NIH Public Access
Author Manuscript
Behav Res Methods. Author manuscript; available in PMC 2014 September 01.

Published in final edited form as:
Behav Res Methods. 2013 September ; 45(3): 623–636. doi:10.3758/s13428-013-0352-z.

The Development and Analysis of Tutorial Dialogues in AutoTutor Lite

Christopher R. Wolfe[1], **Colin L. Widmer**[1], **Valerie F. Reyna**[2], **Xiangen Hu**[3], **Elizabeth M. Cedillos**[1], **Christopher R. Fisher**[1], **Priscilla G. Brust-Renck**[2], **Triana C. Williams**[1], **Isabella Damas Vannucchi**[1], and **Audrey M. Weil**[1]

[1]Miami University, Oxford, Ohio, USA

[2]Cornell University, Ithaca, New York, USA

[3]The University of Memphis, Memphis, Tennessee, USA

Abstract

The goal of Intelligent Tutoring Systems (ITS) that interact in natural language is to emulate the benefits a well-trained human tutor provides to students, by interpreting student answers and appropriately responding 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 to improve performance. The goal of this method was to enable BRCA Gist to interpret and respond to answers in a manner that best facilitates 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 answers found by trained human judges using a reliable rubric. Dialogue quality between users and BRCA Gist, predicted performance on a breast cancer risk knowledge test completed after the tutor. The appropriateness of BRCA Gist feedback also predicted the quality of answers and breast cancer risk knowledge test scores.

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" (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2005; Bloom, 1984) though a recent review of the literature suggests 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 students 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 passively spending time with the material by reading or listening to lectures (Graesser, McNamara, & VanLehn, 2005). A current challenge is to create

---

Corresponding Author: Christopher R. Wolfe, Ph.D., Department of Psychology, Miami University, Oxford, Ohio USA 45056, (513) 529-5670, WolfeCR@MiamiOH.edu.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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 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, 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, (Graesser, Lu, Jackson, et al., 2004; Craig, Sullins, Witherspoon & Gholson, 2006) physics, (Jackson, Ventura, Chewle, Graesser, et al. 2008; VanLehn, Graesser, Jackson, Jordana, Olneyb, & Rose, 2006) 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, Hastings, & Allbritton, 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 et al., 2001; Graesser, Person, & Magliano, 1995). 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. 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 (Jackson et al., 2004; Arnott, Hastings, & Allbritton, 2008) consistently demonstrate that AutoTutor is effective in helping people learn. In ten controlled experiments with over 1000 participants, AutoTutor produced statistically significant gains of .2 to 1.5 standard deviations with a mean of .81 (Graesser et al., 2005).

The research on AutoTutor has extended to a web-based version called AutoTutor Lite (Hu, Han, & Cai, 2008; 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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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 based on expectations using hints and elaboration, use SAC 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, Wiemer-Hastings, Wiemer-Hastings, Harter, 2000; Landauer, Foltz, & Laham, 1998). LSA is a computational technique that mathematically measures the semantic similarity of sets of texts (Hu, Wiemer-Hastings, Graeser, & 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 based on 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 to encourage elaboration and other verbal responses based on 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, et al., 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 SKOs (Sharable Knowledge Objects). 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, Reyna, & Hu, 2012) that teaches women about genetic risk of breast cancer. Our goal was to create an ITS to engage women in a dialogue about the myriad of difficult issues associated with genetic testing for breast cancer risk (Armstrong, Eisen, & Weber, 2000; Berliner, Fay, et al., 2007; Chao, Studts, Abell, 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, and input from medical experts.

We developed BRCA Gist guided by Fuzzy-Trace Theory (FTT; Reyna, 2008a, 2008b, 2012; Reyna & Brainerd, 1995). FTT is a dual process theory (Reyna & Brainerd, 2011; Sloman, 1996) which 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 drilling 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.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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 Intelligent Tutoring Systems, primarily the work of Arthur Graesser with AutoTutor (Graesser, Chipman, Haynes, & Olney, 2005; Graesser, 2011). Three of these questions require participants to create a self-explanation about the material 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 their answers to a set of expectation texts.

