Behav Res (2016) 48:857–868
DOI 10.3758/s13428-015-0681-1

REVIEW

The effectiveness of argumentation in tutorial dialogues with an Intelligent Tutoring System for genetic risk of breast cancer

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

Published online: 28 October 2015

Abstract *BRCA Gist* is an Intelligent Tutoring System that helps women understand issues related to genetic testing and breast cancer risk. In two laboratory experiments and a field experiment with community and web-based samples, an avatar asked 120 participants to produce arguments for and against genetic testing for breast cancer risk. Two raters assessed the number of argumentation elements (claim, reason, backing, etc.) found in response to prompts soliciting arguments for and against genetic testing for breast cancer risk (IRR=.85). When asked to argue for genetic testing, 53.3 % failed to meet the minimum operational definition of making an argument, a claim supported by one or more reasons. When asked to argue against genetic testing, 59.3 % failed to do so. Of those who failed to generate arguments most simply listed disconnected reasons. However, participants who provided arguments against testing (40.7 %) performed significantly higher on a posttest of declarative knowledge. In each study we found positive correlations between the quality of arguments against genetic testing (i.e., number of argumentation elements) and genetic risk categorization scores. Although most interactions did not contain two or more argument elements, when more elements of arguments were included in the argument against genetic testing interaction, participants had greater learning outcomes. Apparently, many participants lack skills in making coherent arguments. These results suggest an association between argumentation ability (knowing how to make complex arguments) and subsequent learning. Better education in developing arguments may be necessary for people to learn from generating arguments within Intelligent Tutoring Systems and other settings.

Keywords Intelligent Tutoring Systems · ITS · *BRCA Gist* · Argumentation · Argumentation discourse

Introduction

The relationship between generating arguments and subsequent learning appears to be mixed in the argumentation and learning literature. Argument generation has been shown to be effective at achieving learning gains in students in some studies (Wiley & Voss, 1999), yet in other studies the results are more ambiguous (Scheuer, Loll, Pinkwart, & McLaren, 2010). The impact of including argument generation within an Intelligent Tutoring System (ITS) has received little research attention. Furthermore, the use of argumentation to facilitate gist understanding that would aid in patient medical decision making has only recently been explored as a component of studies conducted by Wolfe and colleagues (Wolfe, Reyna, Brust-Renck, et al., 2013; Wolfe, Reyna, Widmer, et al., 2013; Wolfe, et al., under review; Wolfe et al., 2015). *BRCA Gist* (Breast Cancer and Genetics Intelligent Semantic Tutoring) appears to be the first ITS applied to lay people's medical decision making (Wolfe et al., 2015). *BRCA Gist* has been found to effectively help women understand and make decisions about genetic testing for breast cancer risk (Widmer et al., 2015; Wolfe, Widmer, Reyna, et al., 2013; Wolfe et al., 2015; Wolfe, Reyna, Brust-Renck, et al., 2013; Wolfe, Reyna, Widmer, et al., 2013). The goal of the research presented in this paper is to better understand the argument-based verbal interactions between *BRCA Gist* and research participants through a fine-grained analysis of tutorial dialogues and to

✉ Elizabeth M. Cedillos-Whynott
cedillem@miamioh.edu

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

[2] Department of Psychology, Cornell University, Ithaca, NY, USA

examine the effectiveness of providing arguments on learning outcomes including content knowledge regarding genetic risk for breast cancer as presented by *BRCA Gist*, the ability to categorize women into levels of breast cancer risk, and gist comprehension.

BRCA Gist: an intelligent tutoring system

An Intelligent Tutoring System (ITS) is a computer-based tutoring program that can be adapted to fit the needs of individual students. An effective ITS can monitor student knowledge, give examples, ask deep level reasoning questions, answer questions, explain answers, and clarify misunderstandings, as well as provide feedback, encouragement, and additional information (Graesser, 2011; Craig, Sullins, Witherspoon, & Gholson, 2006; Glaser & Bassok, 1989; Graesser, Chipman, Haynes, Olney, 2005; Graesser, McNamara, 2010; Ohlsson, 1986; Wolfe & Cedillos, 2015). An effective ITS can lead to learning gains that are comparable to those of human tutors (VanLehn, 2011). A recent meta-analysis has demonstrated ITS's to have an effect size of about .76 sigma above standard classroom teaching methods, similar to that of most non-expert human tutors (VanLehn, 2011).

BRCA Gist was built on the Shareable Knowledge Objects (SKO) platform. SKO (formerly AutoTutor Lite) is a web-based platform for creating web-based ITS's that employs many of the previously mentioned capabilities of ITS (Hu, Han, & Cai, 2008; Wolfe, Fisher, Reyna, & Hu, 2012; Wolfe, Widmer, Reyna, Hu, et al., 2013). Furthermore, because SKO is a web-based ITS, it allows users to interact with it over any internet browser. SKO also allows the creator to generate tutorials from the ground up. This includes choosing from many animated avatars, agents that interact with the user via verbal responses and feedback accompanied by facial expressions and movements (Graesser, VanLehn, Rose, Jordan, & Harter, 2001). SKO also permits additional interactive elements in the form of pictures, videos, and responses consisting of fill-in-the-blank, matching, multiple-choice, and tutorial dialogues (Wolfe, Fisher, Reyna, & Hu, 2012).

