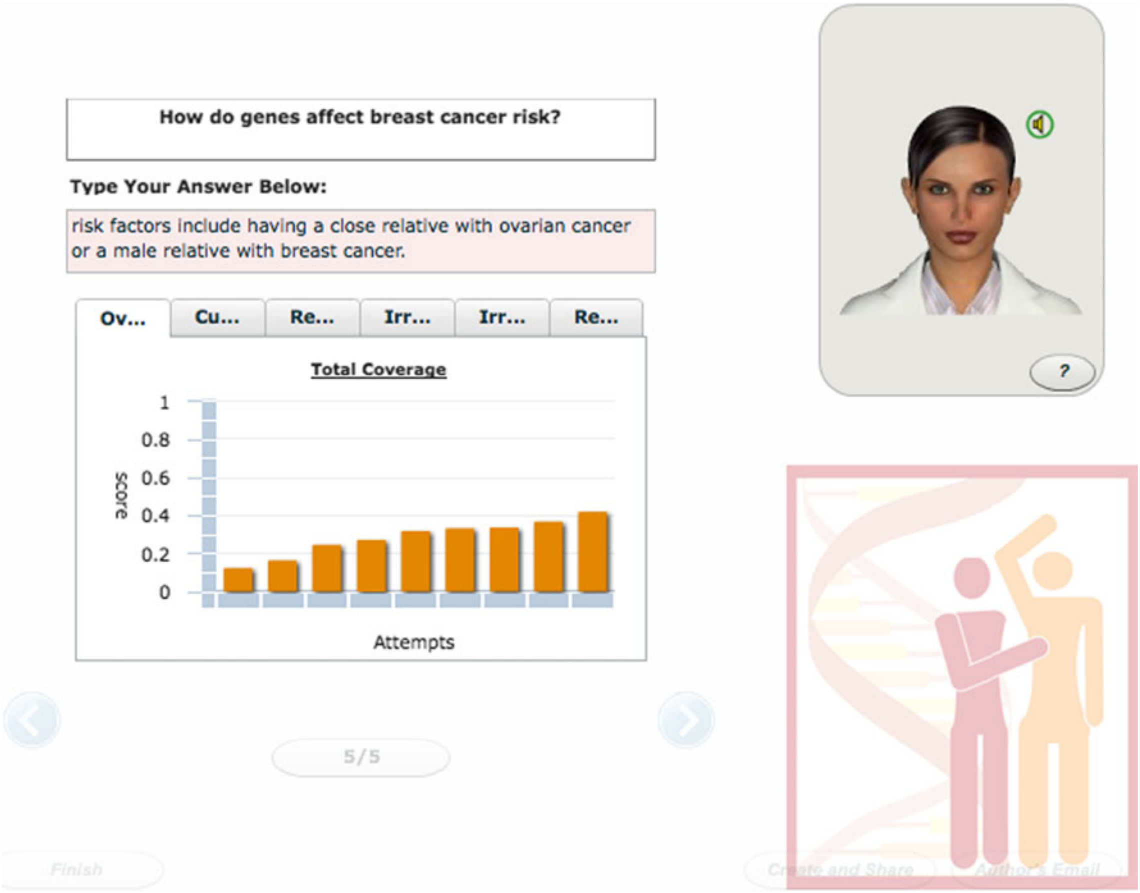
During the course of the BRCA Gist tutorial, participants interact with the tutor to answer five questions about genetic breast cancer risk. These interactions between the participants and BRCA Gist were developed guided by principles from prior work with ITS, primarily from the work of Arthur Graesser with AutoTutor (Graesser, 2011; Graesser, Chipman, et al., 2005). Three of these questions required participants to create a self-explanation about the material that they have just encountered by answering a question such as "What should someone do if she receives a positive result for genetic risk of breast cancer?" In addition to self-explanation, participants interacting with the tutor had to develop arguments and counterarguments (Wolfe, Britt, Petrovic, Albrecht, & Kopp, 2009), addressing questions such as "What is the case for genetic testing for breast cancer risk?" Research has shown that the creation of an argument can produce significant learning gains (Wiley & Voss, 1999) and can be successfully integrated into Web-based learning environments (Wolfe, 2001). Argumentation is key to learning in many disciplines (Wolfe, 2011). In each of these interactions, BRCA Gist is capable of giving responses and feedback using natural language. It does this by comparing the semantic similarity of a respondent's answers to a set of expectation texts.

A screen shot of BCA Gist from the learner's perspective can be found in Fig. 1. Here the avatar has asked, orally and in writing, "How do genes affect breast cancer risk?" The learner has composed a reply of nine sentences (or *turns*, in the parlance of AutoTutor Lite), with the last part of the last sentence reading "risk factors include having a close relative with ovarian cancer or a male relative with breast cancer." The bar graphs indicate that the participant has earned an overall CO score exceeding .4 (explained below).

Creating an effective tutor with natural-language dialogues

In order to successfully enhance learning, BRCA Gist needs to be able to effectively encourage participants to elaborate their answers. This occurs when the tutor responds appropriately to what participants say. AutoTutor Lite, with the ability to interact in natural language, is capable of encouraging elaboration and argumentation, but several challenges must be met in order for it to do so appropriately. AutoTutor Lite uses LSA, and this must be properly configured for the interactions to be successful. BRCA Gist requires expectation texts that reflect the *gist* of a good answer to the tutor's questions, so that it can compare input from participants to those expectations (Reyna, 2008a). If these expectations are not properly constructed, the tutor will be unable to make appropriate comparisons, and thus will be unable to respond appropriately.

BRCA Gist also needs a defined corpus of text that it can use to determine the mathematical similarity of texts. AutoTutor Lite is capable of using several such corpuses, but the most appropriate one must be identified. AutoTutor Lite also has many settings that can be adjusted to determine how BRCA Gist makes comparisons, such as the minimum association strength for words to be considered for comparisons. These settings must be calibrated as well, so that the



How do genes affect breast cancer risk?