A screen shot of BCA Gist from the learner’s perspective can be found in Figure 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 a participant has earned an overall CO score exceeding 0.4 (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 expectations 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 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 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 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

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

human free association (e.g., used by Nelson, McEvoy, and Schreiber, 2004) and college LSA (selected here). There are several domains to choose from including science and mathematics, computer and internet, health, and environment. We found 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 is compared to the expectations text. Weight criteria is the minimal weight associated terms in the expectation text must have to be included in the space. There is a tradeoff 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 expectations text and similar terms, such as synonyms, that is required 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 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 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 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 material.

The next step was the creation of a reliable rubric to judge answers to these questions for the content 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 either present or absent in an answer using a gist scoring approach (Britt, Kurby, Dandotkar, and 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 inter-rater reliability, we counted only instances where at least one rater judged an item as present. (Differences in raters' judgments of "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 both marked as absent were not counted (as this was the vast majority of items and would greatly inflate agreement ratings). Inter-Rater Reliability is thus the number of items the two raters agreed on (both marked as present) divided by the total number of items marked present by at least one rater.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

An important source of data we used to calibrate BRCA Gist were answers to the same questions written by 81 untrained undergraduate participants in order to get examples of the range of answer 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 .89 agreement on their judgments. Thus, the conditional probability that both judges marked an item present given that either one did so was .89.

The next step in the method was to develop the actual expectation text that BRCA Gist would use to evaluate participant 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 everything 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 expectations 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 expectations 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 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 expectations 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 total coverage and generally ranges from 0 to 1 with higher numbers corresponding to better coverage of the expectations text. We operationalized a good fit as a high CO score, indicating a high association between the input text and the expectation text.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

The first text we used was the exact same expectation text (see above) entered one sentence at a time. This process does not guarantee one-to-one matching of identical text. We reasoned that expectations 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 parameters for association strength and item weight, and made modifications to the expectations texts accordingly.

To illustrate, in the case of the expectations text above, we started with a version that was 100 words long, but the best final CO score we could obtain feeding the exact text expectations text back as input one sentence at a time was only 0.44. Moreover, the mean change in CO score from one turn to the next was only 0.03 with the highest being just 0.05. We reasoned that these scores would be inadequate for distinguishing appropriate responses from inappropriate responses (because they mean that the program does not even recognize inputs that exactly match expectations). AutoTutor Lite also produces scores for the extent to which a sentence of input is relevant, but has already been provided previously (described in more detail below). These scores for relevant old information ranged from 0.51 to 0.95 suggesting that the 100 word expectation text had many redundancies. The reduced expectations text of 75 words (above) initially produced a final CO score of 0.53, but by changing the settings to a weight of 0.3, an association strength of 0.2, a minimum item weight of 0.2, and a minimum rank (the product of weight and strength) of 0.2, we obtained a higher final CO score. Those settings produced a final CO score of 0.66 and a mean change of CO score per turn of 0.05. 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 domains combined produced a final CO score of 0.53 and the same procedure using human association norms instead of LSA also produced a final CO score of 0.53.

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 research assistants. These were entered to ensure 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 to make sure the tutor was also capable of appropriately discriminating irrelevant text that does not address its questions. This allows the tutor to give appropriate feedback to try to get participants who are 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 0.019.

We then developed reasonable answers to the questions using 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 exact same words that were contained in the expectation text. This was done to make sure the tutor had been constructed and calibrated to appropriately judge real semantic similarity rather than look for specific words. According to Fuzzy-Trace Theory, 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 participants might write, which may or may not include the words in the expectation text.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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 is capable of responding to actual answers that may be given by participants, and tests that the tutor can correctly differentiate between better and worse answers to the questions. To illustrate, among essays written in response to the positive test result question an essay that we identified as relatively good yielded a final CO score of .474, and an essay that we judged as not as good (but among the better responses) produced a final CO score of .397. One issue that we discovered using these texts is 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 gives in response to participants' statements. AutoTutor Lite generates a number of scores each time a sentence is entered to answer on 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 (± 0.05) values from 0 to 1 in increments of 0.1 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 Figure 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, information that has already been presented in a previous turn, and is not relevant to the expectations. As previously noted, "CO" stands for total 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 appropriate feedback to be given based on a participant's progress in answering a question. Examples of feedback rules we created for the interaction in which we asked participants what should someone do if she finds out that she has inherited an altered BRCA gene can be found in Figure 3.