Guided by Fuzzy Trace Theory (FTT, Reyna, 2008; Reyna 2012), *BRCA Gist* uses a number of techniques to aid in gist comprehension, including using pictures, videos, and engaging users in tutorial dialogues which emphasize both gist explanation and argumentation. Gist comprehension is a focus on the relevant bottom-line semantic meaning of information that does not include more specific verbatim information (Clarian & Koul, 2006). A gist understanding of technical and complex information has been associated with better learning outcomes when compared to verbatim memorization (Lloyd & Reyna, 2009). For example, Clariana and Koul (2006) demonstrated the effect multiple-try feedback has in helping to create gist, or fuzzy, representations of knowledge. The multiple-try feedback allowed for clarification of knowledge and stopped incorrect knowledge from being incorporated during study, leading to learning outcomes that demonstrated better understanding due to gist rather than verbatim knowledge representations. Furthermore, interventions targeting gist representations have been shown to have more persistent effects on behavioral outcomes than verbatim representations, particularly when making risky decisions (Reyna & Mills, 2014; see also Reyna, Estrada, DeMarinis, Myers, Stanisz, & mills, 2011). More specifically, in randomized controlled experiments, *BRCA Gist* has been shown to lead to improvements in gist comprehension, general understanding of information about breast cancer and genetic-risk (as demonstrated on multiple-choice declarative knowledge items), as well as decision making and risk assessment (Wolfe, Reyna, Brust-Renck, et al., 2013; Wolfe, Reyna, Widmer, et al., 2013; Wolfe et al., 2015).

The interactive tutorial dialogues of *BRCA Gist* use Latent Semantic Analysis (LSA; Graesser, Wiemer-Hastings, Wiemer-Hastings, & Harter, 2000), which utilizes a mathematical algorithm to determine the conceptual and semantic similarity of the user's response to the creator's desired response (Graesser, Lu, Jackson, Mitchell, Ventura, Olney, & Louwerse, 2004). *BRCA Gist* uses LSA to interpret responses entered by a user in a natural language (in our case English) to select from pre-programmed hints, prompts, and pumps to encourage users to elaborate on their responses. A typical tutorial dialogue interaction between *BRCA Gist* and a participant proceeds in the following manner:

> *BRCA Gist*: What is the case for genetic testing for breast cancer risk?
> Participant: Genetic testing for breast cancer risk can be both positive and negative.
> *BRCA Gist*: You are off to a good start. Please keep making a case in favor of testing.
> Participant: However, testing can not only make the person aware of his or her condition, but can help that person take steps towards safety precautions that can be taken.

BRCA Gist generates a coverage (CO) score, between 0 and 1, representing the semantic similarity between a user's answer to a question and the tutor's expectation text. This enables the tutor to respond appropriately and help the user understand and better learn the material (Wolfe et al., 2013). Interspersed throughout the *BRCA Gist* tutorial were five tutorial dialogues requiring participants to make gist explanations. Human tutors and ITS often encourage people to generate self-explanations of what they have learned (Chi, Leeuw, Chiu, & LaVancher, 1994). The encouragement of these self-explanations has been shown to account for much of the success of tutors (Chi, 2000). Verbal exchanges between a student

and an ITS such as AutoTutor may be 100 turns long (Graesser, McNamara, & VanLehn, 2005). However, self-explanations in *BRCA Gist* are briefer, with users typically making about seven dialogue moves to form a *gist explanation*. Gist explanation requires a briefer explanation that emphasizes the bottom line meaning rather than verbatim knowledge, which requires an exact representation of stored knowledge (Widmer et al., 2015). The tutorial concluded with two tutorial dialogues in which participants were asked to first provide an argument for and then provide an argument against genetic testing for breast cancer risk. The dialogues participants generated in response to these two argument prompts are the focus of this investigation.

Wolfe, Widmer, Reyna, Hu, and colleagues (2013) compared the efficacy of *BRCA Gist* to that of an irrelevant SKO control group, and screen shots from the National Cancer Institute (NCI) website. These screen shots contained similar information about genetic risk of breast cancer, but lacked the FTT-guided techniques of the ITS (Wolfe, Reyna, Brust-Renck, et al., 2013; Wolfe, Reyna, Widmer, et al., 2013; Wolfe et al., 2015). Findings demonstrate that participants who interacted with *BRCA Gist* performed significantly better than both the NCI and control group in declarative knowledge and in correctly endorsing statements that require gist comprehension. When participants had to make risk assessments by categorizing women into low-, medium-, or high-risk groups, and determine whether they should undergo genetic testing, the *BRCA Gist* group performed significantly better than the control group (Wolfe et al., 2015). Overall, these findings demonstrate the effectiveness of the *BRCA Gist* tutor in guiding users to a gist understanding of information (Wolfe, Reyna, Brust-Renck, et al., 2013; Wolfe, Reyna, Widmer, et al., 2013; Wolfe et al., 2015). Having already demonstrated the effectiveness of *BRCA Gist*, we wanted to examine the different teaching techniques used by *BRCA Gist*. More specifically, we examined the role generating arguments plays in improving learning outcomes after interactions with an ITS.

Argumentation

Our knowledge regarding the role of argumentation in learning has been furthered by research led by cognitive psychologists (Kuhn & Udell, 2003; Voss, Fincher-Kiefer, Wiley, & Silfes; Wiley & Voss, 1999). Being able to produce a well-reasoned and thoughtful argument is a necessary skill for students bound for various fields of study such as the sciences, social sciences, and humanities. Although web-based tutorials have been built to teach argumentation (e.g., Wolfe, Britt, Petrovic, Albrecht, & Kopp, 2009) the potential of ITS's to promote learning in substantive domains through argument generation has received scant attention.

Use of argumentation as a pedagogical method by ITS's and other computerized systems has had inconsistent support (Scheuer, Loll, Pinkwart, & McLaren, 2010). This has largely been caused by the fact that different topics require different methods of implementation. For example, using computerized diagrams to form arguments is helpful in understanding public policy (Easterday, Aleven, & Scheines, 2007) but not better than traditional methods for law students to determine the admissibility of evidence in court (Carr, 2003). Because findings such as these demonstrate the inconsistencies in using technology-based argumentation tools across varying domains, further research is necessary concerning the locus of *BRCA Gist* effects. Specifically, this research seeks to determine if argumentation used in an ITS can aid in understanding genetic risk for breast cancer and offer guidance towards making well-informed medical decisions.