Type Your Answer Below:

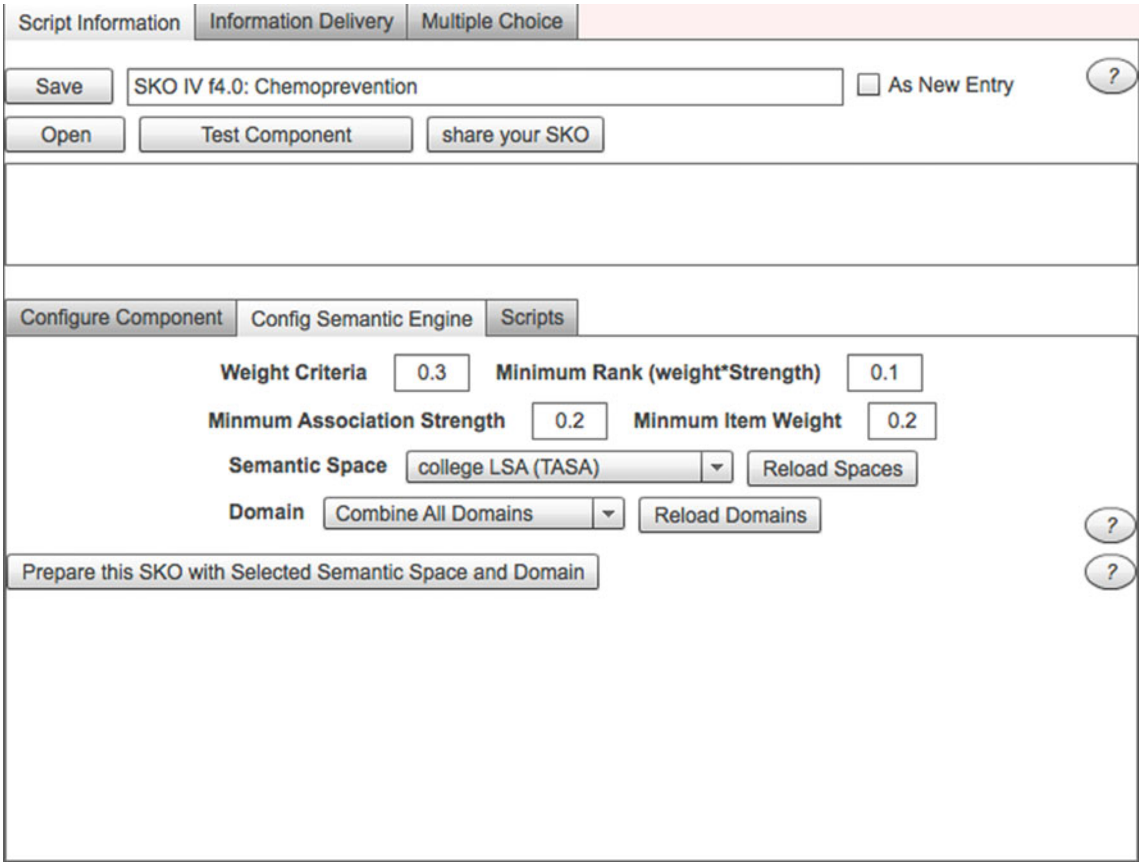
risk factors include having a close relative with ovarian cancer or a male relative with breast cancer.

Fig. 1 A screen shot from BRCA Gist. Ov stands for "Overall Coverage," Cu for "Current Contribution," the first Re for "Relevant New," the first Irr for "Irrelevant Old," the second Irr for "Irrelevant New," and the second Re for "Relevant Old"

tutor is best able to respond to participants' answers. Finally, the actual responses of the tutor must be created so that they will best respond to the potential answers that participants will create. The most accurate comparison between a participant's responses and expectations would be wasted if the tutor could not use it to respond meaningfully.

Figure 2 is a screen shot of the authoring tools used to configure AutoTutor Lite's semantic engine. AutoTutor Lite allows the designer to select from several semantic spaces, such as human free association (e.g., used by Nelson, McEvoy, & Schreiber, 2004) and college LSA (selected here). Several domains are available to choose from, including science and mathematics, computer and Internet, health, and environment. We found that it was best to combine all domains. The designer must decide about four numeric settings that determine the size and scope of the space in which the learner input will be compared to the expectation text. "Weight criteria" is the minimal weight that associated terms in the expectation text must have in order to be included in the space. There is a trade-off between novelty and the speed with which AutoTutor Lite can process the inputs. When AutoTutor Lite encounters a new constellation of domain, term, and space not previously encountered, it slows down the system. Judiciously selecting weights and other parameters thus helps improve performance. "Association strength" is the minimal association between words in the expectation text and similar terms, such as synonyms, that is required for them to be included. "Minimum rank" is the product of weight and strength, and this cutoff score is ultimately used to form the space. Finally, "minimum item weight" is the cutoff score for the weight of terms in the learner's input to be considered. So, for example, selecting a weight score of 0.1 or higher for the user-input sentence "I am sad" would knock out the higher-frequency words "I" and "am." Together, these parameters strongly affect the efficiency of the system and the ability of AutoTutor Lite to behave differently depending on how users respond to questions. Our methods for making these determinations are described below.

To make BRCA Gist interact with learners effectively, we developed an empirical method for creating and calibrating the semantic processing engine used by BRCA Gist. The first step of this process was to identify the appropriate information that could be included in a strong answer to each of the questions. To achieve this aim, several "ideal" answers to each question were written by three research



SKO IV f4.0: Chemoprevention

Weight Criteria 0.3
Minimum Rank (weight*Strength) 0.1
Minmum Association Strength 0.2
Minmum Item Weight 0.2
Semantic Space college LSA (TASA)
Domain Combine All Domains

Fig. 2 AutoTutor Lite authoring tools for configuring the semantic engine

assistants using the information from the National Cancer Institute's website. These ideal answers to questions, such as "How do genes affect breast cancer risk?" included a good deal of information that was relevant to answering each question, and thus were many times longer than the answer that we expected of actual participants interacting with the tutor. An essay written by one of the research assistants in answer to the question "What should someone do if she finds out that she has inherited an altered BRCA gene?" is provided as supplemental materials.

The next step was the creation of a reliable rubric to judge answers to these questions for the content that they contained. The ideal answers were examined for the individual items of relevant information that reflected a good answer. Once these information items were identified, they were used to create a rubric that could be used by raters to make judgments about the content of answers. Each piece of relevant information was included in the rubric as a separate item. The linked supplemental materials include the scoring rubric for the question, "what should someone do if she finds out that she has an inherited altered BRCA gene?" Raters could then judge each item as being either present or absent in an answer, using a gist scoring approach (Britt, Kurby, Dandotkar, & Wolfe, 2008), because gist-level mental representations have been demonstrated to be important for reasoning and decision making (Reyna, 2008a). We used a conditional reliability procedure to assess reliability while controlling for "absent" decisions. Two trained judges independently rated sample essays for each question. They used the rubric to assess whether each item was present or absent (see the supplemental materials for a sample rubric). To assess interrater reliability, we counted only instances in which at least one rater judged an item as being present. (Differences in the raters' judgments of whether an item was present were rare; see below.) Thus, there was agreement on an item if both raters marked it as present, and there was disagreement if one marked it as present and the other as absent. Items that both raters marked as absent were not counted (as this was the vast majority of items and would greatly inflate agreement ratings). The interrater reliability was thus the number of items that the two raters agreed on (both marked as present) divided by the total number of items marked "present" by at least one rater.