AutoTutor Lite cannot detect anything more specific about the learner's input beyond the scores generated above. Thus, a CO score of 0.2 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 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 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 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 participants were considering. Our strategy was to always prompt participants who scored low on the very first turn, to let them know they were off track. Participants who

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

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 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 paper which focuses on the development and assessment of tutorial dialogues) compared the efficacy of BRCA Gist to reading text from the National Cancer Institute website, and an irrelevant nutrition control (Wolfe 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 minutes. 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 are to determine whether BRCA Gist’s assessment of the similarity of answers to the expectations is a reliable measure for the quality of those answers, whether the quality of answers predicts learning, and whether the success of the interactions has an effect on learning.

It is important to establish that BRCA Gist’s judgments about the semantic similarity of participant answers are actually capturing how much content is in those answers. That is, if the tutor is appropriately interpreting answers, an answer which is given a higher measure of semantic similarity should contain more relevant content than in an answer that is 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 if the CO scores accurately measure the amount of content in an answer, we compared BRCA Gist final CO scores to scores obtained applying our rubrics blind to CO score. To ensure 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 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 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 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 floor effect. For the untrained participants Cronbach’s Alpha was 0.67, and the mean was 57% correct. These scores are solid considering that there is a wide range of content ranging from biology to anti-discrimination laws that help women decide whether 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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

made by BRCA Gist. Judgments were made only in relation to the user's previous statement and not 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 inappropriate were the opposite: an inappropriate response discouraged elaboration, did not flow naturally from the previous input, and responded incorrectly to the accuracy of the participant's input. Neutral was used when it was difficult to classify it as clearly one or the other (such as if it encouraged elaboration and responded correctly, but didn't flow as naturally as most appropriate ones did). During the analysis phase, we collapsed to a distinction between appropriate and inappropriate, and neutrals 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 these judgments were reliable, two independent raters made judgments about one third of the responses. To calculate reliability, we looked at the proportion of responses the two judges agreed on over the total number of responses and found .86 agreement.

To illustrate, the following is part of a dialogue between BRCA Gist and a participant (spelling mistakes and typos included) where 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 encourages her to talk about first 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 score, are included in Appendix A.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Results

The goals of these analyses were to determine whether BRCA Gist's assessment of the similarity of answers to the expectations (final CO score) is a reliable measure for the quality of those answers, whether the quality of a participant's verbal statements predicts 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.

BRCA Gist participants produced answers with an average final CO score of 0.41 (SD = 0.14), creating answers using an average of 5.37 sentences (SD = 1.75). Rubric judgments determined that answers contained an average of about a quarter of the total possible content items for each answer (M = 0.25, SD = 0.10). This compares favorably to the previously obtained rubric scores from 81 untrained undergraduates who wrote brief essays without the benefit of any tutorial (M = 0.147, SD = 0.057). Because the testing conditions differ dramatically between these two samples, the use of statistics to make inferences about them is unwarranted. CO scores and rubric scores from the BRCA Gist tutorial dialogues were highly correlated, $r(62) = 0.75$, $p < 0.001$. Thus, the BRCA Gist semantic processing engine's final CO score accounted for over half 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 = 0.85, SD = 0.24) was highly correlated with both CO score, $r(62) = 0.82$, $p < 0.001$, and with rubric score, $r(62) = 0.74$, $p < 0.001$. This finding demonstrates that the greater the extent to which BRCA Gist responded appropriately to participant input, the greater the proportion of expectations were covered by participants' complete answers. The correlations between appropriateness of responses and the quality of participants' final answers remains significant even when controlling for number of sentences in their answers, for both CO score, $t(62) = 6.11$, $p < 0.001$, and rubric score, $t(62) = 4.55$, $p < 0.001$. This indicates that these correlations are more than a simple case of the tutor failing to respond appropriately to incomplete, short answers.

