Behav Res (2015) 47:632–648
DOI 10.3758/s13428-015-0592-1

Tutorial dialogues and gist explanations of genetic breast cancer risk

Colin L. Widmer[1,3] · Christopher R. Wolfe[1] · Valerie F. Reyna[2] · Elizabeth M. Cedillos-Whynott[1] · Priscila G. Brust-Renck[2] · Audrey M. Weil[1]

Published online: 29 April 2015

Abstract The intelligent tutoring system (ITS) *BRCA Gist* is a Web-based tutor developed using the Shareable Knowledge Objects (SKO) platform that uses latent semantic analysis to engage women in natural-language dialogues to teach about breast cancer risk. BRCA Gist appears to be the first ITS designed to assist patients' health decision making. Two studies provide fine-grained analyses of the verbal interactions between BRCA Gist and women responding to five questions pertaining to breast cancer and genetic risk. We examined how "gist explanations" generated by participants during natural-language dialogues related to outcomes. Using reliable rubrics, scripts of the participants' verbal interactions with BRCA Gist were rated for content and for the appropriateness of the tutor's responses. Human researchers' scores for the content covered by the participants were strongly correlated with the coverage scores generated by BRCA Gist, indicating that BRCA Gist accurately assesses the extent to which people respond appropriately. In Study 1, participants' performance during the dialogues was consistently associated with learning outcomes about breast cancer risk. Study 2 was a field study with a more diverse population. Participants with an undergraduate degree or less education who were randomly assigned to BRCA Gist scored higher on tests of knowledge than those assigned to the National Cancer Institute website or than a control group. We replicated findings that the more expected content that participants included in their gist explanations, the better they performed on outcome measures. As fuzzy-trace theory suggests, encouraging people to develop and elaborate upon gist explanations appears to improve learning, comprehension, and decision making.

Keywords Intelligent tutoring system · Risk perception · Risk communication · Discourse technology · Fuzzy-trace theory

Decisions in the medical domain are some of the most important that people face. Patients must make decisions about whether to receive medical tests and about which courses of treatment to pursue. In an ideal situation patients should engage in shared decision making with doctors and other healthcare providers to collectively decide the course of action that best suites the patients' specific needs and situations (Col et al., 2011). However, people generally do not have the medical training or knowledge to accurately assess their situation and appropriately use their understanding to make sound medical decisions. Thus, effective and accessible patient education are needed to provide people facing medical decisions with the knowledge that they require. One medical domain where there is a great need for patient education is breast cancer. Breast cancer is a serious issue for many women; each year more than 170,000 cases are diagnosed in the United States (Armstrong, Eisen, & Weber, 2000). Genetic tests exist that can provide women with information about whether they have an increased risk of developing breast cancer due to gene mutations. However, genetic testing is not appropriate for all women. Many difficult issues are relevant to the decision of whether to receive genetic testing for breast cancer (Berliner & Fay, 2007; Nelson, Huffman, Fu, & Harris, 2005; Wolfe et al., 2014). Our strategy to provide women with the education they need to understand the decision to be tested for genetic risk of breast cancer has been to develop BRCA Gist

✉ Colin L. Widmer
widmercl@miamioh.edu

[1] Miami University, Oxford, OH, USA

[2] Cornell University, Ithaca, NY, USA

[3] Department of Psychology, Miami University, Oxford, OH 45056, USA

("Breast Cancer and Genetics Intelligent Semantics Tutor"), an intelligent tutoring system (ITS) that teaches women about breast cancer risk and interacts with them to generate gist explanations through natural-language dialogues.

Tutoring and self-explanation

There is ample evidence that both human and computer tutors provide impressive benefits to learning in academic domains. For example, learners who study with a tutor consistently show marked learning gains over those learning in traditional classrooms (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Roscoe & Chi, 2008). A major source of the benefit of tutoring is that tutors encourage learners to generate self-explanations. Learners who generate self-explanations show greater and deeper understanding of learned material (Chi, Leeuw, Chiu, & LaVancher, 1994; Magliano, Trabasso, & Graesser, 1999). Learners' initial responses to questions about learned material are often brief and unsophisticated, but with training (McNamara, 2004) or the support of a tutor (Chi, 1996) learners can generate more elaborate self-explanations that allow them to construct a more complete understanding of the content by connecting new insights with prior knowledge and uncovering any conflicting misconceptions in their understanding (Chi, 2000). Tutors encourage the creation of self-explanations by offering feedback and asking questions that lead learners to build on their answer. This interaction between the tutor and the learner that leads to self-explanation appears to be essential for the full learning benefits of tutoring—the benefits cannot be solely attributed to either the tutor or the learner (Graesser, 2011). One account of this phenomenon holds that the benefit of tutoring arises only when student-constructed explanations are supported by tutor scaffolding (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001). Although the necessity of this scaffolding is not universally supported in the literature, generating such self-explanations has been consistently associated with strong learning gains, and can account for much of the success of tutors (Roscoe & Chi, 2008).

Computer-based tutoring systems that provide the same kind of interactions have been in use now for decades (Graesser, 2011). An ITS is a computer system designed to emulate the experience of a human tutor for learners (Ohlsson, 1986). The most successful ITS are those that are able to elicit self-explanations from learners by providing similar feedback and encouragement that a human tutor would give. ITS that capitalize on the learning benefits of self-explanation have shown great potential to increase learning (Aleven & Koedinger, 2002; Graesser, Lu, et al., 2004), and can be nearly as effective at achieving learning gains as human tutors (VanLehn, 2011). The greatest learning is produced by ITS that enable students to generate and elaborate self-explanations by interacting in natural-language dialogues (Graesser, McNamara, & VanLehn, 2005). ITS that use natural-language dialogues are able to communicate with students in a natural-language (such as English). This is accomplished by comparing learner input to expectations provided to the tutor and using discourse processing techniques to assess what the learner has said and determine appropriate feedback (Graesser et al., 2000). One such discourse processing technique is latent semantic analysis (LSA; Landauer, Foltz, & Laham, 1998). LSA is a computational method that measures the semantic similarity of sets of texts (Hu, Cai, Wiemer-Hastings, Graesser, & McNamara, 2007). LSA uses a large corpus of text to create a semantic space that represents the semantic relatedness of words based on co-occurrence in the corpus (Landauer & Dumais, 1997). ITS that use LSA and similar techniques to process discourse use these mathematical measures of semantic relatedness to determine how similar learner input is to the tutor's expectations (Graesser & McNamara, 2010). The tutor can then use this measure of student progress to select the next dialogue move that will most promote learning and encourage self-explanation based on the pedagogically informed strategies that guided the tutor's development (Anderson, Corbett, Koedinger, & Pelletier, 1995; Chi, VanLehn, Litman, & Jordan, 2011).

Two platforms used to build ITS that have successfully utilized natural-language dialogue interactions are the Web-based Shareable Knowledge Objects (SKO—formerly called AutoTutor Lite; Hu et al., 2009; Hu, Han, & Cai, 2008) that our tutor BRCA Gist is built on, and its inspiration, AutoTutor (Graesser, Chipman, Haynes, & Olney, 2005; Graesser, Lu, et al., 2004). SKO and AutoTutor draw heavily on tutoring research about self-explanation and elaboration and has benefitted from continuous research and development for over twenty years. AutoTutor has been shown to be successful in multiple academic domains, such as physics and computer science (Craig, Sullins, Witherspoon, & Gholson, 2006; Gholson et al., 2009; Jackson, Ventura, Chewle, Graesser, & the Tutoring Research Group, 2004). AutoTutor's successful natural-language interaction style of *expectation and misconception tailored dialogue* (Graesser, Chipman, et al., 2005; Graesser, Person, & Magliano, 1995) uses scripts that include a set of expectations as well as misconceptions learners are likely to entertain. AutoTutor uses LSA to track progress the learner makes toward each expectation and provides appropriate hints and prompts to guide a learner toward more elaborate self-explanations of learned material, while also identifying and offering feedback to address any misconceptions found in learner input. AutoTutor also utilizes avatars (animated faces or figures) that present the tutor's dialogue, providing a conversational agent that permits additional means of communication (such as expressions and gestures) to improve the interaction (Graesser, 2011; Graesser, Jeon, & Dufty, 2008).

SKO takes many of the most important features of AutoTutor and is optimized to run online using an ordinary

Web browser. SKO is among the first ITS capable of running on the Web, and is apparently the first Web-based ITS platform to permit interactions through natural-language dialogues. SKO currently does not have all of the features that AutoTutor is capable of (e.g., SKO is currently only able to match learner input to expectations and cannot yet track learner misconceptions), but SKO is under continuous development and consistently adds new features. The ability to run on a Web browser from any computer with an internet connection opens the door for a more widespread use of ITS for real-world interventions.