While many sophisticated and modern models of argumentation have emerged, we have opted to use a version of the classic Toulmin (1958) model of argumentation. The elements of an argument identified by the Toulmin model provide a solid foundation and the terminology is more approachable for identifying elements of argument discourse. While the Toulmin model provides the foundation for identifying an argument, we have further refined this model within a cognitive context (see Wolfe & Britt, 2008; Wolfe, Britt & Butler, 2009). The argumentation model defines an argument as a claim that is supported by one or more reasons, with the claims and reasons being connected by warrants that are often implied but not stated explicitly (Wolfe, Britt, & Butler, 2009), and sometimes further supported by backing, as with appeals to more general principles and in some cases another argument (Voss, 2005). A claim is the conclusion that a person is trying to establish with "data" or reasons, typically factually grounded information. A warrant is a statement that logically follows the reasons and connects it to the claim, and is often not stated. An argument can also have counter-arguments and rebuttals, a counter-statement to the claim that is usually followed up by reasons why the counter-statement is not credible, and qualifiers expressing the degree of certainty about the claim in context (Toulmin, 1958; Wolfe, Britt, Petrovic, Albrecht & Kopp, 2009).

When developing parts of an argument it is clear that several cognitive processes are brought to bear, including logical and deductive reasoning. Kuhn (1993) theorizes that this type of reasoning is quite similar to scientific reasoning, in that it involves skills necessary in the development of critical thinking. The ability to think critically is a valuable skill in a learning environment because it allows a person to evaluate new information and integrate it into what they already know, a process known as knowledge-transforming (Voss & Van Dyke, 2001; Wiley & Voss, 1999). Along with this, Butler and Britt (2011) have shown that argument-type tutorials help students in qualitatively improving essays.

The motivation for the current research was to shed light on discrepancies in the literature on argumentation and learning,

including some of our own research findings. Wiley and Voss (1999) found that students who wrote arguments from web sources demonstrated superior understanding to students in comparison groups. However, this effect was relatively small and inconsistent across dependent measures. Wiley and colleagues concluded that the specific details of essay assignments were less important than the processes in which learners engaged. More recently, Goldman and colleagues (Goldman, Braasch, Wiley, Graesser, & Brodowinska, 2012) compared better and poorer learners on understanding information they read on-line and found that argumentation instructions were themselves not predictive of outcomes but that better learners engaged in more comprehension-monitoring and self-explanation on reliable websites as compared with unreliable websites. Part of the research conducted by Wolfe and colleagues (under review) compared an argument-interaction and no argument-interaction condition along with self-explanation and no self-explanation conditions in interactions with *BRCA Gist*. They found that randomly assigning participants to generate arguments did not help improve declarative knowledge, gist comprehension, or risk assessment ability, whereas asking participants to develop gist explanations was associated with significant improvements in key outcomes.

There are a number of possible explanations for these inconsistent findings about the relationship between generating arguments, learning, and comprehension. One possibility is that people typically produce good arguments in which claims are supported by warranted reasons, counter-arguments are presented and then rebutted, but the relationship between learning outcomes and argumentation is simply weak, inconsistent, or nonexistent. An alternative account is that even college educated people generally have difficulty generating complete arguments and that learning from argumentation is largely a function of argument quality. To address this question, we conducted a reliable fine grain analysis of over 400 argumentative tutorial dialogues between *BRCA Gist* and research participants arguing both for and against genetic testing for breast cancer risk. The present study examines the nuances of the argument interactions by identifying the number of argument elements contained in the interactions with *BRCA Gist* and determines whether participants did indeed make arguments. We examined whether including more elements of an argument in a *BRCA Gist* interaction leads to better performance on declarative knowledge, risk categorization, and gist comprehension tasks.

Hypotheses

There is already strong evidence for the efficacy of the *BRCA Gist* ITS (Wolfe et al., 2015). The goal of this research is to conduct a fine-grained analysis of argument-based interactions in *BRCA Gist*. Furthermore, we want to determine the effect these interactions have on understanding content information about genetic risk for breast cancer provided by the *BRCA Gist* tutorial, risk categorization ability, and understanding the gist of information. First, we hypothesize that there will be a difference between argument and explanation interactions in coverage of rubric items and content coverage (CO) scores given by *BRCA Gist*. Next, we hypothesize that argument scores for both arguments for and arguments against will correlate with coverage of tutorial content and that the outcomes on declarative knowledge scores, risk categorization, and gist comprehension will be positively correlated with argument scores. We will also assess whether participants actually produced arguments when asked to do so, with the minimal definition of an argument being a claim supported by one or more reasons, and determine if this had an effect on performance on any of the dependent measures of learning, comprehension, and decision making. We hypothesize that those participants who produced arguments will perform better on declarative knowledge, risk categorization, and gist comprehension measures than those whose verbal responses did not rise to the minimum definition of an argument. Participants were asked to produce two arguments, one for and another against genetic testing for breast cancer risk, so analysis will be further broken down within each of these argument prompts. Because people do not generally think about the "downside" of genetic testing, we predict there will be better outcomes when participants are asked to produce an argument for genetic testing due to the fact that these arguments are more well-known.

Experiment

Method

We analyzed verbal interactions between research participants and the most recent version of *BRCA Gist* from every study in which *BRCA Gist* helped participants generate arguments for and arguments against testing for genetic breast cancer risk. Over the course of these studies, 120 unique female participants over the age of 18 years and with no history of breast cancer were recruited from Miami University, Cornell University, the web, and an upstate New York community. These women were randomly assigned to interact with *BRCA Gist* while others were assigned to comparison conditions including the National Cancer Institute website, a control tutorial on nutrition and alternative versions of *BRCA Gist* with data from those conditions being excluded from the current analysis because those participants did not generate arguments. For all conditions the first half of the study consisted of 1.5 h of interaction with one of the four versions of the *BRCA Gist* tutorial, a National Cancer Institute (NCI) webpage, or a control tutorial (these will be discussed below) followed by several dependent measures. Respective conditions were

identical for all participants across all studies. Participants from the universities were given class credit for their Introductory to Psychology class while community and web participants received US$50. Upon completion, participants were thanked and debriefed.