There was also a positive correlation between performance on the knowledge test (M = 0.74, SD = 0.15) and final CO score, $r(62) = 0.35$, $p = 0.004$, and performance on the knowledge test and rubric score, $r(62) = 0.67$, $p < 0.001$. The appropriateness of tutor responses was also correlated with performance on the knowledge test, $r(62), = 0.41$, $p < 0.001$. When the tutor responded appropriately, participants produced fuller answers reflecting greater knowledge of Breast Cancer and Genetic Risk. These answers within the tutorial dialogue predict subsequent knowledge as measured by the multiple choice test. Moreover, the mean score of 74% correct compares 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 provided evidence that the tutorial dialogues were suitable and beneficial to learning. One implication of these results is that the semantic similarity of answers as judged by the tutor seems to be capturing 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 demonstrates that an ITS that uses natural language processing to interpret answers, such as BRCA Gist, can

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

capture much of the content of users' statements 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 based on scores generated by a semantic processing engine that correspond to the expectations of a trained human researcher.

The correlation between BRCA Gist final CO scores and human judgments of r = .75 is comparable with 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 r = .48 (Magliano et al., 2011). AutoTutor's evaluation of whether or not students stated expectations is 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, & Boonthum, (2004) compared human judgments to those of iSTART and found that iSTART scores score successfully distinguished between paraphrases and explanations containing some kind of elaboration.

These results also provide evidence that the amount of relevant content 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, Leeuw, Chiu, & LaVancher, 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 (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 is 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 appropriately. 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 answer, rather than the quality of the answer only reflecting the knowledge 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 untrained participants in the test development study, and Wolfe, Reyna and colleagues (2012) report 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 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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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

The results of this analysis also prove useful for the future development of BRCA Gist. The knowledge we gained from this study will allow us to apply these empirical method to improve the tutor's assessments of participant's 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 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?" sounds reasonable to both a person who just entered a sentence about tumors, and a person who had not done so. Because subtle changes in BRCA Gist responses can yield noticeable differences in how much elaboration they produce, careful thought anticipating the sentences people may use to answer the tutor is required.

AutoTutor Lite does not really "understand" what users are saying the way a human tutor does, or even the way some ITS such as AutoTutor understand natural language. In developing BRCA Gist, our solution to this shortcoming is 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 uses 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) 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 and colleagues (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, Sullins, Witherspoon, and Gholson, 2006). Again, the BRCA Gist approach to tutorial dialogues appears well situated to produce deep-level learning; however, systematic experimentation is 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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

such as the NCI web site. 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The project described was supported by Award Number 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 wish to thank Dr. Nananda Col for her expert feedback on our research materials and BRCA Gist. The authors wish to thank Sharjeel Chaudry, Suveera Dang, Amrita Rao, Anna Catherine Morant, Kate Bassolino, Eric Cooke, Jessica Reigrut, and Mandy Withrow for assistance with data collection.

References

Armstrong K, Eisen A, Weber B. Assessing the risk of breast cancer. New England Journal of Medicine. 2000; 342:564–571. [PubMed: 10684916]

Arnott E, Hastings P, Allbritton D. Research Methods Tutor: Evaluation of a dialogue-based tutoring system in the classroom. Behavior Research Methods. 2008; 40:694–698. [PubMed: 18697663]

Azevedo R, Lajoie SP. The cognitive basis for the design of a mammography interpretation tutor. International Journal of Artificial Intelligence in Education. 1998; 9:32–44.

Berliner JL, Fay AM, et al. Risk assessment and genetic counseling for hereditary breast and ovarian cancer: Recommendations of the National Society of Genetic Counselors. Journal of Genetic Counseling. 2007; 16:241–260. [PubMed: 17508274]

Bloom BS. The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educational Research. 1984; 13(6):4–16.

Britt MA, Kurby CA, Dandotkar S, Wolfe CR. I Agreed with What? Memory for Simple Argument Claims. Discourse Processes. 2008; 45:52–84.