Gist explanation

Most of the work on the benefits of self-explanation and elaboration in human tutoring and ITS such as AutoTutor and SKO has been done in academic domains such as solving physics problems (Chi, VanLehn, Litman, & Jordan, 2011; VanLehn, Jones, & Chi, 1992; VanLehn et al., 2007). Physics problem solving requires deep learning and sophisticated skills, but academic learning is fundamentally different from the process of making decisions—particularly in high-stakes medical decision making. Thus, an important part of our task is determining which aspects of tutoring and self-explanation most apply to patients' medical decision making in domains such as breast cancer risk, and then refining these aspects in an ITS that can assist people in making sound health decisions. The literature suggests ITS that teach academic skills typically guide a learner to meet specific, detailed expectations and involve lengthy interactions that can last up to 200 conversational turns of learner input and tutor feedback (Graesser, Chipman, et al., 2005). Following fuzzy-trace theory (FTT; Reyna, 2008), we argue that an ITS designed to teach decision relevant material in a health context may warrant briefer interactions that focus on the bottom-line gist of decision relevant information rather than on specific verbatim details.

We introduce the concept of *gist explanations* to refer to these self-explanations that are briefer and focus primarily on the bottom-line information that is most consequential for decision making as distinct from the longer and more intricate self-explanations used in academic domains. The concept of gist explanations stems from FTT (Reyna, 2008). FTT is a dual-process model of memory and decision making (Reyna, 2004, 2012; Reyna & Brainerd, 2011) that holds that people form multiple representations of information along a continuum. This continuum ranges from verbatim representations, which contain a large amount of specific detail, to gist representations, which lack the detail of verbatim representations but capture the important bottom-line meaning of the information. An important feature of FTT is that people tend to rely on the fuzzier gist representations, and these gist representations are more helpful for people to use when making decisions than detail oriented verbatim representations (Reyna, 2008, 2012). This preference for using bottom-line gist information to make decisions has implications for systems designed to teach people information that is relevant to real-life decision making.

Gist explanations make use of many of the same features as traditional self-explanations to promote learning. In generating the explanations learners are able to synthesize tutorial content and connect it with their own prior knowledge, leading to a deeper understanding of the material. However, gist explanations also differ from traditional self-explanations on a number of points. The focus of gist explanations is the essential bottom-line meaning of information rather than individual details. In a decision making context it is more important for a learner to achieve a succinct, meaningful, and accurate gist representation of decision relevant information than to get each individual verbatim detail correct. In gist explanations the key is for people to include the decision relevant takeaway information in their gist explanations. This should manifest as a learner being able to explain this gist meaning in their own words, rather than recalling specific information such as precise numeric details. A learner's gist explanation should focus on what is ultimately meaningful about the information, particularly aspects that directly relate to decision options. An ITS can support a learner in creating a gist explanation by offering feedback that encourages the learner to use his or her own words and guides the learner to focus primarily on information that is decision relevant and information about the consequences of decisions. We predict that generating gist explanations will not only increase learners' knowledge of tutor content as self-explanation research has suggested, but also specifically promote understanding the appropriate gist of the content, as well as enabling learners to appropriately apply that understanding to make sound decisions.

Because gist explanations focus only on the bottom-line meaning information, a gist explanation in an ITS should require many fewer turns of dialogue than the traditional self-explanations in ITS that teach academic domains and guide learners through all of the precise details. However, it important to stress that a gist explanation is not shorter than traditional self-explanations simply because less content is included. Gist explanations are not the same as text summaries. Rather, gist explanations are shorter because they focus on only the most consequential bottom-line decision relevant information. The details that are peripheral to this major takeaway message are excluded, somewhat akin to how a well-written abstract should convey the essence of a journal article without including all the specific details that make up the rest of the article. Gist explanations should focus only on information that is relevant to the decision that must be made ("should I get tested for genetic risk of breast cancer?") and the consequences of that decision ("what happens if I do get tested?"). In addition to potentially being more effective at assisting

decision making, this reduction in length also offers practical benefits for systems that can be implemented as actual decision aides in actual medical settings. Practical interventions require a greater degree of efficiency in order to be deployed effectively, and there is evidence that well designed ITS that make efficient use of dialogue in briefer interactions can effectively teach learners (Kopp, Britt, Millis, & Graesser, 2012). An ITS informed by FTT that utilizes gist explanations would engage with learners in natural-language dialogue interactions that encourage the learner to synthesize bottom-line gist information and put that information into the learner's own words.

We expect engaging in gist explanation dialogues in an ITS will be particularly helpful for assisting women learn about breast cancer risk and improving their ability to make informed decisions about her risk. There is evidence that focusing on bottom-line gist information leads to better health decision making than focusing on precise details (e.g., Reyna et al., 2011; Reyna & Lloyd, 2006), which has been tested in randomized experiments training gist thinking in health decisions (e.g., Reyna & Mills, 2014), as well as patient medication decisions (e.g., Fraenkel et al., 2012). Teaching precise information can be less effective because people can get all the details right but still fail to comprehend the actual gist meaning that would lead to more informed decisions. Standard, detailed medical reports about risk of breast cancer do not adequately convey the important decision relevant information about the meaning of risk to patients, whereas less complex, gist-based reports can more effectively convey such information, suggesting that facilitating appropriate gist representations of information is more useful in the context of understanding risk of breast cancer (Brewer, Richman, DeFrank, Reyna, & Carey, 2012). Additionally, we have found that a preference for integrating qualitative gist information and quantitative base rates assessed by the fuzzy processing preference index (Wolfe & Fisher, 2013) predicts accuracy at estimating a woman's degree of genetic breast cancer risk (Weil et al., 2015). Another indication that generating gist explanations about breast cancer risk may promote learning are results from VanLehn et al. (2007), who found that generating self-explanations are most beneficial for learners who are novices learning material designed for intermediate learners. The relevant knowledge for understanding genetic risk of breast cancer is relatively sophisticated, as compared to the average person's level of knowledge.

BRCA gist: An intelligent tutor with gist explanations

BRCA Gist appears to be the first ITS developed as a decision aide in the domain of patient decision making (Azevedo & Lajoie, 1998, created a prototype tutor for training radiology residents in diagnosing breast disease). Medical decision-making interventions informed by theories of decision making offer great promise of success (Reyna, Nelson, Han, & Pignone, 2015). To that end, BRCA Gist's development was guided by FTT (Wolfe et al., 2014; Wolfe et al., 2013).

The BRCA Gist tutor consists of a total of approximately 90 min of didactic content and gist explanation dialogue interactions. The didactic content of the tutor covers information about the formation and spread of breast cancer, genetic and other risk factors, genetic testing, and what women should do in the event of a positive or negative test result. The content was adapted from the National Cancer Institute website and was vetted by a medical expert. BRCA Gist includes five gist explanation dialogues. The interface during these gist explanation dialogues can be seen in Fig. 1. Learners are asked a question and enter their answers in the text box. After each learner turn of one sentence, BRCA Gist delivers feedback through an animated avatar to encourage the creation of a thorough gist explanation. Typical gist explanation dialogues in BRCA Gist last approximately six to nine turns of dialogue. As learners continue to build their gist explanation, BRCA Gist indicates progress toward a more complete gist explanation via a bar graph. This bar graph is generated on the basis of the semantic similarity of learner input and expectation texts. The expectation texts consist of approximately 60 to 80 words that best reflect the bottom-line meaning of a good answer to each question (Wolfe et al., 2013). BRCA Gist determines the semantic similarity by using LSA and the semantic space provided in the SKO platform to compare the sentences learners input in their gist explanations and these expectation texts. BRCA Gist is also able to use the semantic similarity of learners' answers and the expectation texts to determine appropriate feedback. At each dialogue turn BRCA Gist compares learners' answers to the expectation texts and selects feedback from a preset list using rules it was provided with during development. Thus, if the tutor detects learners are making good progress by entering input that increases the semantic similarity of their answer and the expectation text it can provide positive feedback ("Good job! Can you say more?"), and if the tutor detects learners are not making progress and are entering input that does not increase the semantic similarity of their answer and the expectation text it can provide feedback to direct learners to include relevant content ("You seem to be off track. What can you add about how genes affect breast cancer risk?"). A more detailed discussion of the iterative process we developed to create and improve the expectation texts and feedback of BRCA Gist's dialogues can be found in Wolfe et al. (2013).

The five gist explanation dialogues cover the important takeaway information about genetic risk of breast cancer and focus on decision relevant material, and are reflective of questions considered important by genetic counselors (Berliner & Fay, 2007). The five questions used in the gist explanation dialogues are "What is breast cancer?," "How does breast cancer spread?," "How do genes affect breast cancer risk?,"

Fig. 1 BRCA Gist interface during the gist explanation dialogues

"What should someone do with a positive test result?," and "What should someone do with a negative test result?" BRCA Gist also includes two argument-based dialogues ("What is the case for genetic testing?" and "What is the case against genetic testing?") that differ from the gist explanations that are the focus of this article. An analysis of the argumentation component of BRCA Gist and the use of argumentation in ITS can be found in Cedillos-Whynott, Wolfe, Widmer, Brust-Renck, and Reyna (2015).