In the first study, which took place at the universities, participants were randomly assigned to one of three conditions. The conditions included the *BRCA Gist* tutor, the NCI condition containing NCI webpages, and a tutorial on nutrition that served as a time-on-task control that utilized an SKO tutorial that contained nutrition and health information.

Similar to the first study, participants in the web and community study were randomly assigned to one of the three conditions. Web participants were recruited through numerous targeted websites with women who would benefit from gaining knowledge about breast cancer and genetic testing. Web participants could complete the study at any computer across the country that had internet access, while community participants were brought into the laboratory to complete the study.

Participants completed the argumentation dialogues at the end of the *BRCA Gist* tutorial session. The NCI and nutrition tutorial conditions data were not used in the present analyses because they do not contain the necessary argument interaction data (for a more indepth examination of the effectiveness of *BRCA Gist* refer to Wolfe and colleagues, 2015, and see Widmer and colleagues, 2015, for an analysis of gist explanation dialogues).

The dependent measures include a 52-item multiple-choice declarative knowledge test, a 12-item risk categorization task, and a 30-item gist comprehension task. Other dependent measures such as anxiety and worry about breast cancer risk are beyond the scope of this investigation and not reported here (see Wolfe et al., 2015 for details). The declarative knowledge test included factual knowledge presented in the *BRCA Gist* tutorial about breast cancer, genetic testing, as well as genetic risk. The knowledge test was created by the researchers as a means of testing knowledge presented in *BRCA Gist*; and has a Cronbach's α of 0.88 (Wolfe et al. 2015). The risk categorization task presented participants with 12 scenarios of women with varying degrees of genetic risk for breast cancer based on Pedigree Assessment Tool (PAT; Hoskins, Zwaagstra, & Ranz, 2006) scores an individual would get, including none, medium, and high. Scenarios were of similar length and included a similar amount of relevant and irrelevant information. Participants read the scenario then categorized the woman as having low risk, medium risk, or high risk. Correct categorization meant risk was accurately assessed. Developers of the PAT have found it to be a reliable tool in identifying genetic risk for breast cancer (Hoskins et al., 2006). The gist comprehension task utilized a 1- to 7-point Likert scale, with 1 being strongly disagree and 7 being strongly agree, and is intended to assess gist understanding of important bottom line meaning of the information about breast cancer and genetic testing. While these items are presented using a Likert scale rating of agreeing with the statement, the items do have independently verifiable correct answers and are not opinions. That is, the more agreement a participant shows to a statement such as: "The base rate for BRCA (breast cancer) genetic mutations is low" demonstrates greater gist comprehension, whereas a lower agreement demonstrates poorer gist comprehension because the statement is true. Furthermore, a little less than half of the items were reverse-scored in order to prevent response bias from occurring when participants answer "Strongly Agree" to all items. The gist comprehension task was created by the researchers to specifically measure gist understanding of material presented by *BRCA Gist*, and has a Cronbach's α of 0.85 (Wolfe et al., 2015).

Participants' argumentation dialogues (interactions with *BRCA Gist*) were assessed by independent raters as well as by *BRCA Gist*. These two methods of assessing the dialogues of participants' interactions with *BRCA Gist* allow us to determine not only if participants are engaging in a meaningful interaction, but also ensures that the *BRCA Gist* ITS is accurately providing feedback and assessing knowledge during interactions. Dialogues were assessed by independent raters using a reliable rubric (see Appendix A for rubrics). The rubric was created using the text from the ideal answer used in *BRCA Gist*. These ideal answers were written by research assistants to reflect the most thorough possible answer that a person could write in response to each interaction with BRCA Gist given the information presented in the tutorial. An ideal answer touched upon all of the key points covered in BRCA Gist. These answers were used to create the rubrics. From these ideal answers, individual elements of relevant information such as cost were identified and included in the rubric as separate items. There were seven rubric items for the question "What is the case for genetic testing for breast cancer risk?" and 17 rubric items for the question "What is the case against genetic testing for breast cancer risk?" Raters reviewed argumentation dialogues and gave participants a point when a gist understanding of one of the rubric items was stated. In other words, participants did not have to provide a verbatim statement in the argumentation dialogue that matched a rubric item precisely. Instead demonstrating an understanding of the bottom-line gist meaning was awarded a point. Two independent raters scored about one-third of the argumentation dialogues together to discuss any disagreements and came to a consensus about what type of statements demonstrated a gist understanding. We used a conditional scoring procedure such that an item was counted in the denominator only if it was marked as present by at least one rater. This procedure avoids inflating agreement based on items that are clearly absent in participants' dialogues, which account for the majority of judgments. The percent of agreement between raters' scores

produced an inter-rater reliability score of .80 or greater on both argumentation interactions. The same method was employed for the five dialogues which utilized an explanation interaction (Widmer, et al., 2015). Because these explanation interactions are only used as a comparison to argumentation interactions and not the focus of this manuscript, the rubrics are not included. However, inter-rater reliability scores for the explanation interactions were greater than .87 (for a more detailed discussion of the explanation interactions see Widmer et al., 2015).

Because *BRCA Gist* can provide feedback to participants, it must assess the accuracy of their knowledge during an interaction. To do this, *BRCA Gist* uses LSA to determine the amount of semantically related information given by participants when compared to an ideal answer as identified by the researchers. The amount of semantically similar text is given a Coverage Score (CO Score) from 0 to 1, with higher scores indicating better coverage of the preprogrammed text. Furthermore, using CO scores *BRCA Gist* determined which preprogrammed feedback (i.e., a hint, prompt, or pump) participants should receive to guide them towards a more complete answer, thus a higher CO score. Researchers developed preprogrammed text by identifying the most valid information in response to the questions "What is the case for genetic testing for breast cancer risk?" and "What is the case against genetic testing for breast cancer risk?" allowing us to create expectations for each question.