Chao C, Studts JL, Abell T, Hadley T, Roetzer L, Dineen S, Lorenz D, Agha YA, McMasters KM. Adjuvant chemotherapy for breast cancer: How presentation of recurrence risk influences decision-making. Journal of Clinical Oncology. 2003; 21(23):4299–4305. [PubMed: 14581440]

Chi MT. Self-explaining expository texts: The duel processes of generating inferences and repairing mental models. Advances in instructional psychology. 2000; 5:161–238.

Chi MTH, Bassok M, Lewis MW, Reimann P, Glaser R. Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science. 1989; 15:145–182.

Chi MTH, de Leeuw N, Chiu MH, LaVancher C. Eliciting Self-Explanations Improves Understanding. Cognitive Science. 1994; 18:439–477.

Chi MTH, Siler SA, Jeong H, Yamauchi T, Hausmann RG. Learning from human tutoring. Cognitive Science. 2005; 25:471–533.

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

Graesser AC. Learning, thinking, and emoting with discourse technologies. American Psychologist. 2011; 66(8):746–757. [PubMed: 22082403]

Graesser AC, Chipman P, Haynes BC, Olney A. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. IEEE Transactions on Education. 2005; 48(4):612–618.

Graesser AC, Lu SL, Jackson GT, Mitchell HH, Ventra M, Olney A, Louwerse MM. AutoTutor: A tutor with dialogue in natural language. Behavioral Research Methods, Instruments & Computers. 2004; 36(2):180–192.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Graesser A, McNamara D. Self-Regulated Learning in Learning Environments With Pedagogical Agents That Interact in Natural Language. Educational Psychologist. 2010; 45(4):234–244.

Graesser AC, McNamara DS, VanLehn K. Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. Educational Psychologist. 2005; 40(4):225–234.

Graesser AC, Person NK, Magliano JP. Collaborative dialogue patterns in naturalistic one-on-one tutoring. Applied Cognitive Psychology. 1995; 9:359–387.

Graesser AC, VanLehn K, Rose C, Jordan P, Harter D. Intelligent tutoring systems with conversational dialogue. AI Magazine. 2001; 22(4):39–51.

Graesser AC, Wiemer-Hastings P, Wiemer-Hastings K, Harter D. Tutoring Research Group & Person N. Using latent semantic analysis to evaluate the contributions of students in AutoTutor. Interactive Learning Environments. 2000; 8(2):129–147.

Hu, X.; Cai, Z.; Han, L.; Craig, SD.; Wang, T.; Graesser, AC. Proceedings of the 2009 conference of Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modeling. Amsterdam: IOS Press; 2009 Jul. AutoTutor Lite; p. 802-802.

Hu, X.; Cai, Z.; Wiemer-Hastings, P.; Graeser, AC.; McNamara, DS. Strengths, limitations, and extensions of LSA. In: McNamara, D.; Landauer, T.; Dennis, S.; Kintsch, W., editors. LSA: A Road to Meaning. Mahwah, NJ: Erlbaum; 2007. p. 401-425.

Hu, X.; Han, L.; Cai, Z. 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; 2008. https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxzY2lwd3N8Z3g6NmZhMTE2NWI3OGRiMzBlMw&pli=1

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

Jackson, GT.; Ventura, MJ.; Chewle, P.; Graesser, AC. the Tutoring Research Group. The Impact of Why/AutoTutor on learning and retention of conceptual physics. In: Lester, JC.; Vicari, RM.; Paraguacu, F., editors. Intelligent Tutoring Systems 2004. Berlin, Germany: Springer; 2004. p. 501-510.

Kopp KJ, Britt MA, Millis K, Graesser AC. Improving the efficiency of dialogue in tutoring. Learning and Instruction. 2012; 22(5):320–330.

Lairson DR, Chan W, Chang Y-C, Junco DJ, Vernon SW. Cost-Effectiveness of Targeted vs. Tailored Interventions to Promote Mammography Screening Among Women Military Veterans in the United States. Evaluation and Program Planning. 2011; 34:97–104.10.1016/j.evalprogplan.2010.07.003 [PubMed: 20810168]

Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse Processes. 1998; 25:259–284.