Previous research on BRCA Gist has revealed participants randomly assigned to BRCA Gist consistently outperform participants assigned to the same content in static form and a control group that receives a comparable tutor on irrelevant content (Wolfe et al., 2015; Wolfe et al., 2014; Wolfe et al., 2013) on measures of knowledge of genetic risk of breast cancer. The most recent study of BRCA Gist examined the locus of the tutor's efficacy and found that the gist explanations account for a substantial part of the tutor's success. A version of the tutor including gist explanations resulted in a significant increase in participants' knowledge on several outcome measures, as compared to an otherwise identical version that excluded the gist explanation dialogues (Wolfe et al., 2015). In the two studies presented here, we provide fine-grained analyses of verbal interactions between BRCA Gist and adult women in response to five questions requiring gist explanations pertaining to breast cancer and genetic risk, testing the hypothesis that gist explanations lead to improved learning outcomes.

Study 1: University laboratory study

The analyses presented in Study 1 focus on the specific effects of the gist explanation dialogues with the BRCA Gist tutor under laboratory conditions. The gist explanation dialogues analyzed in Study 1 were embedded in a larger randomized, controlled study on the effectiveness of BRCA Gist, much of which has been presented by Wolfe et al. (2014) and is outside the scope of this article. Wolfe et al. (2014) found that BRCA Gist is more effective at teaching women about genetic testing and breast cancer risk than learning the same information presented in static form from the National Cancer Institute website or a control group. The focus of this article is to provide a fine-grained analysis of the gist explanation dialogues in the BRCA Gist condition.

Method

Participants in Study 1 were adult undergraduate women who had not themselves had breast cancer gathered from two sources: a university in the Eastern United States and a university in the Midwestern United States. Participants in both laboratories received course credit in return completing the study. Of 59 women in the study randomly assigned to the BRCA Gist condition, 17 were excluded from analyses here due to technical problems resulting in unsaved or unidentifiable verbal dialogue data, leaving 42 participants with dialogues for analysis.

Participants in Study 1 completed the BRCA Gist tutor (including the five gist explanation dialogues), and then completed several tasks intended to evaluate knowledge about genetic risk of breast cancer. These dependent measures include a 52-item multiple-choice test on declarative knowledge, a 30-item gist comprehension task, and a 12-item risk categorization task (Wolfe et al., 2014). The multiple-choice declarative knowledge test consists of original questions about the factual knowledge presented in the BRCA Gist tutor about

breast cancer, genetic risk, and genetic testing. The gist comprehension task captures a participant's gist understanding of the important bottom-line meaning information of breast cancer and genetic testing. For example, one of the important takeaway pieces of information about genetic risk of breast cancer is that only a very small amount of the cases of breast cancer are caused by genetic mutations, but if a woman does have a genetic mutation present her own risk of developing breast cancer is high. The gist comprehension task is designed to assess if participants understood such bottom-line gist meanings. Participants answer these gist comprehension questions using a 1- to 7-point Likert scale (from *strongly disagree* to *strongly agree*). However, although these items are presented with a Likert scale, they are not opinion-based—all of the items have independently verifiable correct answers (the correct gist meaning). Thus, a participant who shows low agreement to the statement "The greatest danger of dying from breast cancer is when a tumor grows larger in the location where it started" can be said to possess a stronger gist understanding about breast cancer risk than one who shows higher agreement because this is not the correct gist meaning conveyed in the tutorial. The risk categorization task consists of 12 scenarios profiling women with varying degrees of genetic risk of breast cancer. The Pedigree Assessment Tool (PAT; Hoskins, Zwaagstra, & Ranz, 2006) was used to create scenarios representing women of high risk, medium risk, and no risk (meaning no increase beyond the base rate). Participants read each scenario and then had to categorize each woman on the basis of which risk category she belongs in. These reliable instruments were vetted by a medical expert for accuracy and appropriateness. See Wolfe et al. (2014) for further discussion of these instruments.

The gist explanation dialogues of BRCA Gist were assessed both by researcher and computer measures using a reliable method developed in Wolfe et al. (2013). Researchers used a rubric to score the content included by participants in the gist explanation dialogues. Rubrics were developed for each of the five dialogue questions, and each rubric contained between 13 and 18 items that reflected the possible relevant pieces of information that a participant could include in a good complete answer (see Appendix A for the complete rubrics). Each item was scored as present or absent in each participant's answer, and the rubrics were used to score for the gist of each item. That is, participants did not need to match precise wording in their answer for a rubric item to be marked present, but had to show the bottom-line gist meaning of an item. The participant input and tutor feedback from each gist explanation dialogue were extracted from log files saved by SKO. Two independent raters used the rubric to score one third of the gist explanation dialogues together blind to outcomes and had an interrater reliability of .87. Raters also used a rubric to evaluate the feedback provided by BRCA Gist to participants. Each piece of feedback was classified as appropriate, inappropriate, or neutral using a gist scoring procedure. In order for feedback to be classified as appropriate it needed to correctly respond to the accuracy of a participant's input, encourage elaboration, and flow naturally from the previous verbal statement. Inappropriate feedback had the opposite criteria. An inappropriate response incorrectly responded to the accuracy of input, discouraged elaboration, and did not flow naturally from the previous input. The neutral classification was used when feedback was not clearly appropriate or inappropriate, such as if feedback encouraged elaboration but did not flow as naturally as most responses. Neutral classifications were rare (less than 1 % of feedback was classified as neutral) and were collapsed into the inappropriate category during analysis. See Appendix B for the rubrics used to make appropriateness judgments. All judgments about appropriateness were made in the context of the participant's previous input, and not in the context of the entire dialogue. Two independent raters used the rubric to score one third of the tutor feedback together blind to outcomes and had an interrater reliability of .93.

The gist explanation dialogues were also assessed by the BRCA Gist ITS itself using measures built into the SKO platform. SKO tracks student progress toward it expectations by using LSA to compare the semantic similarity of input to those expectations. This results in coverage scores (CO scores) that represent the progress a participant makes toward a complete answer as measured by the semantic similarity of the expectation text and all of the sentences entered as input up to that point. CO scores in BRCA Gist range from 0 to less than 1, with higher scores representing an answer that is more semantically related to the expectation text. BRCA Gist uses these CO scores to monitor learner progress and determine which preprogrammed feedback to deliver at a given turn of dialogue. These CO scores can also be used as a representation of the amount of content a learner includes in an answer to one of the gist explanation dialogues as measured by BRCA Gist. CO scores were recorded and extracted from log files for comparison with the researcher rated rubric scores as well as outcome measures.

Results

Participants produced gist explanations using an average of 5.8 sentences (SD = 2.0). The researcher rubric judgments determined that participants' answers covered an average of more than a quarter of all rubric items (M = .29, SD = .12). This is comparable to the results found in Wolfe et al. (2013), evaluating the first version of BRCA Gist, in which participants' answers covered a similar proportion of rubric items (M = .25, SD = .10) using an average of 5.4 sentences (SD = 1.8). Twenty-four participants ended one or more of the dialogues without even beginning their gist explanations. Participants who wrote at least one sentence in their gist explanation

produced answers of an average of 7.0 sentences (*SD* = 0.75) that covered a greater proportion of rubric items (*M* = .37, *SD* = .10). This translates to participants covering on average five or six rubric items in their gist explanations, indicating that they produced gist explanations that contained a high proportion of good decision relevant information. Some rubric items were included in a greater proportion of participants' gist explanations than others. For example, nearly all participants included Rubric Item 5 "When BRCA1 and BRCA2 genes are harmfully mutated, the risk of developing breast cancer increases" in their gist explanations addressing "How do genes affect breast cancer risk?" (Dialogue 3). In contrast, no participants included Rubric Item 12 "Mutations in other genes are also associated with the development of breast cancer. . . ." Figure 2 displays the proportions of participant answers of at least one sentence that included each rubric item (the full rubrics with the content covered by each item can be found in Appendix A).

Judgments about the appropriateness of tutor feedback determined that the BRCA Gist tutor responded appropriately 94 % of the time across all gist explanation dialogues. The appropriateness of tutor feedback varied slightly between dialogues. The tutor responded appropriately 87 % of the time in Dialogue 1 ("What is breast cancer?"), 94 % of the time in Dialogue 2 ("How does breast cancer spread?"), 95 % of the time in Dialogue 3 ("How do genes affect breast cancer risk?"), 96 % of the time in Dialogue 4 ("What should someone do with a positive test result?"), and 89 % of the time in Dialogue 5 ("What should someone do with a negative test result?"). In the vast majority of cases for each of the dialogues the tutor feedback correctly recognized the accuracy of participant input, encouraged participants to elaborate their answers, and flowed naturally from the prior input. These results represent an improvement over the first iteration of the BRCA Gist tutor that responded appropriately to participant input 85 % of the time (Wolfe et al., 2013).