A separate rubric was also created to assess the strength of the argument interactions. Using the elements of an argument outlined by Wolfe, Britt, and Butler (2009), an argument rating system was created for the argument interactions using the elements of an argument (see Appendix B). Raters scored arguments using a 0–4 scale. A score of 0 indicated no argument elements were included or irrelevant information was present. A score of 1 indicated that one or more reasons were simply listed or stated with no attempt to link them to claims. A score of 2 indicated that both a claim and supporting reasons were included, with the warrant connecting the claim and reasons implied. Thus, a score less than 2 indicated that the response failed to meet the minimum definition of an argument (a claim supported by reasons; Toulmin, 1958; Wolfe & Britt, 2008, Wolfe, Britt & Butler, 2009). A score of 3 included at least three argument elements that were implied, such as counter-argument, rebuttal, or backing (warrants were excluded from the list of elements that could be implied to receive a score of 3 since the inclusion of a claim supported by reasons already necessitates at least an implied warrant). Finally, a score of 4 indicated that at least three argument elements that were explicitly stated (including an explicit statement of the warrant). Table 1 contains the percent of participants that received each argument rating. Providing multiple or various claims, reasons, warrants, counterarguments, or rebuttals and backing was not awarded a higher argument rating, so that responses that consisted of multiple simple arguments did not receive the same score as response that contained better developed complex arguments. Two raters trained on 40 interactions to discuss any disagreements and come to a consensus about what was considered an element of an argument. Inter-rater reliability scores were computed for 60 additional random interactions by taking the number of scores agreed on and dividing by the total number of scores. Using the same conditional scoring procedure outlined above, an inter-rater reliability score of .89 was achieved for the "argument against" interaction and a score of .80 was achieved for the "argument for." The remaining interactions were rated by one of the raters.

Table 1 Percent of responses with each argument rating score

	Argument score				
	0	1	2	3	4
Argument for	1.3 %	52 %	21.3 %	22.7 %	2.7 %
Argument against	2.3 %	57 %	8.1 %	25.6 %	7 %

Below is a sample of an argument score of 1 when *BRCA Gist* asked a participant "What is the case against genetic testing for breast cancer risk?" In this interaction, the participant only listed reasons genetic testing could be bad and did not include any other elements necessary for producing an argument.

> Participant: Even if you test negative, it does not mean you do not have risk of getting cancer.
> Participant: If your results are positive, it will be very hard to tell your family the news.
> *BRCA Gist*: Can you give me some reasons why someone would not want to be tested?
> Participant: Testing positive can lead to anxiety and other emotions.
> *BRCA Gist*: Can you make any arguments against testing using genetic discrimination?
> Participant: Confidentiality may also not be kept.

The next sample received an argument score of 3 when asked by *BRCA Gist* "What is the case for genetic testing for breast cancer risk?" First, the overall interaction implies the claim that genetic testing is beneficial and a person should have it done. The response also listed many reasons against genetic testing. Finally, in the last response, there is an implied rebuttal that there are laws in place to protect someone from genetic discrimination.

> Participant: Getting tested for genetic testing will allow you to plan your future so that you are prepared for possible chances of developing breast cancer.

BRCA Gist: You are off to a good start. Please keep making a case in favor of testing.
Participant: Your test result could also allow other family members to realize possible risks of breast cancer development since family history is an important factor in calculating risk.
Participant: There are also some emotional benefits to being tested for breast cancer; if your test is negative, you may feel relieved.
BRCA Gist: Good job. Can you talk about any benefits of testing if the results turn out negative?
Participant: Also, if your test result is negative, it is possible that you do not have to undergo any surgery, limiting the amount of treatment costs.
Participant: Additionally, the GINA protects you from being discriminated by employees even if your test result comes out positive.
BRCA Gist: Very good. Please click on the finish button to continue.

Results

Using the rater's rubric scores we examined the differences between coverage scores of rubric items in explanation interactions and argument interactions. There were five explanation interactions (Widmer et al., 2015) and two argument interactions in the *BRCA Gist* tutorial. In order to test whether differences in the number of rubric items covered between explanation and argumentation interactions were present, the average rubric score for the five explanation interactions was computed, as well as the two argumentation interactions. The two averages were then compared using a paired-samples t-test. A significant difference between explanation type (M = 19.87, SD = 14.95) and argumentation type (M = 25.89, SD = 21.55) interactions in their mean coverage of rubric items were found, t (116) = 4.66, p = < .01. Argument interactions covered more rubric items than explanation interactions. We also averaged then compared the CO scores generated by *BRCA Gist* for each interaction type. A paired-samples t-test revealed a significant difference between gist explanation type (M = .38, SD = .2) and argument type (M = .35, SD = .22) interactions in their mean CO scores, t (94) = −2.11, p = .038. In this case, the *BRCA Gist* tutor did a better job detecting coverage for gist explanations than arguments.

Correlations were conducted between argument ratings, rubric coverage scores, CO scores, and all dependent measures. Rubric coverage scores were significantly correlated with CO scores given by *BRCA Gist*, r = .854, p < .01 for the argument for interaction r = .870, p = .01 for the argument against interaction. This suggests that *BRCA Gist* was sensitive to the degree to which expectations for content were met in in the argumentation interactions. Additionally, as can be seen in Table 2, CO scores for both the argument for and argument against interactions were significantly correlated with the three dependent measures. Thus, the more arguments for and against genetic testing included expected content, the better participants did on post-test measurers of declarative knowledge, comprehension, and risk assessment. The argument ratings in the argument for genetic testing did not produce any significant correlations. However, argument ratings for the case against genetic testing had small to medium significant correlations with the "con side" CO score, r = .533, p = .01, rubric coverage scores, r = .379, p = .01, declarative knowledge scores, r = .324, p = .01, and risk categorization scores, r = .267, p = .05. The greater the extent to which participants made elaborate arguments against genetic testing including elements of argumentation such as warrants, counterarguments, and rebuttals, the better they did on measurers of knowledge and risk assessment.