An important source of data that we used to calibrate BRCA Gist was answers to the same questions written by 81 untrained undergraduate participants, in order to get examples of the range of answers that participants might come up with when interacting with the tutor. We used these essays in a number of ways described in more detail below. Two independent raters used the rubric to make judgments about the content of the answers collected from the 81 untrained participants; the raters had an agreement of .89 on their judgments. Thus, the conditional probability that both judges would mark an item as being present, given that either of them had done so, was .89.

The next step in the method was to develop the actual expectation text that BRCA Gist would use to evaluate participants' input. The information in the ideal answers and the rubric was taken and reduced to a manageable size that reflected the core gist content of a good answer. This was done first by removing common function words such as "of" or "as," because these words are highly associated with all types of text and do not reflect specific content. Next, we removed instances of repetition and redundancy. Then additional text was removed in stages until the text that remained was composed of only words that reflected the core ideas that a good answer would include. In our experience with BRCA Gist, the best expectation texts were somewhat shorter than the 100 words recommended by AutoTutor Lite. To illustrate, the three ideal essays written by research assistants in response to the question "What should someone do if she finds out that she has inherited an altered BRCA1 or BRCA2 gene (meaning a positive test result for genetic breast cancer risk)?" (see the supplemental materials for an example) were ultimately "boiled down" to this 75-word expectation text:

> Genetic predisposition development breast cancer, positive test BRCA1, BRCA2, BRCA not necessarily cancer. Talk physician, genetic counselor. Measures prevent breast cancer. Manage cancer risk active surveillance, watching frequent cancer screenings, cancerous cells detected early, mammography, frequent clinical breast exams, examination MRI. Methods reduce risk breast, ovarian cancer, chemoprevention prophylactic mastectomy surgery, ovary removal. Goal reducing, eliminating risk cancer remove breast tissue, operation prophylactic mastectomy, chemo, Tamoxifen, chemoprevention. Cases breast cancer, ovarian cancer after prophylactic surgery.

When people interact with BRCA Gist, each sentence that they type in response to the question "What should someone do if she finds out that she has inherited an altered BRCA gene?" is compared with the expectation text above. BRCA Gist responds differently to different people, depending on the scores generated by the comparison of their verbal input to the expectation text. Therefore, BRCA Gist is a tailored health intervention that does not incur the typical added costs of tailoring, because the tailoring process is automated (Lairson, Chan, Chang, Junco, & Vernon, 2011).

A successive refinement process was then applied as a method to calibrate the LSA settings and corpus so that the tutor would best recognize appropriate answers to its questions. In this process, several different texts were entered multiple times as answers in the tutor for each question. As each text was entered, BRCA Gist's measure of the relatedness of the answer and the expectation text were noted, and settings were adjusted accordingly. In order to determine the degree to which a text covered the expectations according to the current settings, we looked at a text's CO score, a variable generated by AutoTutor Lite that represents the cumulative degree to which the expectations have been met by all of the learner's responses, combined across all sentences entered by the participant. CO stands for "overall coverage" and generally ranges from 0 to 1, with higher numbers corresponding to better coverage of the expectation text. We operationalized a good fit as a high CO score, indicating a high association between the input text and the expectation text.

The first text that we used was the exact same expectation text given above, entered one sentence at a time. This process does not guarantee one-to-one matching of identical text. We reasoned that expectation texts (and criteria settings) that did a poor job of recognizing identical text by giving it a low CO score would be unlikely to appropriately score input from actual learners. We adjusted the parameters for association strength and item weight and made modifications to the expectation texts accordingly.

To illustrate, in the case of the expectation text above, we started with a version that was 100 words long, but the best final CO score that we could obtain by feeding the same exact expectation text back as input one sentence at a time was only .44. Moreover, the mean change in CO score from one turn to the next was only .03, with the highest being just .05. We reasoned that these scores would be inadequate for distinguishing appropriate from inappropriate responses (because they meant that the program did not even recognize inputs that exactly matched expectations). AutoTutor Lite also produces scores for the extent to which a sentence of the input is relevant, but has already been provided previously (described in more detail below). These scores for relevant old information ranged from .51 to .95, suggesting that the 100-word expectation text had many redundancies. The reduced expectation text of 75 words (above) initially produced a final CO score of .53, but by changing the settings to a weight of .3, an association strength of .2, a minimum item weight of .2, and a minimum rank (the product of weight and strength) of .2, we obtained a higher final CO score. Those settings produced a final CO score of .66 and a mean change of the CO score per turn of .05. This higher final CO score and steady rise from one turn to the next suggested that this expectation text with these settings might be adequate to allow BRCA Gist to distinguish good answers from poor answers. By way of contrast, the same procedure and settings using only health rather than all of the domains combined produced a final CO score of .53, and the same procedure using human association norms instead of LSA also produced a final CO score of .53.