BRCA Gist determined that participants produced gist explanations with an average CO score of .48 (*SD* = .16) across all five gist explanation dialogues. Average CO score varied between the dialogues: Dialogue 1 had an average CO score of .75 (*SD* = .24), Dialogue 2 had an average CO score of .51 (*SD* = .21), Dialogue 3 had an average CO score of .41 (*SD* = .17), Dialogue 4 had an average CO score of .45 (*SD* = .25), and Dialogue 5 had an average CO score of .31 (*SD* = .18).

The researcher rubric scores were highly correlated with BRCA Gist's internal CO scores. This was true with all gist

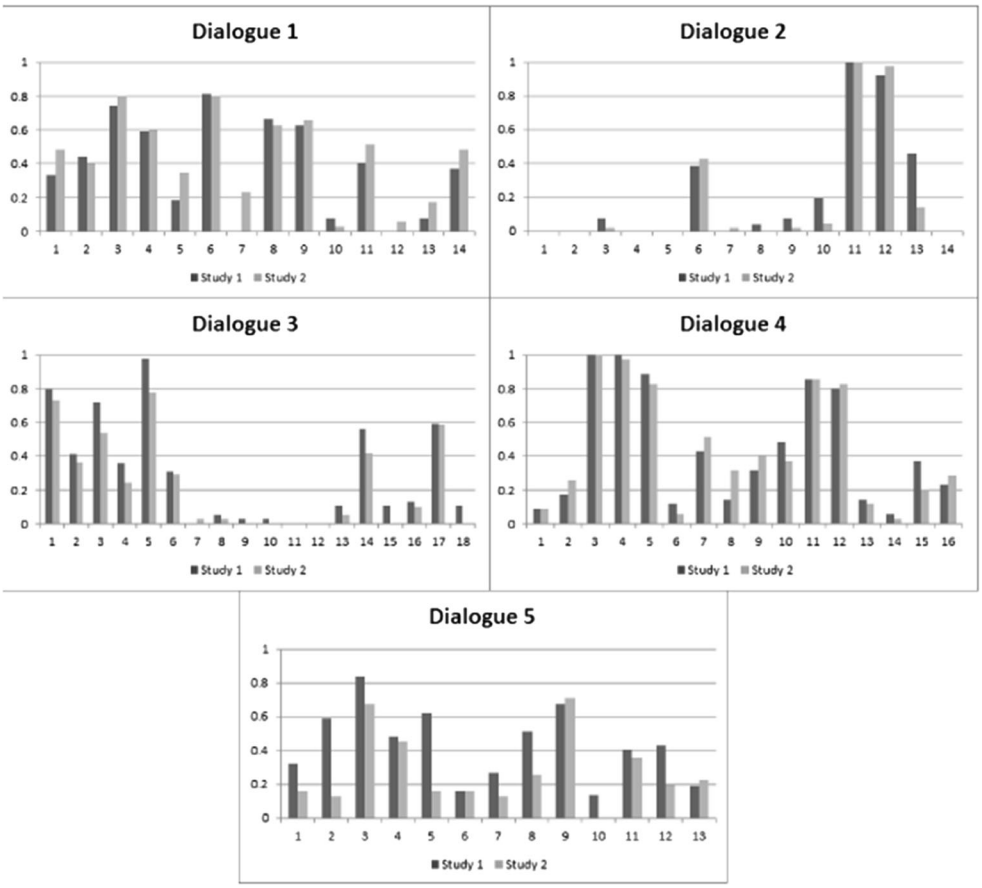

Fig. 2 Proportions of answers including each rubric item for all five gist explanation dialogues in Studies 1 and 2 (scoring rubrics for each dialogue can be found in Appendix A)

explanation dialogues combined, $r(40) = .788$, $p < .001$, as well as with each individual dialogue taken alone [Dialogue 1, $r(40) = .751$, $p < .001$; Dialogue 2, $r(40) = .747$, $p < .001$; Dialogue 3, $r(40) = .642$, $p < .001$; Dialogue 4, $r(40) = .849$, $p < .001$; Dialogue 5, $r(40) = .84$, $p < .001$]. This indicates that BRCA Gist's CO scores capture a large portion of the content covered in the rubrics and can successfully assess the amount of expected content a participant includes in an answer in a gist explanation dialogue. Because some participants did not complete one or more gist explanation dialogues, we wanted to make sure that these correlations are not caused by participants who skipped dialogues (which by default received a both a 0 CO score from BRCA Gist and a 0 rubric score from the researchers). To assess this possibility the correlations were recalculated removing skipped dialogues. All correlations remained significant [combined, $r(39) = .624$, $p < .001$; Dialogue 1, $r(24) = .611$, $p = .001$; Dialogue 2, $r(23)$, $= .586$, $p = .002$; Dialogue 3, $r(35) = .484$, $p = .002$; Dialogue 4, $r(34) = .735$, $p < .001$; Dialogue 5, $r(32) = .523$, $p = .001$], indicating that the effect was not simply inflated by skipped dialogues.

As predicted, performance on the gist explanation dialogues also had an impact on the outcome measures. The amount of content included in participant gist explanations was associated with increased ability to correctly answer questions about breast cancer and genetic risk on the multiple-choice declarative knowledge task. This is true both when content covered is measured by BRCA Gist's CO scores, $r(40) = .546$, $p < .001$, and by the researcher rubric, $r(40) = .517$, $p < .001$. The more content participants included in their gist explanations the more knowledge they were able to show on this task. Participants who included more content in their gist explanations also showed an increased understanding of the important bottom-line meaning of genetic risk of breast cancer as measured by the gist comprehension task, both when measured by CO score, $r(40) = .391$, $p = .01$, and the researcher rubric, $r(40) = .449$, $p = .003$. Participants who included more content in the gist explanations also showed increased ability to correctly classify profiles of women at varying levels of risk of breast cancer on the risk categorization task, again when measured by both CO score, $r(40) = .416$, $p = .006$, and researcher rubric, $r(40) = .379$, $p = .012$. Amount of content covered in each individual gist explanation was also associated with performance on the outcome measures, except for Dialogue 5 ("What should someone do with a negative test result?.") The correlations for each of the individual dialogues can be found in Table 1. The more that participants talked about the expected content in their verbal interactions with BRCA Gist, the better they performed on the outcome measures.

Discussion

BRCA Gist successfully evaluated participant's gist explanations. BRCA Gist's assessment of content captured much of the same content assessed using the reliable researcher rubrics. This was true both overall and for each individual dialogue, indicating that each of the five gist explanation dialogues are able to assess how much content participants include in their gist explanations. BRCA Gist is able to use this assessment to respond appropriately to participants as they generate gist explanations and provide feedback that encourages elaboration in a natural conversational manner. Also encouraging is that improvements to the latest version of BRCA Gist's rules for choosing feedback increased the tutor's proportion of appropriate responses from 85 % to 94 %.

Combined with the results of randomized, controlled experiments that have demonstrated that people randomly assigned to BRCA Gist perform significantly higher on outcome measures (Wolfe et al., 2014), the results of these fine-grained analyses strongly suggest that the gist-explanation dialogues play a major role in the success of BRCA Gist. Participants who generated more complete gist explanations showed improved performance on tasks that measured declarative knowledge and gist understanding, as well as the ability to apply that gist understanding to make decisions. This suggests that engaging in gist explanations effectively promotes acquiring the appropriate gist of the information. It is

Table 1 Study 1 dialogue measure correlations with outcome measures

	Dialogue					
	1	2	3	4	5	Combined
CO & Declarative Knowledge	.408**	.509**	.377*	.469**	.251	.546**
Rubric & Declarative Knowledge	.458**	.567**	.373*	.464**	.212	.517**
CO & Gist Comprehension	.555**	.394*	.095	.312*	.090	.391*
Rubric & Gist Comprehension	.566**	.438*	.182	.415**	.173	.449**
CO & Risk Categorization	–.069	.374*	.560**	.410**	.337*	.416**
Rubric & Risk Categorization	–.113	–.021	.453**	.372*	.210	.379*

* Correlation is significant at the .05 level. ** Correlation is significant at the .01 level.

important to remember that while participants generate gist explanations, the tutor offers feedback that encourages elaboration and prompts participants to include more content. Thus, participants who may enter the gist explanation dialogue with less initial knowledge will still be encouraged to generate a more complete answer, which promotes learning (Graesser, 2011). This conclusion is further supported by another study that examined the locus of the success of BRCA Gist, which showed that participants who received a version of BRCA Gist with the gist explanation dialogues outperformed participants who received a version with no gist explanation dialogues on the same outcome measures (Wolfe et al., 2015).