Argument rubric scores were classified as to whether participants made an argument or not. Scores of 0 or 1 were classified as not an argument while 2, 3, and 4 were considered an argument because they contained at least minimal argument elements (e.g., a claim supported by reasons with the warrant implied if not stated; Wolfe & Britt, 2008; Wolfe, et al., 2009). This comparison was broken into argument versus not an argument, rather than comparing the different scores, for two reasons. First, there was not an even distribution of argument scores (see Table 1), disrupting the homogeneity of variance assumption for conducting an ANOVA. Second, we were particularly interested in examining outcomes for those who made an argument compared to those who did not. We view the question of whether an argument was made (i.e., the difference between a score of 1 and 2) to be fundamental whereas the additional presence of argumentation elements such as rebuttal, counter-argument, backing, and qualification (i.e., the difference between a score of 2 and 3) to be of secondary interest. When asked to produce arguments for genetic testing, an independent samples t-test revealed no significant differences between those classified as no argument and argument on declarative knowledge test, t (70) = −1.4, p = .167, the risk categorization task t (70) = −.27, p = .787, and the gist comprehension task t (70) = −.85, p = .4. See Table 3 for all means and standard deviations of argument classifications.

Similar to the argument for, the argument against rubric scores were classified based on whether participants produced an argument or not. An independent-samples t-test revealed a significant difference between those that made an argument (M = .82, SD = .08) and those that did not make an argument (M = .73, SD = .19) on the declarative knowledge test, t (80) = −2.5, p = .015. Those that made an argument had better knowledge of genetic risk than those that did not make an argument. Furthermore, the analysis approached significance for the risk categorization, t (79) = −1.86, p = .067. There was

Table 2 Correlations of argument scores, coverage scores, and dependent measures

	Argument ratings for	Argument ratings against	Rubric coverage for	Coverage score for	Rubric coverage against	Coverage score against	Declarative knowledge	Risk category.	Gist composition
Argument ratings for	1	.237**	.089	.111	.140	.313*	.206	.114	.219
Argument ratings against		1	.318*	.377*	.379*	.533*	.324*	.267**	.170
Rubric coverage for			1	.854*	.821*	.761*	.323*	.236*	.311*
Coverage score for				1	.748*	.775*	.430*	.418*	.438*
Rubric coverage against					1	.870*	.335*	.183*	.376*
Coverage score against						1	.469*	.353*	.564*
Declarative knowledge							1	.620*	.775*
Risk category								1	.459*
Gist composition									1

*Correlation significant at $p < .01$

**Correlation significant at $p < .05$

no significant difference on gist comprehension tasks, $t\,(78) = -1.19$, $p = .24$. See Table 3 for all means and standard deviations of argument classifications.

Discussion

These findings suggest that the *BRCA Gist* instructions to produce an argument for and against genetic testing failed to get people to do so. Fewer than half of the argumentation dialogues resulted in verbal responses that met the minimal criteria for being classified an argument (Wolfe, Britt, & Butler, 2009). However, participants who actually did generate an argument against genetic testing did outperform those who did not on the declarative knowledge test. Furthermore, when responses were classified as either argument or no argument, significant differences were found for the declarative knowledge test when participants made an argument against genetic testing. In these cases, those that were classified as producing an argument performed better than those that did not produce an argument. This is consistent with the notion that many participants lacked the requisite argumentation skills to produce arguments, but those who did benefited from making elaborate arguments in the case against genetic testing for breast cancer risk.

Argument ratings, in the argument against genetic testing for breast cancer risk, but not the argument for testing, had a positive correlation with the declarative knowledge and risk categorization outcome measures. However, rubric coverage and CO scores for both the argument for and against were correlated with all outcome measures. Argument ratings in the argument against were also correlated with both rubric coverage and CO scores. Thus, when people actually produced arguments against genetic testing they improved on measures of knowledge and risk assessment. Furthermore, the more argumentation elements in the argument against genetic testing that were included in their dialogues, the better the performance on outcome measures.

It is interesting to note that most of the significant findings from these studies were found in the argument against genetic testing interactions. One possible reason for this is that there could be a fundamental difference when one has to make a case for or against a position. An alternative reason is that arguments against genetic testing require participants to draw upon more unfamiliar decision-relevant information, which would cause participants to gain more knowledge in this interaction drawing attention to less familiar information, compared to the more familiar information needed for arguments for genetic testing. For example, the arguments for genetic testing are relatively well known and

Table 3 Mean scores and standard deviations of outcome measures by argument classification

	Argument for			Argument against		
	Declarative knowledge Percent correct	Risk categorization Percent correct	Gist comprehension (1 low – 7 high)	Declarative knowledge Percent correct	Risk categorization Percent correct	Gist comprehension (1 low – 7 high)
Argument	.76 (.13)	.60 (.13)	5.37 (.62)	.82 (.08)	.63 (.12)	5.68 (.68)
No argument	.76 (.15)	.61 (.13)	5.31 (.64)	.73 (.19)	.57 (.16)	5.52 (.68)

obvious (such as gaining additional information about risk) whereas arguments against testing are less intuitively obvious to laypeople (such as those involving cost, the potential for genetic discrimination, and conflict with family members).