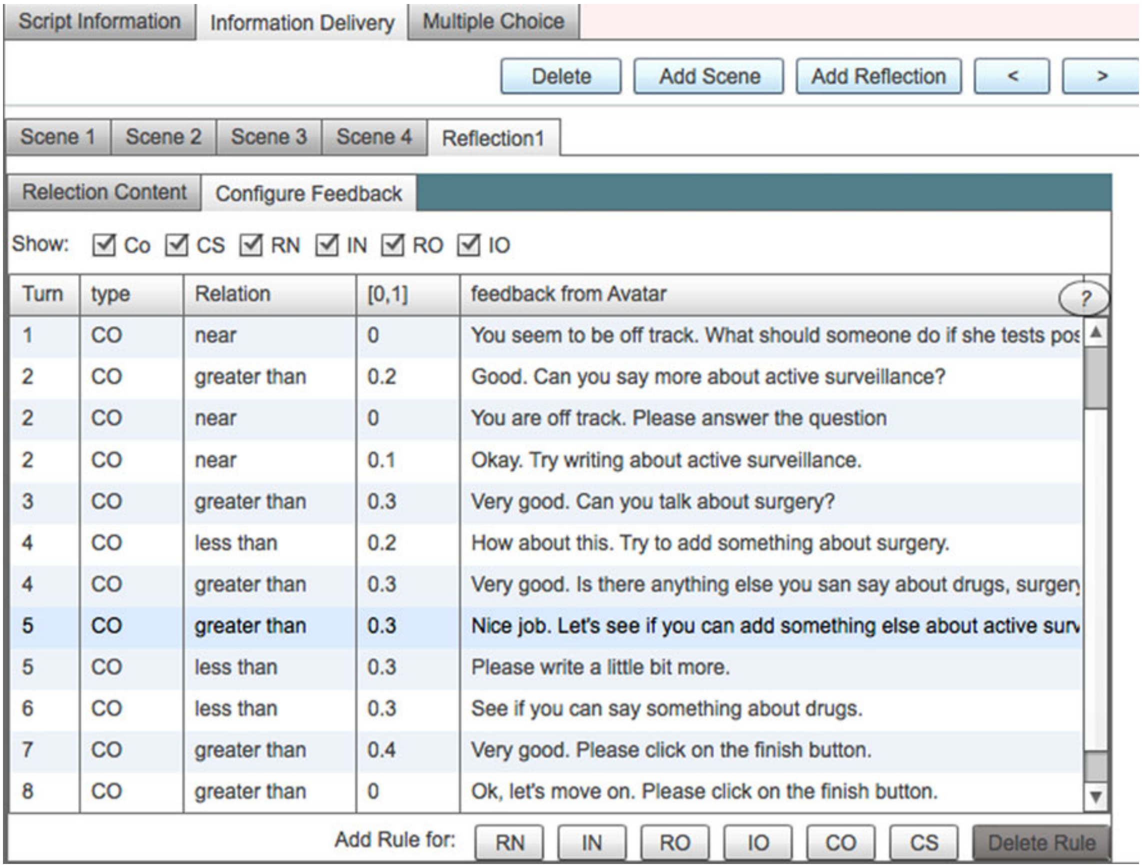
After the settings were adjusted such that the tutor was recognizing the identical expectation text at the highest level possible, the next texts used as answers were the full ideal answers generated by the research assistants. These were entered in order to ensure that the tutor could recognize as similar texts that were written to reflect a very good answer, and additional adjustments were made. Although having an expectation text and settings that produce successively higher CO scores is good, the real trick is to distinguish good responses from poor ones. Toward this end, next, irrelevant texts such as the lyrics to "Take Me Out to the Ball Game" were entered into the tutor. This was done in order to make sure the tutor was also capable of appropriately discriminating irrelevant text that did not address its questions. This would allow the tutor to give appropriate feedback in order to try to get participants who were off track to answer the questions appropriately. Fortunately, this was accomplished without difficulty for each of the five tutorial interactions. For example, for the question about positive test results, "Take Me Out to the Ball Game" yielded a final CO score of .21 with a mean change per turn of just .019.

We then developed reasonable answers to the questions that used synonyms of the terms found in the ideal answers. These answers appropriately addressed the questions with good information, but did not use many of the same words that were contained in the expectation text. This was done in order to make sure the tutor had been constructed and calibrated to appropriately judge real semantic similarity, rather than look for specific words. According to FTT, it is important to credit gist comprehension of the curriculum (as opposed to simply verbatim parroting of the curriculum). This is also important because the tutor must respond appropriately to the variety of possible correct answers that participants might write, which might or might not include the words in the expectation text.

One final step to calibrate the settings was to enter samples from the answers created by some of the 81 untrained participants. Both good and poor examples from these answers were entered into the tutor. This ensured that the tutor was capable of responding to actual answers that might be given by participants, and tested that the tutor could correctly differentiate between better and worse answers to the questions. To illustrate, among the essays written in response to the positive-test-result question, an essay that we identified as being relatively good yielded a final CO score of .474, and an essay that we judged as being not as good (but among the better responses) produced a final CO score of .397. One issue that we discovered using these texts was that most of the responses provided by untrained participants were not particularly good.

In addition to changing the settings at each step of this process, we also adjusted the feedback that the tutor gave in response to participants' statements. AutoTutor Lite generates a number of scores each time that a sentence is entered to answer one of the questions. After a user types in each sentence (or *turn*, in the lingo of AutoTutor Lite) in addition to CO, the semantic processing engine provides four other numeric scores that the system can use in generating a verbal response. For each rule on each turn, the authoring tools of AutoTutor Lite permit the designer to select greater than, less than, or near (±.05) values, from 0 to 1 in increments of .1, to characterize information as RN (relevant new), RO (relevant old), IN (irrelevant new), and IO (irrelevant old), which sum to 1.0 on each turn and for CO (see Fig. 3). If more than one rule matches on a given turn, AutoTutor selects randomly among them. However, BRCA Gist was designed to provide only one match on a given turn by using mutually exclusive feedback rules: Specifically, information is scored as relevant to the extent that it conforms to the established expectations. New information is simply that which the learner has not already entered in a previous turn. Thus, a high RN score indicates that the learner has met the expectations to a high degree on that turn. RO refers to repeated information that is relevant to the expectations, but has also been covered by a previous turn. IN is new information that deviates from the expectations. IO stands for "irrelevant old information"—that is, information that has already been presented in a previous turn—and is not relevant to the expectations. As previously noted, CO stands for "overall coverage." In developing BRCA Gist, we found coverage score (CO) to be the most useful score for setting rules early in the dialogue, and sometimes used the other scores to guide the tutor's verbal interactions toward the end of a dialogue. Using rules based on CO scores allowed for appropriate feedback to be given on the basis of a participant's progress in answering a question. Examples of the feedback rules we created for the interaction in which we asked participants what someone should do if she finds out that she has inherited an altered BRCA gene can be found in Fig. 3.

AutoTutor Lite cannot detect anything more specific about the learner's input beyond the scores generated above. Thus, a CO score of .2 on the first turn should be considered an appropriate response. However, in the case of the question about positive test results, AutoTutor Lite could not tell the difference between a good sentence suggesting that a woman should talk to her physician or genetic counselor about risk and a sentence suggesting that she should reduce environmental risk factors for developing breast cancer such as decreasing alcohol consumption and eating healthier foods. In creating the tutorial dialogue feedback for BRCA Gist, we used the essays generated by the research assistants and those produced by the untrained participants to guide



Show: ☑ Co ☑ CS ☑ RN ☑ IN ☑ RO ☑ IO

Turn	type	Relation	[0,1]	feedback from Avatar
1	CO	near	0	You seem to be off track. What should someone do if she tests pos
2	CO	greater than	0.2	Good. Can you say more about active surveillance?
2	CO	near	0	You are off track. Please answer the question
2	CO	near	0.1	Okay. Try writing about active surveillance.
3	CO	greater than	0.3	Very good. Can you talk about surgery?
4	CO	less than	0.2	How about this. Try to add something about surgery.
4	CO	greater than	0.3	Very good. Is there anything else you san say about drugs, surger
5	CO	greater than	0.3	Nice job. Let's see if you can add something else about active surv
5	CO	less than	0.3	Please write a little bit more.
6	CO	less than	0.3	See if you can say something about drugs.
7	CO	greater than	0.4	Very good. Please click on the finish button.
8	CO	greater than	0	Ok, let's move on. Please click on the finish button.