Study 2: Field study

To better establish the consequences of interacting with BRCA Gist and the robustness of the effects of its gist explanation dialogues, we replicated previous experiments on BRCA Gist's effectiveness with a broader and more diverse population. This is particularly relevant in the case of genetic testing for breast cancer risk as many women for whom this information is important are well beyond the age of the average university student. Although younger women may want to know their genetic risk prior to child bearing, older women are at higher risk of breast cancer. With this in mind, a field study was conducted to replicate prior experimental results about the effectiveness of BRCA Gist with participants recruited from a more diverse pool: from the community in upstate New York and on the Web. Using a broader sample also enabled us to examine the role of educational level, a factor that cannot be examined with samples exclusively composed of undergraduate participants.

Method

There were two arms to this study, such that 74 participants were women from the upstate New York community, and 106 participants were women recruited on the Web through www.facebook.com, www.hispanic.com, and www.avonfoundation.org. Participants recruited from both sources received $50 in return for completing the 3-h study (90 min of content presentation and 90 min of dependent measures). Of the women recruited, 46 reported having an advanced degree (beyond a B.A.), 85 reported having completed a bachelor's degree, 31 reported completing some college, 14 reported being high school graduates with no college experience, and four reported not graduating high school. The data from ten participants were excluded due to technical problems. Dropout rate (the proportion of participants who leave a study early) has the potential to be problematic for Web studies, particularly when dropout is unequal across conditions (Reips, 2002). The dropout rate for participants recruited on the Web was found to be 47 %, which is not unreasonable for Web experiments (Musch & Reips, 2000), and was approximately equal in all conditions.

Participants recruited from the upstate New York community completed the study in the laboratory, whereas participants recruited on the Web completed the study from their own computers. Participants were randomly assigned to one of three groups: the BRCA Gist tutor ($N = 57$), the National Cancer Institute (NCI) website ($N = 56$), and a control condition ($N = 67$). These are the same three conditions used in Study 1. Participants assigned to the BRCA Gist condition received the same version of the BRCA Gist tutor as was used in Study 1. Participants assigned to the NCI condition reviewed screenshots of the NCI website that presented the same information that is included in BRCA Gist. Participants reviewed the screenshots for 90 min (the same duration of the BRCA Gist tutor). Participants in the control condition were presented with an ITS also developed in SKO that delivers information about nutrition and is approximately as effortful and time consuming as BRCA Gist, but does not teach anything about breast cancer and genetic risk. After completing the content phase of the experiment all participants completed the same outcome measures used in Study 1: the multiple-choice declarative knowledge task, the gist comprehension task, and the risk categorization task.

The gist explanation dialogues were assessed using the same method as in Study 1. Participants' answers, BRCA Gist's feedback, and BRCA Gist's CO scores were extracted from logs save by the SKO system. Two independent raters gist scored one third of participants' answers together using the same content rubric (see Appendix A) and had an inter-rater reliability of .89. Two independent raters also reviewed on third of BRCA Gist's feedback using the appropriateness rubric (see Appendix B) and had an inter-rater reliability of .91.

Results

BRCA Gist participants scored the highest on the declarative knowledge task with a mean percent correct of .770 (SD = .172), followed by the NCI website group (M = .666, SD = .253), with the control condition scoring lowest (M = .570, SD = .202) on the multiple-choice knowledge task. Both the BRCA Gist and NCI groups scored significantly higher than the control group, but BRCA Gist participants did not score significantly higher than NCI participants, $F(2, 144) = 8.76$, $p < .0001$, on declarative knowledge. Planned comparisons revealed that highly educated participants (those with an advanced degree beyond the bachelor's) appeared to receive roughly equal benefit from BRCA Gist and the NCI website, whereas BRCA Gist was significantly more effective than the NCI site for participants without advanced degrees. Tukey's HSD revealed that those in the NCI group with an advanced degree (M = .824) scored significantly higher than those in the

NCI groups with a bachelor's degree or less education (M = .633), whereas educational differences were not significant among those with advanced degrees (M = .794) or those with less education (M = .772) in the BRCA Gist group, and those in the NCI group with advanced degrees were not significantly higher than either BRCA Gist group at $p < .05$. Control group participants with advanced degrees (M = .644) scored slightly but not significantly higher than those with less education (M = .564).

Participants who interacted with BRCA Gist also showed the greatest ability to recognize the important bottom-line information about breast cancer and genetic risk on the gist comprehension task (M = 5.64, SD = 0.680), followed by NCI participants (M = 5.19, SD = 0.797) and control participants (M = 4.60, SD = 0.550). Both BRCA Gist and NCI participants scored significantly higher than control participants, but the difference between BRCA Gist and the NCI website was not significant, $F(2, 144) = 27.88$, $p < .0001$. Planned comparisons again reveal that for gist comprehension, participants with an advanced degree beyond the bachelor's receive roughly equal benefit from BRCA Gist and the NCI website, whereas BRCA Gist is significantly more effective than the NCI for participants without advanced degrees. Tukey's HSD reveals that those in the NCI group with an advanced degree (M = 5.77) scored significantly higher than those in the NCI groups with a bachelor's degree or less education (M = 4.94), whereas differences in education were not significant for those with advanced degrees (M = 5.91) or less education (M = 5.57) in the BRCA Gist group, and those in the NCI group with advanced degrees were not significantly higher than either BRCA Gist group at $p < .05$. Control group participants with advanced degrees (M = 4.81) scored slightly but not significantly higher than those with less education (M = 4.53).

BRCA Gist participants were best able to correctly categorize descriptions of women in the risk categorization task, with a mean percent correct of .590 (SD = .142), followed by NCI participants (M = .557, SD = .162), and control participants (M = .489, SD = .143). Both the BRCA Gist and NCI groups performed significantly better at assessing risk than control participants, but were not significantly different from each other, $F(2, 144) = 6.43$, $p = .0016$. Planned comparisons did not reveal any additional findings with respect to education.

Participants in the BRCA Gist condition produced gist explanations using an average of 4.3 sentences (SD = 2.8) and covered an average of about one fifth of the rubric items (M = .21, SD = .15). Forty-two participants ended one or more of the dialogues before starting a gist explanation. Participants who wrote at least one sentence in their gist explanation produced answers of an average of 6.6 sentences (SD = 1.7) and covered a greater proportion of rubric items (M = .31, SD = .10). This translates to participants including on average four or five rubric items in their gist explanations. As in Study 1, participants produced gist explanations containing relevant content. As in Study 1, some rubric items were included in the gist explanations more frequently than others. Figure 2 also displays the proportions of participants' answers of at least one sentence that include each of the rubric items (see Appendix A). As Fig. 2 shows, the items that were more frequently covered by participants in Study 2 closely match those from Study 1, indicating that the more diverse population from Study 2 generated dialogues that are substantively similar to those generated by undergraduates in Study 1. We found that women without a college education were quite capable of generating appropriate gist explanations pertaining to genetic risk and breast cancer, which is illustrated in two sample gist explanation dialogues generated by women in the field study who reported completing some or no college (see Appendix C).

Judgments using the appropriateness rubric found that the tutor provided appropriate feedback to participants 95 % of the time across all gist explanation dialogues. This is comparable to the rate that the tutor was found to provide appropriate feedback in Study 1. The tutor responded appropriately 91 % of the time in Dialogue 1 ("What is breast cancer?"), 92 % of the time in Dialogue 2 ("How does breast cancer spread?"), 96 % of the time in Dialogue 3 ("How do genes affect breast cancer risk?"), 97 % of the time in Dialogue 4 ("What should someone do with a positive test results?"), and 92 % of the time in Dialogue 5 ("What should someone do with a negative test result?").

BRCA Gist determined that participants in Study 2 produced gist explanations with an average CO score of .34 (SD = .21) across all five dialogues. The average CO score of Dialogue 1 was .42 (SD = .36), the average CO score of Dialogue 2 was .39 (SD = .27), the average CO score of Dialogue 3 was .29 (SD = .20), the average CO score of Dialogue 4 was .41 (SD = .22), and the average CO score of Dialogue 5 was .35 (SD = .26).