Magliano JP, Graesser AC. Computer-based assessment of student-constructed responses. Behavior Research Methods. 2012; 44:608–621. [PubMed: 22581494]

Magliano JP, Millis KK, Levinstein I, Boonthum C. the RSAT Development Team. Assessing comprehension during reading with the Reading Strategy Assessment Tool (RSAT). Metacognition and Learning. 2011; 6:131–154. [PubMed: 23901332]

Malatesta, K.; Wiemer-Hastings, P.; Robertson, J. Beyond the short answer question with research methods tutor. In: Cerri, SA.; Gouardères, G.; Paraguaçu, F., editors. Intelligent Tutoring Systems 6th International Conference LNCS 2363. Berlin/Heidelberg: Springer; 2002. p. 562-573.

McNamara DS, Levinstein IB, Boonthum C. iSTART: Interactive strategy training for active reading and thinking. Behavior Research Methods, Instruments, & Computers. 2004; 36:222–233.

Nelson DE, McEvoy CL, Schreiber TA. The University of South Florida free association, rhyme, and word fragment norms. Behavior Research Methods, Instruments, & Computers. 2004; 36:402–407.

Reyna VF. A theory of medical decision making and health: Fuzzy trace theory. Medical Decision Making. 2008a; 28(6):850–865. [PubMed: 19015287]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Reyna VF. Theories of medical decision making and health: An evidence-based approach. Medical Decision Making. 2008b; 28:829–833. [PubMed: 19020341]

Reyna VF. A new intuitionism: Meaning, memory, and development in Fuzzy-Trace Theory. Judgment and Decision Making. 2012; 7:332–359.

Reyna VF, Brainerd CJ. Dual processes in decision making and developmental neuroscience: A fuzzy-trace model. Developmental Review. 2011; 31:180–206. [PubMed: 22096268]

Reyna VF, Brainerd CJ. Fuzzy-trace theory: An interim synthesis. Learning and Individual Differences. 1995; 7(1):1–75.

Sloman SA. The empirical case for two systems of reasoning. Psychological Bulletin. 1996; 119:3– 22.

Stefanek M, Hartmann L, Nelson W. Risk-reduction mastectomy: Clinical issues and research needs. Journal of the National Cancer Institute. 2001; 93:1297–306. [PubMed: 11535704]

Roscoe RD, Chi MTH. Tutor learning: the role of explaining and responding to questions. Instructional Science. 2008; 36(4):321–350.

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

VanLehn K, Graesser AC, Jackson GT, Jordan P, Olney A, Rose CP. When are tutorial dialogues more effective than reading? Cognitive Science. 2007; 31:3–62. [PubMed: 21635287]

VanLehn K, Jones RM, Chi MTH. A model of the self-explanation effect. Journal of the Learning Sciences. 1992; 2:1–59.

Wiley J, Voss JF. Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. Journal of Educational Psychology. 1999; 91(2):301–311.

Wolfe CR. Plant a tree in cyberspace: Metaphor and analogy as design elements in Web-based learning environments. CyberPsychology & Behavior. 2001; 4:67–76. [PubMed: 11709909]

Wolfe CR. Argumentation across the curriculum. Written Communication. 2011; 28(2):193–219.

Wolfe CR, Britt MA, Petrovic M, Albrecht M, Kopp K. The efficacy of a web-based counterargument tutor. Behavior Research Methods. 2009; 41:691–698. [PubMed: 19587180]

Wolfe CR, Fisher CR, Reyna VF, Hu X. Improving internal consistency in conditional probability estimation with an Intelligent Tutoring System and web-based tutorials. International Journal of Internet Science. 2012; 7:38–54.

Wolfe, CR.; Reyna, VF.; Cedillos, EM.; Widmer, CL.; Fisher, CR.; Brust-Renck, PG. 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. 2012 Oct.

Appendix A. Examples of Successful and Unsuccessful Dialogues Between BRCA Gist and Research Participants

Note: Typos included from original dialogue.