Fig. 3 AutoTutor Lite authoring tools for configuring verbal feedback. CO stands for "overall coverage," RN stands for "relevant new," RO stands for "relevant old," IN stands for "irrelevant new," and IO stands for "irrelevant old"

the order of feedback. In the case of the question about positive test results, we started prompting participants about active surveillance, then surgery, and then drugs, because this was the order most typically found in the essays (for an example, see the supplemental materials). Using this approach, we were more likely to provide specific feedback about breast cancer and genetic risk that also matched the specific issues that participants were considering. Our strategy was to always prompt participants who scored low on the very first turn, to let them know that they were off track. Participants who scored very well on the first turn also often received comments, but those in between did not receive feedback until the second turn.

The efficacy of the BRCA Gist tutorial dialogues

Once we created BRCA Gist's tutorial dialogues by our empirical method, we assessed their quality in interacting with research participants. The assessment of the dialogues was embedded in a larger randomized, controlled study of the effectiveness of BRCA Gist in teaching women about genetic risk of breast cancer. A report of that experiment (in contrast to the scope of this article, which focuses on the development and assessment of the tutorial dialogues) compared the efficacy of BRCA Gist to that of reading text from the National Cancer Institute website and from an irrelevant nutrition control (Wolfe, Reyna, et al., 2012). In the study, 64 undergraduate women at Miami University and Cornell University interacted with the BRCA Gist tutor and participated in the natural-language tutorial dialogues. Participants self-reported their age as between 18–22 years, 76 % described themselves as white, 15 % Asian, 6 % African American, and 6 % Latina in non-mutually-exclusive categories. Participants received the entire BRCA Gist tutorial, which lasts approximately 90 min. After completing the tutorial, participants completed a multiple choice knowledge test about breast cancer and genetic risk (and a number of other tasks). Our goals in this analysis were to determine whether BRCA Gist's assessment of the similarity of answers to the expectations was a reliable measure of the quality of those answers, whether the quality of the answers predicted learning, and whether the success of the interactions had an effect on learning.

It was important to establish that BRCA Gist's judgments about the semantic similarity of participants' answers were actually capturing how much content was in those answers. That is, if the tutor were interpreting answers appropriately, an answer that was given a higher measure of semantic similarity should contain more relevant content than would an answer that was given a lower score. Our method to examine this was to use the final CO score for the last sentence entered by each participant. This score measures the tutor's judgment of the semantic similarity of the entire answer to the expectation text. To determine whether the CO scores accurately measured the amount of content in an answer, we compared BRCA Gist's final CO scores to scores obtained by applying our rubrics blind to CO score. To ensure that the rubric measures were reliable, two independent trained raters used the rubric to make judgments of about one third of the answers. Applying the same conditional reliability procedure that had been used with the essays by untrained participants, the two judges had .87 agreement.

To assess the effect of the dialogues on learning, we used the score on a 32-item multiple choice test that measures knowledge of genetic risk of breast cancer. To develop reliable knowledge items, we drew upon pages from comparable sections of the National Cancer Institute website to write potential items and had all potential items vetted by a medical expert. Originally, we developed 45 items drawing on the full range of content. We then tested these items on 82 untrained participants while we were developing BRCA Gist. We selected the 32 items with the best psychometric properties—specifically, those that produced the highest value of Cronbach's alpha and did not produce either a ceiling or a floor effect. For the untrained participants, Cronbach's alpha was .67, and the mean was 57 % correct. These scores are solid, considering that a wide range of content, ranging from biology to antidiscrimination laws, helps women decide whether or not to undergo genetic testing. The 32-item multiple choice test is provided in the linked supplemental materials.

A method to assess the success of the interactions between BRCA Gist and participants was also needed. To measure this, we judged the appropriateness of each response made by BRCA Gist. At the time of scoring, responses could be judged appropriate, inappropriate, or neutral. We used a gist-scoring procedure to make one of three judgments for each response made by BRCA Gist. Judgments were made only in relation to the user's previous statement, and not to the entirety of the dialogue. The criteria for the appropriateness of responses was that the tutor's response did each of the following: encouraged elaboration, flowed naturally from the previous input, and responded correctly to the accuracy of the participant's input. The criteria for inappropriateness were the opposites: An inappropriate response discouraged elaboration, did not flow naturally from the previous input, and responded incorrectly to the accuracy of the participant's input. The neutral option was used when it was difficult to classify a response as clearly one or the other (such as if it encouraged elaboration and responded correctly, but didn't flow as naturally as most appropriate responses did). During the analysis phase, we collapsed to a distinction between appropriate and inappropriate, and neutral judgments were categorized as inappropriate responses. Thus, to be considered appropriate, the BRCA Gist response ultimately had to meet all three criteria. Approximately one third of the responses were used to train the judges. To ensure that their judgments were reliable, two independent raters made judgments about one third of the responses. To calculate reliability, we looked at the proportions of responses that the two judges agreed on over the total number of responses, and found .86 agreement.

To illustrate, the following selection is part of a dialogue between BRCA Gist and a participant (spelling mistakes and typos included) in which each response was judged to be appropriate.

> Participant: They should also meet with the genetic counselor again to discuss their options for what to do next.
> BRCA Gist: Good. Can you say more about active surveillance?
> Participant: Active survelliance closely watches women and their risks, looking for any changes in their breasts or bodies that would hint at breast cancer forming
> BRCA Gist: Good. Can you talk about surgery?

Here, BRCA Gist recognized that her response was good and encouraged her to talk first about active surveillance and then surgery. By way of contrast, the following dialogue was found to be inadequate:

> Participant: Surveillance includes mammograms, MRI
> BRCA Gist: You could be making better progress. Try writing about active surveillance
> Participant: This surgery does not ensure that she will not get breast cancer, but the removal of healthy tissue that could potentially lead to breast cancer can lower someone's risk of breast cancer
> BRCA Gist: Good. Can you talk about surgery?
> Participant: There is also medication that can be taken to help reduce the risk of develping breast cancer
> BRCA Gist: See if you can add something about drugs.

Unfortunately, BRCA Gist did not recognize that the participant made an accurate statement about surveillance. Moreover, it prompted her to say something about surgery and drugs after she had just mentioned these options—without acknowledging that they were already part of the conversation. Annotated examples of complete appropriate and inappropriate dialogues, including changes in CO scores, are included in the Appendix.

Results

The goals of these analyses were to determine whether BRCA Gist's assessment of the similarity of answers to the expectations (i.e., final CO score) was a reliable measure for the quality of those answers, whether the quality of a participant's verbal statements predicted scores on a subsequent knowledge test, and whether the success of BRCA Gist in responding appropriately during the dialogue interactions had an effect on knowledge test scores.