The researcher rubric scores were strongly correlated with BRCA Gist's CO scores, as was the case in Study 1. This was both true with all gist explanation dialogues combined, $r(55) = .911$, $p < .001$, as well as with each individual dialogue alone [Dialogue 1, $r(55) = .930$; Dialogue 2, $r(55) = .877$, $p < .001$; Dialogue 3, $r(55) = .842$, $p < .001$; Dialogue 4, $r(55) = .71$, $p < .001$; Dialogue 5, $r(55) = .512$, $p = .001$]. The correlations were checked as in Study 1 to be sure that they were not being driven by skipped dialogues. Skipped dialogues were remove and all correlations remained significant except the fifth dialogue, "What should someone do with a negative test result?" when taken alone [combined, $r(45) = .589$, $p < .001$; Dialogue 1, $r(33) = .704$, $p < .001$; Dialogue 2, $r(40) = .484$, $p < .001$; Dialogue 3, $r(40) = .534$, $p < .001$; Dialogue 4, $r(35) = .616$, $p < .001$; Dialogue 5, $r(33) = .295$, $p = .085$].

The amount of content covered in gist explanations was again associated with the outcome measures. Participants who included more expected content in gist explanations performed better on the multiple-choice knowledge task, measured by both BRCA Gist's CO score, $r(55) = .618$, $p < .001$, and the researcher rubric, $r(55) = .623$, $p < .001$. Content covered in the gist explanations was also associated with better understanding of the gist of breast cancer risk as measured by the gist comprehension task, measured by both CO score, $r(55) = .668$, $p < .001$, and the researcher rubric, $r(55) = .664$, $p < .001$. Participants who covered more content also showed better ability to correctly assess women's genetic risk as measured in the risk categorization task, measured both by CO score, $r(55) = .428$, $p = .001$, and by the researcher rubric, $r(55) = .487$, $p < .001$. As in Study 1, including more content in most of the individual gist explanations was also correlated with performance on the outcome measures, except for the fourth and fifth gist explanations, "What should someone do with a positive test result?" and "What should someone do with a negative test result?" (the correlations for the individual dialogues of Study 2 can be found in Table 2).

Discussion

The experimental results show that BRCA Gist is effective with a broader population. BRCA Gist was more effective at delivering information about breast cancer risk than reading static information from the NCI website and a control tutor for women without advanced degrees (i.e., a bachelor's degree or less). However, women with an advanced degree benefited roughly equally from BRCA Gist and the NCI website text. This demonstrates that the effect of BRCA Gist is more robust across education levels than the information on the NCI website. BRCA Gist is able to convey the gist of information about genetic risk of breast cancer to women of all educational levels. However, women with advanced degrees appear to be better able to extract the key information from simply reading the text of the NCI website. BRCA Gist may thus be more effective at achieving learning gains in the general population.

BRCA Gist consistently evaluated participants' gist explanations successfully and captured much of the same information covered in the reliable researcher rubrics. This was true taking all dialogues together and for each individual dialogue, again indicating that each of the five gist explanation dialogues is able to successfully evaluate participants' content. It is encouraging to replicate this finding with participants from a broader background as this demonstrates that BRCA Gist is able to successfully evaluate any differences in the kinds of gist explanations generated by women of different backgrounds. As Fig. 2 suggests, it appears that women in this more diverse field study had interactions with BRCA Gist that are substantively similar to those conducted in the lab with undergraduate participants. BRCA Gist was also able to respond appropriately to the gist explanations generated by participants in the broader sample, delivering appropriate feedback that encouraged elaboration 95 % of the time.

As in Study 1, participants who included more of the expected material in their answers to the gist explanation dialogues also performed better on the outcome measures. A participant who covered more content in her gist explanations was able to show more declarative knowledge, show a better gist understanding of risk of breast cancer, and better categorize women of different levels of risk of breast cancer. Combined with the overall evidence for the efficacy of BRCA Gist, these results again suggests that the gist explanation dialogues are an important component of BRCA Gist's success.

General discussion

The results of the fine-grained analyses of the gist explanation dialogues of BRCA Gist are consistent with the hypothesis that engaging in gist explanation dialogues with a well-developed ITS promotes learning gains. FTT suggests that encouraging people to develop and elaborate upon gist explanations is an avenue to improving learning, comprehension and decision-making. The success of BRCA Gist's dialogues relies on the ability of its natural-language processing engine

Table 2 Study 2 dialogue measure correlations with outcome measures

	Dialogue					
	1	2	3	4	5	Combined
CO & Declarative Knowledge	.483**	.560**	.533**	.170	.129	.618**
Rubric & Declarative Knowledge	.495**	.527**	.613**	.280	.164	.623**
CO & Gist Comprehension	.483**	.577**	.564**	.187	.173	.668**
Rubric & Gist Comprehension	.520**	.561**	.671**	.200	.097	.664**
CO & Risk Categorization	.411**	.368**	.414**	.096	–.145	.428**
Rubric & Risk Categorization	.442**	.268*	.527**	.226	.071	.487**

* Correlation is significant at the .05 level. ** Correlation is significant at the .01 level.

to adequately assess learner input. The results of these analyses indicate that BRCA Gist's assessment of the amount of expected content covered in learner's gist explanation dialogues is comparable to human raters. The CO scores used by BRCA Gist to measure the semantic similarity of learner input to its expectation texts capture much of the good content of learner's answers, and this much of this content is the same as that measured by researchers using a reliable rubric. This success can attributed to the quality of the script the tutor uses to assess learner input and provide appropriate responses. The reliable process used to develop BRCA Gist's script involved iterative empirical testing of the tutor to carefully construct the most effective expectation texts and calibrate the tutor's natural-language engine to most appropriately evaluate the similarity of learner input to those expectations. See Wolfe et al. (2013) for a detailed description of this process. Because each gist explanation differs in the amount and kind of content it covers, different final CO scores can be reflective of a good gist explanation. For example, good answers to the question "What is breast cancer?" (Dialogue 1) frequently achieve final CO scores of .7 or higher, whereas good answers to the question "How do genes affect breast cancer risk?" (Dialogue 3) rarely exceed .5. However, this does not hinder BRCA Gist's ability to assess participant gist explanations as each dialogue has been properly calibrated such that a better gist explanation that includes more expected content will score higher than an explanation that includes less content on the individual dialogue's scale.

BRCA Gist is also able to use its assessment of learner input to provide appropriate feedback to encourage elaboration of learner gist explanation dialogues. Appropriate feedback does not necessarily indicate optimal feedback, so there is still room to improve the feedback delivered by BRCA Gist. However, this improvement is especially notable given the streamlined nature of BRCA Gist's natural-language processing. More sophisticated systems such as AutoTutor employ expectation and misconception tailored dialogues that are able to track learner answers for multiple expectations and monitor for the presence of misconceptions, and can use any of these sources to select feedback (Graesser, Chipman, et al., 2005). The SKO platform on which BRCA Gist operates does not yet have this capability.

We employed several strategies to predict the feedback that would be most beneficial to learners and what situations learners might be in when they receive it. One strategy was to tailor feedback to provide prompts for content that previous learners tend to leave out of their gist explanations, or content that is frequently missed on outcome measures. Another strategy was to improve appropriateness by predicting the order learners would include content based on dialogues generated by prior learners. For example, in Dialogue 3 ("How do genes affect breast cancer risk?") learners frequently include the idea that having family members with breast cancer factors into a woman's genetic risk early in their gist explanations. Thus, feedback that prompts for this content would be most helpful and sound most natural to more learners when delivered early in the dialogue rather than late. Feedback about content typically included later in gist explanations (such as that women of certain ethnic backgrounds may be at greater genetic risk) is most likely to be helpful and appropriate later in the dialogue. We also found it useful to examine each line of feedback and try to make it sound more appropriate in a variety of contexts. Finally, in order to help learners generate the most beneficial gist explanations for understanding the important takeaway information about the decision to receive genetic testing, feedback encouraged learners to include decision relevant content and content about consequences of decisions.

Participants who generated more successful gist explanation dialogues after interacting with BRCA Gist were able to demonstrate greater knowledge (as measured by the declarative knowledge task), showed greater understanding of the essential bottom-line meaning of the content (as measured by the gist comprehension task), and were better able to apply that knowledge to make decisions about risk (as measured by the risk comprehension task. This is an essential finding for establishing that there are benefits to engaging in gist explanations. However, because these results are correlational, it could be that participants who are more intelligent or have more knowledge about the topic are both able to produce better gist explanations and perform better on the outcome measures. Although this may be logically possible, observing that participants with more successful gist explanation dialogues show more knowledge is an important step in establishing the benefits of engaging in gist explanations. In addition, there are reasons to believe that the gist explanation dialogues contribute to the success of BRCA Gist. The tutor consistently prompts participants to include more content and elaborate on their gist explanations. The feedback presented by the tutor is designed to encourage participants to focus on the relevant content measured by the outcome tasks, so participants who are more engaged with the tutor should generate better gist explanations and reap the benefits of self-explanation regardless of initial knowledge. A recent study on the locus of the efficacy of BRCA Gist also determined that the inclusion of the gist explanation dialogues significantly increased performance on knowledge measures (Wolfe et al., 2015). In future work, we plan to experimentally manipulate whether participates engage in dialogues that ask for bottom-line gist information or that ask for specific verbatim information to further investigate the specific benefits of engaging in a gist explanation. This will allow us to more precisely evaluate that it is the focus on gist content (bottom-line meaning information, decision relevant information, and information about the consequences of decisions) that is responsible for the success of BRCA Gist and gist explanations. Additionally, conducting reliable fine-grained analyses of

tutorial dialogues is often challenging and time consuming. These data suggest that BRCA Gist's coverage scores can be used as a reliable index of content coverage. This is an encouraging finding for the development of ITS that interact in natural language. Conversational ITS that are developed with well-constructed gist explanation scripts can both aide learners by responding with appropriate feedback and provide accurate assessments of the content covered by learners. Other discourse analysis tools such as Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) may also prove useful in advancing our understanding of gist explanations in tutorial dialogues.