Some argue that the effective element of tutorial dialogues may simply be the interaction between the tutor and the learner, the I-hypothesis put forth by Chi, Siler, Jeong, Yamauchi, and Hausmann (2001). Essentially, the I-hypothesis claims that tutoring is successful specifically because of the interaction between the tutor and the student, and this success cannot be isolated to the actions of either party alone. This type of scaffolded interaction allows for meaningful dialogues, elaboration of content, and checking for comprehension of material (Chi et al., 2001, 2008; Pressley et al., 1992; Roscoe & Chi, 2008). Indeed, the interactive dialogues in *BRCA Gist* aided participants in discussion and coverage of more content. Our findings demonstrate that simple interaction is insufficient to produce learning gains, however content coverage as measured by coverage of rubric items and ideal answers to questions posed by the ITS is associated with better learning outcomes. Specific findings regarding producing well-structured arguments during argument against interactions had more modest associations with improved learning outcomes.

There are some shortcomings of the present research. It may not have been clear to some participants that *BRCA Gist* was asking for formal arguments to be made. Sometime before the interaction it might also prove useful to provide a general outline of how to make an argument, which may have encouraged more participants to generate well developed arguments in their responses. It also should be noted that there were five explanation interactions and only two argument interactions. This means that there was much more content covered in the explanation interactions reflected in the dependent measures. It would be of interest to see if making an argument had better long-term effects with a follow-up a few weeks or months later. Alternatively, due to the fact that there is still a debate amongst argumentation discourse researchers, the method in which the argument interactions were coded may have led to the finding that the majority of participants failed to construct arguments. However, we used a clearly defined and reliable rubric in order to operationalize what we considered an interaction that contained a true argument. Furthermore, this rubric and the associated elements of an argument were informed using empirically vetted definitions of an argument (Toulmin, 1958; Wolfe & Britt, 2008, Wolfe, Britt & Butler, 2009). While some may disagree with our definition of an argument, we feel this is an adequate first step towards assessing the effectiveness of argumentation by an ITS.

Previous research from controlled experiments using the *BRCA Gist* ITS strongly suggests that it improves knowledge, comprehension, and risk assessment (Wolfe et al., 2015). There is also clear evidence that the *BRCA Gist* tutorial dialogues, which focus on self-explanation in the form of gist explanations, lead to appreciable learning gains (Widmer et al., 2015; Wolfe et al., under review). The results of this fine grained analysis of *BRCA Gist*'s argumentation dialogues indicates that most participants did not produce arguments when prompted to do so. However, to the extent that they did produce full arguments against genetic testing, participants improved on key outcome measures. A challenge for future research and development efforts is to find better ways for this ITS to encourage argumentation which may produce larger gains in knowledge, comprehension, and risk assessment. A challenge for future research on argumentation in ITS's is to implement automated scoring of tutorial dialogues by adapting tools such as Coh-Metrix (Graesser, McMamara, Louwerse, & Cai, 2004). Reliable fine-grained analyses of tutorial dialogues are potentially fruitful, but also labor intensive and the results of our analyses suggest that such an approach is warranted.

Author Note This project 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. We also wish to thank Rachel Aron, Andrew Circelli, Cecelia Favede, Jeremy Long, Mitch McDaniel, Ian Murphy, Kendall Powell, Michael Thomas, and Audrey Weil for their capable assistance with data collection.

Appendix A

Rubric of propositions in essays/responses

Directions for Gist scoring. Assign one point if the participant's answer captures the essence of the statements below even if they use different words.

What is the case for genetic testing for breast cancer risk?

Case for testing

1. There are reasons to undergo genetic testing (credit any reason for testing).

Pro case negative result

2. A negative result can ease the patient's anxiety, especially if a woman is worried and looking for a sense of relief.
3. A negative result may allow the patient to save time and/or money.

4. A negative result may reduce unnecessary tests, exams, and surgery (credit any).

Pro case positive result

5. A positive result enables the woman to have a full understanding of her risks and to take action accordingly. If the woman finds out she is genetically predisposed to develop cancer she can take certain measures to prevent that from happening (credit any appropriate action).
6. If a woman testes positive, it may help other relatives who may have these harmful mutations.
7. Women with positive results may be able to participate in clinical trials and research.

What is the case against genetic testing for breast cancer risk?

Case against testing

1. There are reasons to avoid genetic testing (credit any reason against testing).
2. There is a very small direct medical risk from drawing blood (however it would be a misconception if this was presented as a large risk).
3. There is a chance of receiving ambiguous results. About 10 % of women who undergo genetic testing for BRCA1 and BRCA2 receive an ambiguous result. Include inaccuracy of test.
4. Genetic tests are expensive tests that may not be covered by health insurance.
5. There may be better uses of the money that testing costs for health maintenance or other purposes (mention other better uses of money).
6. The number of women who have harmful BRCA mutation is small (0.1–1 %) thus the probability of a positive result is small.
7. There can be illegal genetic discrimination from employers or prospective employers because they may have a higher risk of developing cancer.
8. Counterargument: In 2008, the Genetic Information Nondiscrimination Act was passed to protect American citizens from genetic discrimination in relation to health insurance and employment. Don't need to be specific about type of discrimination, only need to mention discrimination legislation
9. The Genetic Information Nondiscrimination Act does not cover life insurance, disability insurance, long-term health insurance, or members of the military (credit any of these missing from the bill or as a problem). Also include results may lead to insurance discrimination. Any discrimination

Con case negative result

10. A woman receiving negative results may feel "survivor guilt" as well as a sense of isolation from less fortunate family members who have cancer or test positive for BRCA mutations.

Con case positive result

11. A woman who receives a positive result will probably have increased anxiety, stress, depression, and/or anger (credit any psychological issues).
12. A woman who receives a positive result may choose to explore questionable or inappropriate alternative treatments.
13. A woman who receives a positive result may also choose to undergo preventive measures (even prophylactic surgery) whose effectiveness is uncertain or has serious long-term implications.
14. A positive result may have an impact on several family members (who didn't want testing or knowledge of BRCA mutations) and can create tension within families (must be negative).
15. A positive result may have a negative impact on personal choices, such as marriage and childbearing (credit for self or family members).
16. Information gathered in a person's medical record from positive results could adversely impact purchase of some forms of insurance.
17. Positive test results in a person's medical records may not be kept private or confidential.