The BRCA Gist participants produced answers with an average final CO score of .41 (*SD* = .14), creating answers using an average of 5.37 sentences (*SD* = 1.75). Rubric judgments determined that these answers contained an average of about a quarter of the total possible content items for each answer (*M* = .25, *SD* = .10). This compares favorably to the previously obtained rubric scores from 81 untrained undergraduates who wrote brief essays without the benefit of any tutorial (*M* = .147, *SD* = .057). Because the testing conditions differed dramatically between these two samples, the use of statistics to make inferences about them was unwarranted. CO scores and rubric scores from the BRCA Gist tutorial dialogues were highly correlated, $r(62) = .75$, $p < .001$. Thus, the BRCA Gist semantic processing engine's final CO score accounted for over half of the variance in assessments of the thoroughness of participants' verbal responses made by trained human judges using a reliable rubric.

The proportion of tutor responses judged to be appropriate (*M* = .85, *SD* = .24) was highly correlated with both CO score, $r(62) = .82$, $p < .001$, and rubric score, $r(62) = .74$, $p < .001$. This finding demonstrates that the greater the extent to which BRCA Gist responded appropriately to participant input, the greater was the proportion of expectations covered by participants' complete answers. The correlations between the appropriateness of responses and the quality of participants' final answers remained significant even when the number of sentences in their answers was controlled for, in terms of both CO score, $t(62) = 6.11$, $p < .001$, and rubric score, $t(62) = 4.55$, $p < .001$. This indicates that these correlations were more than a simple case of the tutor failing to respond appropriately to incomplete, short answers.

We also found positive correlations between performance on the knowledge test (*M* = .74, *SD* = .15) and final CO score, $r(62) = .35$, $p = .004$, and between performance on the knowledge test and rubric score, $r(62) = .67$, $p < .001$. The appropriateness of tutor responses was also correlated with performance on the knowledge test, $r(62), = .41$, $p < .001$. When the tutor responded appropriately, participants produced fuller answers reflecting greater knowledge of breast cancer and genetic risk. These answers within the tutorial dialogue predicted subsequent knowledge, as measured by the multiple choice test. Moreover, the mean score of 74 % correct compared favorably with the 57 % correct found among untrained participants in the earlier study. These results were found in the context of a larger experiment in which participants who were randomly assigned to the BRCA Gist condition scored significantly higher than those in other conditions (Wolfe, Reyna, et al., 2012).

Discussion

These results provide evidence that the tutorial dialogues were suitable and beneficial for learning. One implication of these results is that the semantic similarity of answers as judged by the tutor seems to have captured much of the same information as trained human judgments. This is encouraging, because the rubrics were found to be a reliable method of assessing the amount of good content provided in an answer. Our results demonstrate that an ITS that uses natural-language processing to interpret answers, such as BRCA Gist, can capture much of the content of users' statements by using a carefully constructed and calibrated semantic processing engine. That is, once the tutor's expectations were deliberately created through an empirical process, BRCA Gist was capable of making responses on the basis of scores generated by a semantic processing engine that corresponded to the expectations of a trained human researcher.

The correlation of .75 between BRCA Gist final CO scores and human judgments is comparable to those obtained with more sophisticated systems. For example, the Reading Strategy Assessment Tool (RSAT) yielded processing scores that correlated with human judgments from $r = .78$ to .48 (Magliano, Millis, the RSAT Development Team, Levinstein, & Boonthum, 2011). AutoTutor's evaluation of whether or not students stated expectations was correlated with the evaluations of a human expert at $r = .50$, nearly as high as the $r = .63$ correlation between two experts (Magliano & Graesser, 2012, p. 612). Similarly, McNamara, Levinstein, and Boonthum (2004) compared human judgments to those of iSTART and found that iSTART scores successfully distinguished between paraphrases and explanations containing some kind of elaboration.

These results also provide evidence that the amount of relevant content that people show in answers to specific questions asked by the tutor can predict their general knowledge of the content taught by the tutor. Participants who wrote more elaborate, content-heavy answers when interacting with the tutor also performed better on the knowledge test taken later. This finding is consistent with what is known about the learning gains provided by self-explanation and elaborative answers (Chi, 2000; Chi et al., 1994; Roscoe & Chi, 2008).

The success of the interaction between the tutor and the participants was associated with greater knowledge. Participants with more successful interactions (measured as a greater proportion of appropriate tutor responses) showed more content in their answers, according to both the tutor's CO scores and human judgments using the rubric. This relationship held true even when controlling for the length of participants' answers. Of course, this suggests that the converse was also true: To the extent that the tutor responds inappropriately, people are less likely to provide good answers and more likely to score poorly on a subsequent knowledge test. These findings suggest that BRCA Gist has not reached a ceiling with respect to responding appropriately to participant input. We readily found examples of BRCA Gist responding appropriately to good and poor answers, with about 85 % of BRCA Gist's responses being judged appropriate. Thus, it is not the case that only better participants received appropriate responses from BRCA Gist.

It appears that the quality of the interactions between the participant and the tutor affects the quality of the participant's answers, rather than the quality of the answers only reflecting the knowledge that the participant brought to the interaction. Successful interaction with the tutor was associated with participants writing more complete, elaborated answers. Not only did participants with more successful interactions include more content in their answers, but they showed better performance on the knowledge test. The evidence seems to suggest that interaction with the tutor had a positive effect on knowledge. Participants who interacted with BRCA Gist performed better on the knowledge test than did the untrained participants in the test development study, and Wolfe, Reyna, et al. (2012) reported that BRCA Gist participants did better than participants randomly assigned to two comparison groups. However, it is logically possible that all of the benefits of BRCA Gist could stem from the didactic portions of the tutorial (i.e., images and clear explanations presented orally and in text), and that the association between tutorial dialogues and outcome variables simply reflect smarter participants producing better verbal answers, yielding fewer inappropriate responses on the part of BRCA Gist, and also producing better answers on the knowledge test. The beneficial effects of interaction cannot be fully assessed without randomized experiments that present the same didactic information with and without tutorial dialogues.

Another limitation is that our design included only a posttest, without a pretest. Although this would be adequate for a fuller randomized, controlled study, without a pretest the effects presented here could theoretically be due to greater initial knowledge rather than learning. However, the results presented by Wolfe, Reyna, et al. (2012) seem to rule out this interpretation.

The results of this analysis will also prove useful for the future development of BRCA Gist. The knowledge that we gained from this study will allow us to apply these empirical methods to improve the tutor's assessments of participants' answers. We plan to make further adjustments to the expectation text, settings, and responses using these results. We also learned some practical lessons from this study about what kinds of responses promote more or less elaboration. For example, an unanticipated problem with BRCA Gist responses was that they too often suggested a specific topic that proved puzzling to participants who were already talking about that topic. For example, the suggestion "Can you say anything about cells and tumors?" will sound odd to a person who just entered a sentence about tumors. However, the suggestion "Can you say *more* about cells and tumors?" would sound reasonable to both a person who had just entered a sentence about tumors and a person who had not done so. Because subtle changes in the BRCA Gist responses can yield noticeable differences in how much elaboration they produce, careful thought anticipating the sentences that people may use to answer the tutor is required.