Author note The project described was supported by Award Number R21CA149796 from the National Cancer Institute (NCI). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NCI or the National Institutes of Health. We thank the NCI for its support. We also thank Rachel Aron, Andrew Circelli, Cecelia Favede, Jeremy Long, Mitch McDaniel, Ian Murphy, Kendall Powell, and Michael Thomas for capable assistance with the data collection.

Appendixes

Appendix A: Dialogue rubrics

Dialogue 1: "What is breast cancer?" and Dialogue 2: "How does it start, grow, and spread?"

1. Breast cancer is often found in the ducts (tubes that carry milk to the nipple).
2. Breast cancer is often found in the lobules (glands that make milk).
3. Cancer is a malfunction in cell production or cell division or uncontrolled cell growth (mentions cells).
4. Breast cancer can be due to mutations to the BRCA1 and BRCA2 genes (mentions BRCA).
5. BRCA1 and BRCA2 genes (in their non-mutated form) act as tumor suppressors.
6. The damaged cells form a mass, lump, or tumor.
7. As cancer/malfunction continues undetected, the mass will often grow larger (growth in cancer, tumors, etc.).
8. These masses can be benign, in which case they are relatively harmless.
9. Tumors or cancer or masses can be malignant.
10. Malignant tumors may do damage to other parts of the body.
11. A metastatic tumor is cancer that grows and spreads from the original site to other areas of the body (any including lymph nodes).
12. Cancer cells spread/metastatic tumors travel to other areas of the body via blood vessels and lymph vessels.
13. Malignant tumors can be a threat can be a threat to life, because the cancerous cells can then invade and damage vital organs and tissues.
14. The cause of cancer can be genetic (more loosely runs in families, hereditary).

Dialogue 3: "How do genes affect breast cancer risk?"

1. Women who have a first-degree relative (mother, sister, or daughter) or other close relative with breast cancer may be at increased risk of developing breast cancer.
2. Women who have a first-degree relative (mother, sister, or daughter) or other close relative with ovarian cancer may be at increased risk of developing breast cancer.
3. If a family member had breast cancer before age 50, genetic breast cancer risk is higher.
4. BRCA1 and BRCA2 are genes that function as tumor suppressors to prevent unregulated cell growth in normal cells.
5. When BRCA1 and BRCA2 genes are harmfully mutated, the risk of developing breast cancer increases.
6. BRCA1 and BRCA2 mutations are inherited (heritability of genetic mutations).
7. BRCA mutations can be detected with blood tests (any test mentioned).
8. A woman is five times more likely to develop breast cancer is she has a harmful mutation in her BRCA1 or BRCA2 genes (relative risk compared to the population/base rate).
9. Women with BRCA1 and BRCA2 mutations have a higher chance of developing breast cancer before the age of 50.
10. Most women who have breast cancer under the age of 50 have BRCA mutations.
11. The majority of inherited breast cancer cases result from mutations in BRCA1 or BRCA2 genes.
12. Mutations in other genes are also associated with the development of breast cancer, including TP53, PTEN, STK11/LKB1, CDH1, CHEK2, ATM, MLH1, and MSH2 ("other" or any of these specific genes).
13. Autosomal dominant inheritance occurs through either the maternal or paternal side of the family (mother's side and / or father's side).
14. The frequency of BRCA mutations for Ashkenazi Jewish women is higher than that found in the general population (increased genetic risk for Ashkenazi Jewish women).
15. Other ethnic groups around the world, such as the Norwegian, Dutch, and Icelandic peoples (any of these or "other") also have higher frequencies of specific BRCA1 and BRCA2 mutations.
16. About 60 % of women (600 out of 1,000) who have inherited a harmful mutation in BRCA1 or BRCA2 will

develop breast cancer (absolute risk for breast cancer with BRCA mutations).
17. Not everyone who has inherited a BRCA mutation will develop cancer (any indication of chance of breast cancer is less than 100 %).
18. Male breast cancer, is associated with BRCA2 (and to a lesser extent BRCA1) genetic mutations (any male breast cancer & BRCA connection).

Dialogue 4: What should someone do if she finds out that she has an inherited altered BRCA1 or BRCA2 gene (meaning a positive test result for genetic breast cancer risk)?

1. Just because a woman has a genetic predisposition for the development of cancer (positive test for BRCA1 or BRCA2) does not mean that she will develop cancer.
2. The woman should talk to her physician or genetic counselor about the risk and measures they should take to prevent breast cancer.
3. Based on a positive genetic test, the woman should take steps to manage her cancer risk (any one or more risk management steps).
4. There are many methods in which a woman can manage her breast cancer risk (many or more than one specific risk management steps).
5. Active Surveillance is getting more frequent cancer screenings, so that any cancerous cells can be detected early (mammography, breast exams, or MRI).
6. Active Surveillance does not reduce the risk of getting cancer.
7. Risk management method: mammography.
8. Risk management method: clinical breast exams or more frequent breast exams.
9. Risk management method: magnetic resonance imaging or MRI.
10. Methods to reduce or eliminate the risk of getting cancer (chemoprevention or prophylactic mastectomy surgery)—Credit for that these methods reduce risk.
11. Risk management method: chemoprevention (any natural or synthetic substances).
12. Risk management methods with the goal of reducing or eliminating the risk of developing cancer to remove areas of "at-risk" breast tissue (prophylactic mastectomy surgery) or reduce the risk of cancer development (chemoprevention).
13. Even with prophylactic surgery (mastectomy) there is still a chance of developing breast cancer. There have been cases where women have developed breast cancer, ovarian cancer, or primary peritoneal carcinomatosis, even after prophylactic surgery (risk > 0 % after surgery).
14. A woman who is positive for BRCA mutations can also take steps to reduce environmental risk factors that increase her chances of developing breast cancer (decrease alcohol consumption, eat healthier, any lifestyle or environmental factor).
15. An understanding of a woman's risk profile can provide valuable information for her relatives; she can share the information with family members.
16. More loosely, a woman who receives a positive test results should "see her doctor."

Dialogue 5: What should someone do if she finds out that she does not have an inherited altered BRCA1 or BRCA2 gene (meaning a negative test result for genetic breast cancer risk) / what does a negative test result mean?

1. If a woman has a family history of a specific gene mutation (BRCA) and she tests negative for this mutation, this is called a "true negative."
2. A woman who receives a true negative test result can be sure she probably has not inherited a mutated BRCA gene.
3. A true negative test result indicates there is no BRCA mutation
4. A true negative means that the person's risk of cancer is the same as that of the general population (actually slightly less).
5. Even with a true negative a woman may still develop cancer but is at no additional risk when compared to the average woman.
6. Most women who have breast cancer do not have BRCA1 or BRCA2 mutations.
7. If a woman's family does have a history of breast cancer or ovarian cancer but those relatives have been tested and do not have BRCA mutation, a negative test result has less/no meaning and may be a "false negative."
8. A woman who receives a false negative test result is still at increased risk
9. A false negative test result means the test did not detect a BRCA mutation but there actually is one
10. An ambiguous test result means there is a change in BRCA that is not associated with breast cancer in other people.
11. The woman may still have some other unknown harmful mutations that can lead to breast cancer.
12. The woman can provide valuable information for her relatives; she can share the information with family members.
13. She should continue to follow standard practice (regular checkups, self-exams, etc.).

Appendix B: Appropriateness rubrics

Appropriate feedback (must meet all three criteria):

1. Properly recognizes correct / incorrect information in participant input

2. Encourages participants to elaborate their answers
3. Sounds natural following the previous input

Inappropriate feedback (meets any of the criteria):

1. Does not properly recognize correct / incorrect information in participant input
2. Discourages participants from elaborating on answers
3. Does not sound natural following the previous input

Neutral feedback:

Any feedback that does not clearly meet all three criteria to be classified as appropriate but does not clearly meet any of the criteria to be classified as inappropriate.