Appendix B

Content Coverage and Argumentation Rubrics for Pro and Con Arguments

Scores

0. Irrelevant information. Not even reasons are stated
1. Reason(s) stated
2. Re-statement of claim, reason(s), (warrant is implied NOT stated) (gist of claim: testing is good/advantageous, etc.)
3. At least three argument elements implied: claim, grounds/reason, rebuttal, counter argument, backing, qualification (excluding warrant)
4. Three argument elements explicitly stated

Argument elements:

Claim
Grounds/reasons
Warrant
Backing
Rebuttal
Counter-argument
Qualification

References

Butler, J. A., & Britt, M. A. (2011). Investigating instruction for improving revision of argumentative essays. *Written Communication, 28,* 70–96. doi:10.1177/0741088310387891

Carr, C. S. (2003). Using computer supported argument visualization to teach legal argumentation. In P. A. Kirschner, S. J. Buckingham Shum, & C. S. Carr (Eds.), *Visualizing argumentation: Software tools for collaborative and educational sense-making* (pp. 75–96). London, UK: Springer.

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., Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439–477.

Chi, M. T. H., Roy, M., & Hausman, R. G. M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science, 32,* 301–341. doi:10.1080/03640210701863396

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

Clariana, R. B., & Koul, R. (2006). The effects of different forms of feedback on fuzzy and verbatim memory of science principles. *British Journal of Educational Psychology, 76,* 259–270.

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.

Easterday, M. W., Aleven, V., & Scheines, R. (2007). 'Tis better to construct or to receive? Effect of diagrams on analysis of social policy. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AI-ED 2007)* (pp. 83–100). Amsterdam, The Netherlands: IOS Press.

Glaser, R., & Bassok, M. (1989). Learning theory and the study of instruction. *Annual Review of Psychology, 40,* 631–666. doi:10.1016/0364-0213(89)90002-5

Goldman, S. R., Braasch, J. L. G., Wiley, J., Graesser, A. C., & Brodowinska, K. (2012). Comprehending and learning from internet sources: Processing patterns of better and poorer learners. *Reading Research Quarterly, 47,* 356–381.

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

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005a). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education, 48*(4), 612–618.

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., 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.

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

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

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Tutoring Research Group, T. R. G., & 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., Han, L., & Cai, Z. (2008). *Semantic decomposition of student's contributions: an implementation of LCC in AutoTutor Lite*. Paper presented to the Society for Computers in Psychology, Chicago, Illinois: November 13, 2008.

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education, 77,* 319–337.

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development, 74,* 1245–1260.

Lloyd, F. J., & Reyna, V. F. (2009). Clinical gist and medical education: Connecting the dots. *Journal of American Medical Association, 302,* 1332–1333. doi:10.1001/jama.2009.1383

Ohlsson, S. (1986). Some principles of intelligent tutoring. *Instructional Science, 14,* 293–326. doi:10.1007/BF00051825

Pressley, M., Wood, E., Woloshyn, V. E., Martin, V., King, A., & Menke, D. (1992). Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychologist, 27,* 91–109. doi:10.1207/s15326985ep2701_7

Reyna, V. F. (2008). A theory of medical decision making and health: Fuzzy trace theory. *Medical Decision Making, 28,* 850–865. doi:10.1177/0272989X08327066

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., Estrada, S. M., DeMarinis, J. A., Myers, R. M., Stanisz, J. M., & Mills, B. A. (2011). Neurobiological and memory models of risk 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., & 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.

Roscoe, R. D., & Chi, M. T. H. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science, 36,* 321–350. doi:10.1007/s/11251-007-9034-5

Scheuer, O., Loll, F., Pinkwart, N., & McLaren, B. M. (2010). Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning, 5,* 43–102.

Toulmin, S. (1958). *The uses of argument*. New York, New York: Cambridge University Press.

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

Voss, J. F. (2005). Toulmin's model and the solving of ill-structured problems. *Argumentation, 19,* 321–329.

Voss, J. F., & Van Dyke, J. A. (2001). Argumentation in psychology: Background comments. *Discourse Processes, 32,* 89–111. doi:10.1080/0163853X.2001.9651593

Widmer, C. L., Wolfe, C. R., Reyna, V. F., Cedillos-Whynott, E. M., Brust-Renck, P. G., & Weil, A. M. (2015). Tutorial dialogues and gist explanations of genetic breast cancer risk. *Behavior Research Methods, 47,* 632–648.

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91,* 301–311. doi:10.1037/0022-0663.91.2.301

Wolfe, C. R., & Britt, M. A. (2008). Locus of the my-side bias in written argumentation. *Thinking and Reasoning, 14,* 1–27.

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

Wolfe, C. R., Britt, M. A., & Butler, J. A. (2009b). Argumentation schema and the myside bias in written argumentation. *Written Communication, 26,* 183–209.

Wolfe, C. R., & Cedillos, E. M. (2015). E-Learning and E-Communications Platforms. In *The international encyclopedia of the social and behavioral sciences* (2nd ed., pp. 895–902). Philadelphia, PA: Elsevier.

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

Wolfe, C. R., Reyna, V. F., Brust-Renck, P. G., Weil, A. M., Widmer, C. L., Cedillos, E. M., …, Circelli, A. M. (2013). *Efficacy of the BRCA Gist intelligent tutoring system to help women decide about testing for genetic breast cancer risk*. Paper presented to the 35th Annual Meeting of the Society for Medical Decision Making, Baltimore, MD.

Wolfe, C. R., Reyna, V. F., Widmer, C. L., Cedillos, E. M., Fisher, C. R., Brust-Renck, P. G., …, Damas Vannucchi, I. (2013). *Efficacy of a web-based intelligent tutoring system on genetic testing for breast cancer risk*. Presentation to the *6th* Annual Scientific Meeting of the International Society for Research on Internet Interventions, Chicago, IL.

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

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.