AutoTutor Lite does not really "understand" what users are saying in the way that a human tutor does, or even in the way that some ITS such as AutoTutor understand natural language. In developing BRCA Gist, our solution to this shortcoming was to focus on the verbal behavior of the learner. BRCA Gist interacts with learners and encourages them to expand upon key points first presented in theoretically motivated didactic lessons. It uses specific hints such as "can you talk about surgery," rather than vague prompts such as "good job" or "please continue." BRCA Gist is designed to tell users whether or not they are on track, and it can use linguistic devices such as "say *more* about active surveillance" to make the same response appear appropriate to a wider array of potential verbal input.

Although AutoTutor Lite lacks the ability to diagnose and address misconceptions, BRCA Gist allows learners to experience some of the benefits of self-explanation (Roscoe & Chi, 2008) and argument generation (Wiley & Voss, 1999). It appears that the interactions between learners and BRCA Gist are at a suitable grain size or level of interaction granularity (VanLehn, 2011) to be consistent with learning. It is also likely that the materials taught by BRCA Gist are relatively sophisticated for lay people, though far from a level of true expertise. VanLehn et al. (2007) found that eliciting an explanation from participants was superior to providing them with an explanation when novices were taught content appropriate for intermediate students. However, an emerging literature on vicarious learning suggests that tutorial dialogues are not always necessary for an ITS to produce learning, especially when learners are given deep-level reasoning questions (Craig et al., 2006). Again, the BRCA Gist approach to tutorial dialogues appears well situated to produce deep-level learning; however, systematic experimentation will be needed to test these hypotheses.

To guide the communication of future generations of BRCA Gist, we are collecting texts about breast cancer and genetic risk from a variety of published sources to be used to create a corpus for LSA that is specifically about the knowledge domain of genetic risk and breast cancer. To date, we have collected over one million words of meaningful texts from sources such as the NCI website. We hope that the semantic space created by this corpus will be even more effective at assessing users' answers, and will allow future generations of BRCA Gist to engage in mixed-initiative tutorial dialogues (Graesser, Chipman, et al., 2005), in the sense that both a person and BRCA Gist will be able to initiate strands of conversation.

Author note The project described was supported by Award No. R21CA149796 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. We thank the National Cancer Institute for its support. The authors thank Nananda Col for her expert feedback on our research materials and on BRCA Gist. The authors also thank Sharjeel Chaudry, Suveera Dang, Amrita Rao, Anna Catherine Morant, Kate Bassolino, Eric Cooke, Jessica Reigrut, and Mandy Withrow for assistance with data collection.

Appendix: Examples of successful and unsuccessful dialogues between BRCA Gist and research participants

Participant 62, with a good interaction and high CO score

Turn 1: There are multiple options for someone who has tested positive for BRCA mutations (CO = .149)
BRCA Gist Response 1: What should someone do if she tests positive? (Appropriate)
Turn 2: First of all, they can undergo frequent surveillance, which is done through clinical testing, MRI or mammograms (CO = .213)
BRCA Gist Response 2: Good. Can you say more about active surveillance? (Appropriate)
Turn 3: This does not change someone's risk of developing breast cancer, but it enables the clinician to discover the cancre at an early stage, when it is more treatable and less likely to be fatal (CO = .319)
BRCA Gist Response 3: Good. Can you talk about surgery?
Turn 4: Another option is to undergo prophylactic surgery, which entails removing as much of the ovarian or breast tissue as possible to reduce the chances of developing breast cancer (CO = .396)
Turn 5: Another option is to undergo prophylactic surgery, which entails removing as much of the ovarian or breast tissue as possible to reduce the chances of developing breast cancer or ovarian cancer (CO = .429)
BRCA Gist Response 4: Very good. Is there anything else you can say about drugs, surgery, and active surveillance? (Appropriate)
Turn 6: However, this does not completely eliminate the risk of developing breast cancer (CO = .438)
Turn 7: Lastly, there is a drug that can be taken which greatly reduces the chances of developing breast cancer (CO = .448)
BRCA Gist Response 7: Very good. Please click on the finish button. (Appropriate)
Turn 8: However, it can cause stroke, DVT or endometrial cancer (CO = .460)

Participant 268, with a poor interaction and low CO score

Turn 1: She shoud do a survillence (CO = .010)
BRCA Gist Response 1: You seem to be off track. What should someone do if she tests positive for BRCA mutations. (Inappropriate)
Turn 2: She should survillence herself so that she can catch the cancer early (CO = .131)
BRCA Gist Response 2: You could be making better progress. Try writing about active surveillance (Inappropriate)
Turn 3: In active surveillance you are able to test yourself for cancer (CO = .181)
Turn 4: In catching the cacner earlier you are at a higher risk for survival (CO = .220)
BRCA Gist Response 4: Try to add something about surgery. (Appropriate)
Turn 5: Then you are able to have surgery to take away the cancer (CO = .238)
Turn 6: When you have surgery it will get rid of the cancer cells (CO = .250)
BRCA Gist Response 6: See if you can add something about drugs. (Appropriate)
Turn 7: When you have surgery it will get rid of the cancer cells (CO = .250)
Turn 8: Having surgery at a earlier point in the cancer can get rid of the majority of the cancer before it divides (CO = .285)
BRCA Gist Response 8: Try to add something about active surveillance, drugs, or surgery. (Appropriate)

Note: Typos are included from original dialogue.

References

Armstrong, K., Eisen, A., & Weber, B. (2000). Assessing the risk of breast cancer. *The New England Journal of Medicine, 342,* 564–571.

Arnott, E., Hastings, P., & Allbritton, D. (2008). Research methods tutor: Evaluation of a dialogue-based tutoring system in the classroom. *Behavior Research Methods, 40,* 694–698. doi:10.3758/BRM.40.3.694

Azevedo, R., & Lajoie, S. P. (1998). The cognitive basis for the design of a mammography interpretation tutor. *International Journal of Artificial Intelligence in Education, 9,* 32–44.

Berliner, J. L., Fay, A. M., & the Practice Issues Subcommittee of the National Society of Genetic Counselors' Familial Cancer Risk Counseling Special Interest Group. (2007). Risk assessment and genetic counseling for hereditary breast and ovarian cancer: Recommendations of the National Society of Genetic Counselors. *Journal of Genetic Counseling, 16,* 241–260. doi:10.1007/s10897-007-9090-7

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Research, 13,* 4–16.

Britt, M. A., Kurby, C. A., Dandotkar, S., & Wolfe, C. R. (2008). I agreed with what? Memory for simple argument claims. *Discourse Processes, 45,* 52–84.