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 = 0.149)

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 = 0.213)

BRCA Gist Response 2: Good. Can you say more about active surveillance? (Appropriate)

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

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 = 0.319)

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 = 0.396)

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 = 0.429)

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 = 0.438)

Turn 7: Lastly, there is a drug that can be taken which greatly reduces the chances of developing breast cancer (CO = 0.448)

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 = 0.460)

Participant 268 with a Poor Interaction and Low CO Score

Turn 1: She shoud do a survillence (CO = 0.010)

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 = 0.131)

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 = 0.181)

Turn 4: In catching the cacner earlier you are at a higher risk for survival (CO = 0.220)

R4: Try to add something about surgery. (Appropriate)

Turn 5: Then you are able to have surgery to take away the cancer (CO = 0.238)

Turn 6: When you have surgery it will get rid of the cancer cells (CO = 0.250)

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 = 0.250)

Turn 8: Having surgery at a earlier point in the cancer can get rid of the majority of the cancer before it divides (CO = 0.285)

BRCA Gist Response 8: Try to add something about active surveillance, drugs, or surgery. (Appropriate)

NIH-PA Author Manuscript

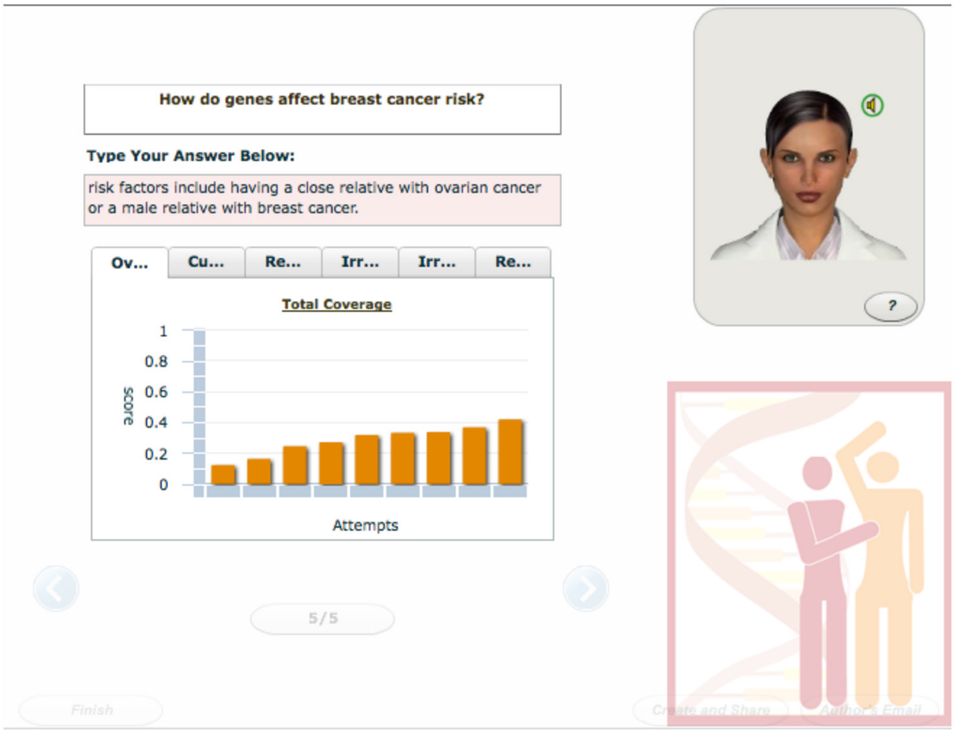
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript



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.