Appendix C: Sample gist explanation dialogue scripts from the field study

Sample Dialogue #1

Participant Age: 38

Participant Education: High School Graduate.

Participant: Breast cancer can spread through the lymph system.

BRCA Gist: Good job. Can you add something about the ways breast cancer can spread?

Participant: It can also spread through the blood stream.

BRCA Gist: Very good. Can you say something about the process of metastasis?

Participant: Cancer can also metastasize to your vital organs.

Participant: Breast cancer can be invasive meaning it spreads out of the breast.

BRCA Gist: Okay. Can you say something new about blood vessels and the way breast cancer spreads?

Participant: It has the ability to circulate through the bloodstream and then invade and grow into the surrounding tissues.

BRCA Gist: Here's something. Can you add something you have not said before about the surrounding tissues?

Participant: Cancer can colonize and cause spread cancer to any organ.

BRCA Gist: Okay. Try adding one more thing about how breast cancer spreads.

Participant: Metastis cancer can spread to the bones.

BRCA Gist: Alright, we're done. Please click on the finish button to continue.

Sample Dialogue #2

Participant Age: 55

Participant Education: Some College

Participant: Breast cancer spreads when cells duplicate out of control and develop too many cells for the body to use, causing a lump of mutated cells to form.

BRCA Gist: Good job. Can you add something about the ways breast cancer can spread?

Participant: The cells can travel through the body through lymphnodes or blood.

BRCA Gist: Very good. Can you say something about the process of metastasis?

Participant: Cancer cells continue to duplicate throughout the body.

Participant: The mutated cells attach to healthy parts of the body, they can cause normal, healthy parts of the body to stop functioning normally.

BRCA Gist: You are doing fine. Can you say more about how breast cancer spreads?

Participant: Breast cancer cells can metasticize into other parts of the body, but they are still breast cancer cells and should be treated as such.

Participant: The more mutated cells spread through out the body, the great your life rishk.

BRCA Gist: Good job. Please click on finish to continue.

References

Aleven, V. A., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26,* 147–179.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences, 4,* 167–207.

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

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. (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.

Brewer, N. T., Richman, A. R., DeFrank, J. T., Reyna, V. F., & Carey, L. A. (2012). Improving communication of breast cancer recurrence risk. *Breast Cancer Research and Treatment, 133,* 553–561.

Cedillos-Whynott, E. M., Wolfe, C. R., Widmer, C. L., Brust-Renck, P. G., & Reyna, V. F. (2015). *The effectiveness of argumentation in tutorial dialogues with an Intelligent Tutoring System*. Manuscript under review.

Chi, M. T. (1996). Constructing self-explanations and scaffolded explanations in tutoring. *Applied Cognitive Psychology, 10,* 33–49.

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., 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, 13,* 145–182.

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

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25,* 471–533.

Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011). Empirically evaluating the application of reinforcement learning to the induction

of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction, 21,* 137–180.

Col, N., Bozzuto, L., Kirkegaard, P., Koelewijn-van Loon, M., Majeed, H., Ng, C. J., & Pacheco-Huergo, V. (2011). Interprofessional education about shared decision making for patients in primary care settings. *Journal of Interprofessional Care, 25,* 409–415. doi:10.3109/13561820.2011.619071

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.

Fraenkel, L., Peters, E., Charpentier, P., Olsen, B., Errante, L., Schoen, R. T., & Reyna, V. (2012). Decision tool to improve the quality of care in rheumatoid arthritis. *Arthritis Care Research, 64,* 977–985. doi:10.1002/acr.21657

Gholson, B., Witherspoon, A., Morgan, B., Brittingham, J. K., Coles, R., Graesser, A. C., . . . Craig, S. D. (2009). Exploring the deep-level reasoning questions effect during vicarious learning among eighth to eleventh graders in the domains of computer literacy and Newtonian physics. *Instructional Science*, *37*, 487–493.

Graesser, A. C. (2011). Learning, thinking, and emoting with discourse technologies. *American Psychologist, 66,* 746–757. doi:10.1037/a0024974

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

Graesser, A. C., Jeon, M., & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes, 45,* 298–322.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004a). 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., Louwerse, M. M., & Cai, Z. (2004b). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36,* 193–202. doi:10.3758/BF03195564

Graesser, A. C., McNamara, D. S., & VanLehn, K. (2005b). 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-to-one tutoring. *Applied Cognitive Psychology, 9,* 495–522.

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.

Hoskins, K. F., Zwaagstra, A., & Ranz, M. (2006). Validation of a tool for identifying women at high risk for hereditary breast cancer in population-based screening. *Cancer, 107,* 1769–1776.

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* (pp. 802–802). 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 T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 401–426). Mahwah, NJ: Erlbaum.

Hu, X., Han, L., & Cai, Z. (2008). *Semantic decomposition of student's contributions: an implementation of LCC in AutoTutor Lite.* Article presented to the Society for Computers in Psychology, Chicago, Illinois.

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* (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.

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). An introduction to latent semantic analysis. *Discourse Processes, 25,* 259–284. doi:10.1080/01638539809545028

Magliano, J. P., Trabasso, T., & Graesser, A. C. (1999). Strategic processes during comprehension. *Journal of Educational Psychology, 91,* 615–629.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38,* 1–30. doi:10.1207/s15326950dp3801_1

Musch, J., & Reips, U.-D. (2000). A brief history of Web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 61–87). San Diego, CA: Academic Press.

Nelson, H. D., Huffman, L. H., Fu, R., & Harris, E. L. (2005). Genetic risk assessment and BRCA mutation testing for breast and ovarian cancer susceptibility: Systematic evidence review for the US Preventive Services Task Force. *Annals of Internal Medicine, 143,* 362–379.

Ohlsson, S. (1986). Some principles of intelligent tutoring. *Instructional Science, 14,* 293–326.

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology, 49,* 243–256. doi:10.1027/1618-3169.49.4.243

Reyna, V. F. (2004). How people make decisions that involve risk: A dual-processes approach. *Current Directions in Psychological Science, 13,* 60–66.

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

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. (2011). Dual processes in decision making and developmental neuroscience: A fuzzy-trace model. *Developmental Review, 31,* 180–206.

Reyna, V. F., Estrada, S. M., DeMarinis, J. A., Myers, R. M., Stanisz, J. M., & Mills, B. A. (2011). Neurobiological and memory models of risky decision making in adolescents versus young adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1125–1142. doi:10.1037/a0023943

Reyna, V. F., & Lloyd, F. J. (2006). Physician decision making and cardiac risk: Effects of knowledge, risk perception, risk tolerance, and fuzzy processing. *Journal of Experimental Psychology: Applied, 12,* 179–195. doi:10.1037/1076-898X.12.3.179

Reyna, V. F., & Mills, B. A. (2014). Theoretically motivated interventions for reducing sexual risk taking in adolescence: A randomized controlled experiment applying fuzzy-trace theory. *Journal of Experimental Psychology: General, 143,* 1627–1648. doi:10.1037/a0036717

Reyna, V. F., Nelson, W. L., Han, P. K., & Pignone, M. P. (2015). Decision making and cancer. *American Psychologist, 70,* 105–118. doi:10.1037/a0036834

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

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. (1992). A model of the self-explanation effect. *Journal of the Learning Sciences, 2,* 1–59.

Weil, A. M., Wolfe, C. R., Reyna, V. F., Widmer, C. L., Cedillos-Whynott, E. M., & Brust-Renck, P. G. (2015). *Proficiency of FPPI and objective numeracy in estimating breast cancer risk*. Manuscript under review.

Wolfe, C. R., & Fisher, C. R. (2013). Individual differences in base rate neglect: A fuzzy processing preference index. *Learning and Individual Differences, 25,* 1–11. doi:10.1016/j.lindif.2013.03.003

Wolfe, C. R., Reyna, V. F., Widmer, C. L., Brust-Renck, P. G., Cedillos, E. M., Hu, X., & Weil, A. M. (2015). *Understanding genetic breast cancer risk: Processing loci of the BRCA Gist intelligent tutoring system*. Manuscript under review.

Wolfe, C. R., Reyna, V. F., Widmer, C. L., Cedillos, E. M., Fisher, C. R., Brust-Renck, P. G., & Weil, A. M. (2014). Efficacy of a Web-based intelligent tutoring system for communicating genetic risk of breast cancer: A fuzzy-trace theory approach. *Medical Decision Making, 35,* 46–59.

Wolfe, C. R., Widmer, C. L., Reyna, V. F., Hu, X., Cedillos, E. M., Fisher, C. R., . . . Weil, A. M. (2013). The development and analysis of tutorial dialogues in AutoTutor Lite. *Behavior Research Methods, 45*, 623–636. doi:10.3758/s13428-013-0352-z