Chao, C., Studts, J. L., Abell, T., Hadley, T., Roetzer, L., Dineen, S., . . . McMasters, K. M. (2003). Adjuvant chemotherapy for breast cancer: How presentation of recurrence risk influences decision-making. *Journal of Clinical Oncology*, *21*, 4299–4305.

Chi, M. T. (2000). Self-explaining expository texts: The duel processes of generating inferences and repairing mental models. *Advances in Instructional Psychology, 5,* 161–238.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 15,* 145–182.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439–477.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2005). Learning from human tutoring. *Cognitive Science, 25,* 471–533.

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). The deep-level-reasoning-question effect: The role of dialogue and deep-level-reasoning questions during vicarious learning. *Cognition and Instruction, 24,* 565–591.

Graesser, A. C. (2011). Learning, thinking, and emoting with discourse technologies. *American Psychologist, 66,* 746–757.

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education, 48,* 612–618.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers, 36,* 180–192. doi:10.3758/BF03195563

Graesser, A., & McNamara, D. (2010). Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist, 45,* 234–244.

Graesser, A. C., McNamara, D. S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point & Query, AutoTutor, and iSTART. *Educational Psychologist, 40,* 225–234.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology, 9,* 359–387.

Graesser, A. C., VanLehn, K., Rose, C., Jordan, P., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine, 22,* 39–51.

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Tutoring Research Group, & Person, N. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments, 8,* 129–147.

Hu, X., Cai, Z., Han, L., Craig, S. D., Wang, T., & Graesser, A. C. (2009). AutoTutor Lite. In *Proceedings of the 2009 Conference of Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modeling* (p. 802). Amsterdam, The Netherlands: IOS Press.

Hu, X., Cai, Z., Wiemer-Hastings, P., Graesser, A. C., & McNamara, D. S. (2007). Strengths, limitations, and extensions of LSA. In D. McNamara, T. Landauer, S. Dennis, & W. Kintsch (Eds.), *LSA: A road to meaning* (pp. 401–425). Mahwah, NJ: Erlbaum.

Hu, X., Han, L., & Cai, Z. (2008). *Semantic decomposition of student's contributions: An implementation of LCC in AutoTutor Lite*. Paper presented to the Society for Computers in Psychology, Chicago, Illinois: November 13, 2008. https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxzY2lwd3N8Z3g6NmZhMTE2NWI3OGRiMzBlMw&pli=1 Abstract accessed December 15, 2012.

Hu, X., & Martindale, T. (2008). Enhance learning with ITS style interactions between learner and content. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC). Proceedings paper accessed December 12, 2011, from http://legacy.adlnet.gov/SiteCollectionDocuments/files/8218Paper.pdf

Jackson, G. T., Ventura, M. J., Chewle, P., Graesser, A. C., & the Tutoring Research Group. (2004). The impact of why/AutoTutor on learning and retention of conceptual physics. In J. C. Lester, R. M. Vicari, & F. Paraguacu (Eds.), *Intelligent tutoring systems 2004* (pp. 501–510). Berlin, Germany: Springer.

Kopp, K. J., Britt, M. A., Millis, K., & Graesser, A. C. (2012). Improving the efficiency of dialogue in tutoring. *Learning and Instruction, 22,* 320–330.

Lairson, D. R., Chan, W., Chang, Y.-C., Junco, D. J., & Vernon, S. W. (2011). Cost-effectiveness of targeted vs. tailored interventions to promote mammography screening among women military veterans in the United States. *Evaluation and Program Planning, 34,* 97–104. doi:10.1016/j.evalprogplan.2010.07.003

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240. doi:10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284. doi:10.1080/01638539809545028

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. *Behavior Research Methods, 44,* 608–621. doi:10.3758/s13428-012-0211-3

Magliano, J. P., Millis, K. K., the RSAT Development Team, Levinstein, I., & Boonthum, C. (2011). Assessing comprehension during reading with the Reading Strategy Assessment Tool (RSAT). *Metacognition and Learning, 6,* 131–154.

Malatesta, K., Wiemer-Hastings, P., & Robertson, J. (2002). Beyond the short answer question with research methods tutor. In S. A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Intelligent Tutoring Systems 6th International Conference LNCS 2363* (pp. 562–573). Berlin, Germany: Springer.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers, 36,* 222–233. doi:10.3758/BF03195567

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers, 36,* 402–407. doi:10.3758/BF03195588

Reyna, V. F. (2008a). A theory of medical decision making and health: Fuzzy trace theory. *Medical Decision Making, 28,* 850–865.

Reyna, V. F. (2008b). Theories of medical decision making and health: An evidence-based approach. *Medical Decision Making, 28,* 829–833.

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision Making, 7,* 332–359.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7,* 1–75. doi:10.1016/1041-6080(95)90031-4

Reyna, V. F., & Brainerd, C. J. (2011). Dual processes in decision making and developmental neuroscience: A fuzzy-trace model. *Developmental Review, 31,* 180–206.

Roscoe, R. D., & Chi, M. T. H. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science, 36,* 321–350.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22. doi:10.1037/0033-2909.119.1.3

Stefanek, M., Hartmann, L., & Nelson, W. (2001). Risk-reduction mastectomy: Clinical issues and research needs. *Journal of the National Cancer Institute, 93,* 1297–1306.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46,* 197–221.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31,* 3–62. doi:10.1080/03640210709336984

VanLehn, K., Jones, R. M., & Chi, M. T. H. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences, 2,* 1–59.

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91,* 301–311.

Wolfe, C. R. (2001). Plant a tree in cyberspace: Metaphor and analogy as design elements in Web-based learning environments. *Cyberpsychology & Behavior, 4,* 67–76.

Wolfe, C. R. (2011). Argumentation across the curriculum. *Written Communication, 28,* 193–219.

Wolfe, C. R., Britt, M. A., Petrovic, M., Albrecht, M., & Kopp, K. (2009). The efficacy of a Web-based counterargument tutor. *Behavior Research Methods, 41,* 691–698. doi:10.3758/BRM.41.3.691

Wolfe, C. R., Fisher, C. R., Reyna, V. F., & Hu, X. (2012). Improving internal consistency in conditional probability estimation with an Intelligent Tutoring System and Web-based tutorials. *International Journal of Internet Science, 7,* 38–54.

Wolfe, C. R., Reyna, V. F., Cedillos, E. M., Widmer, C. L., Fisher, C. R., & Brust-Renck P. G. (2012, October). *An Intelligent Tutoring System to help women decide about testing for genetic breast cancer risk*. Paper presented to the 34th Annual Meeting of the Society for Medical Decision Making, Phoenix, AZ.