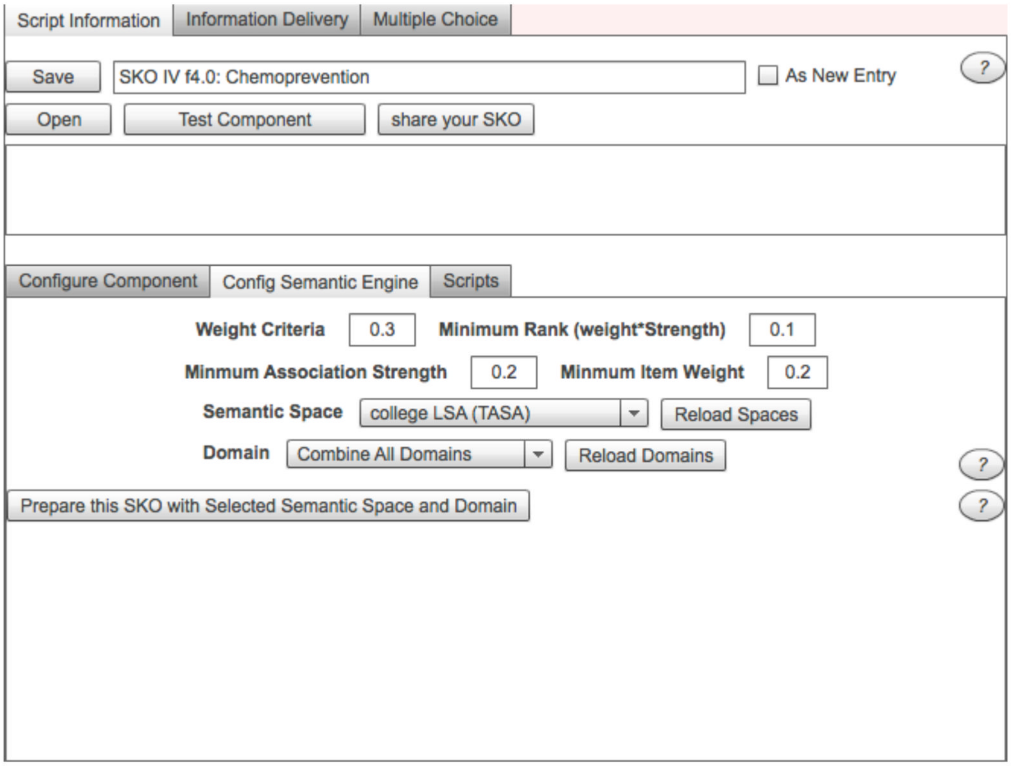
Figure 1.

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

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript



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

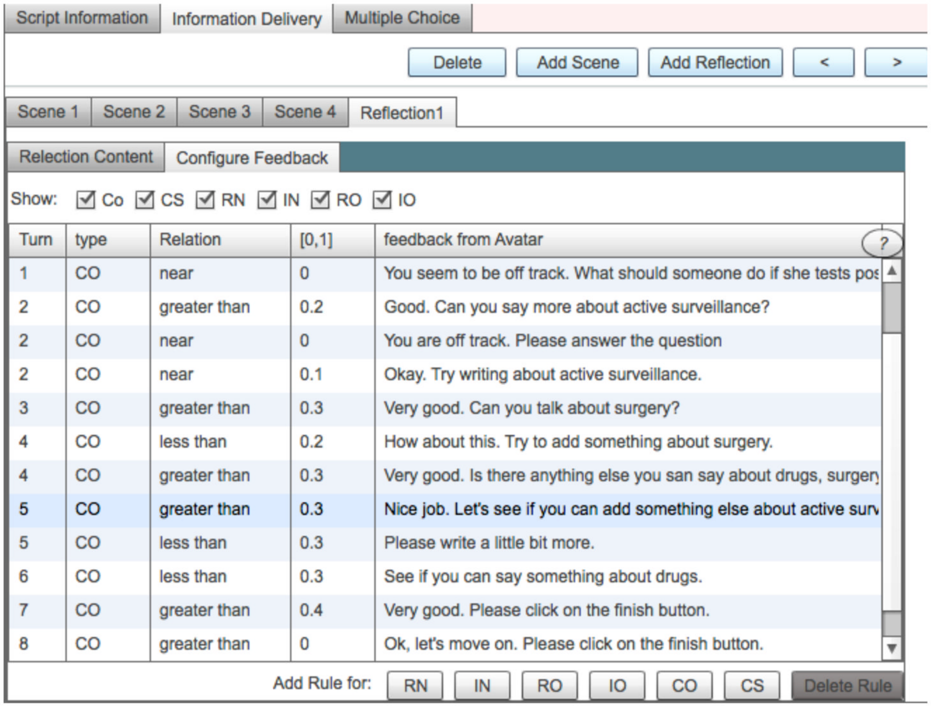
Semantic Space college LSA (TASA)

Domain Combine All Domains

Figure 2.

AutoTutor Lite authoring tools for configuring the semantic engine.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript



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.

Figure 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).