



Pumps and Prompts for Gist Explanations in Tutorial Dialogues About Breast Cancer

Christopher R. Wolfe^a, Valerie F. Reyna^b, Colin L. Widmer^c, Elizabeth Cedillos-Whynott^c, Audrey M. Weil^c, and Priscila G. Brust-Renck^b

^aDepartment of Psychology, Miami University; ^bDepartment of Human Development, Cornell University; ^cDepartment of Psychology, Miami University

ABSTRACT

Fuzzy-Trace Theory (FTT) generalizes research on discourse to predict how health messages can be better understood and remembered, thereby influencing decision making. Applying FTT, *BRCA Gist* delivers messages interactively through tutorial dialogues and is the first Intelligent Tutoring System designed to help laypeople make sound medical decisions. Previous research indicates that *BRCA Gist* helps people form useful “gist explanations,” which leads to improved knowledge, comprehension, and risk assessment. The present research examined the effectiveness of different *BRCA Gist* dialogue moves, including general pumps for information and prompts for specific information. Participants were randomly assigned to a control group or one of four *BRCA Gist* conditions evoking gist or verbatim representations crossed with general pumps or specific information prompts. Gist-evoking pumps by themselves produced significant gains in knowledge and risk assessment. Specific verbatim prompts increased knowledge without affecting risk assessment. Results are explained in light of memory research and FTT principles.

Introduction

Research on discourse processes is highly relevant to the communication of health messages designed to facilitate decision making, but such research is underused. In the present work, we apply a theory grounded in large part in research on memory and comprehension of discourse to the communication of health risks and benefits of genetic testing for cancer (e.g., Reyna & Kiernan, 1994; Singer & Remillard, 2008). According to the theory, levels of representation that were initially identified in research on discourse, namely, gist and verbatim, are encoded from health messages but have different effects on decision making (e.g., Reyna, 2008b). Varying general and specific retrieval cues in the context of tutorial dialogues between users and avatars can thus be used to examine different kinds of learning from discourse and their consequences for decision making.

Advances in discourse technologies (Graesser, 2011) are beginning to allow everyday people to engage in dialogue with intelligent agents to help them understand complicated discourse and to make decisions in accordance with their individual values, goals, and constraints. How best to structure those dialogues is a fundamental and complex question. Much of what we know about engineering verbal interactions among people and intelligent agents (Cohn, Olde, Bolton, Schmorow, & Freeman, 2015) comes from research on one-on-one human tutoring (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Roscoe & Chi, 2008) and on Intelligent Tutoring Systems (ITS) in academic domains (Arnott, Hastings, & Allbritton, 2008; Craig, Sullins, Witherspoon, & Gholson, 2006; Graesser, Chipman, Haynes, &

CONTACT Christopher R. Wolfe, Ph.D. ✉ Wolfecr@MiamiOH.edu Department of Psychology, Miami University, Oxford, OH 45056. Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hdsp.

Olney, 2005; Graesser et al., 2004; VanLehn, Graesser, Jackson, Jordana, Olneyb, & Rose, 2007; Sottolare, Graesser, Hu, & Brawner, 2015). One-on-one human tutoring is arguably the best way to produce deep conceptual understanding (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2005). However, effect sizes for common metrics of learning are approximately the same for artificial intelligence (i.e., the best ITSs) as for human tutoring, both typically producing effect sizes of about .8 (VanLehn, 2011).

Perhaps the most effective feature of discourse with human tutors is the ability to interact and encourage people to generate self-explanations of what they have learned (Chi, Leeuw, Chiu, & LaVancher, 1994). Effective tutors offer feedback and ask questions to get people to elaborate on these self-explanations beyond the initial answer (McNamara, Jacovina, Snow, & Allen, 2015; Roscoe & Chi, 2008; VanLehn et al., 2007). This helps people connect new insights with information they already understand and helps reveal any misunderstandings (Chi, 2000). Graesser et al. (2005) persuasively argue that “learning is more effective and deeper when the learner must actively generate explanations” (p. 612).

One explanation of why individual tutoring is so effective is the Interaction Hypothesis (e.g., Chi et al., 2001; Graesser, Person, & Magliano, 1995; van de Sande & Greeno, 2010) that the more a learner and tutor build on one another’s dialogue moves, the greater the learning. However, more recent research suggests that it is not how much interaction takes place in a tutoring session tutoring that is crucial but, rather, the quality of those interactions (e.g., Chi et al., 2010, 2011). Snow, Allen, Jacovina, and McNamara (2015) discovered a link between user control or agency and the quality of self-explanations in a gaming environment. Graesser and colleagues have found that in helping people form self-explanations in academic domains, it is more effective to focus on causal mechanisms and the functional underpinnings of complex systems (Graesser et al., 2005).

Graesser (2011) notes that specific dialogue moves that an ITS can make to effectively promote self-explanation include *pumps* that encourage learners to express more information (e.g., “What else?”), *hints* that guide the learner to express ideas that are important answers to the main question or problem (e.g., in a physics problem, “What about the forces of two colliding vehicles on each other?”), and *prompts* that guide the learner to fill in missing information in an explanation (e.g., “The forces of the two vehicles on each other are equal in what?”). Thus, an important distinction between pumps and prompts is the level or specificity with pumps asking for more information at a general level and prompts asking for specific information.

The research on ITS in academic domains is important in helping us understand and construct discourse technologies to help people make decisions. However, there are also fundamental differences between discourse for academic learning and discourse for learning and comprehension in the context of decision making. With respect to the decision about whether to undergo genetic testing for breast cancer risk, we argue that it is not so important for laypeople to understand the causal mechanisms and functional underpinnings of blood tests themselves, but it is very important to understand the possibility of ambiguous results, the conditional probabilities of developing breast cancer with positive and negative test results for *BRCA1* and *BRCA2* (BRest CAncer 1 and 2) genetic mutations, and the base rates for breast cancer and *BRCA* mutations in the general population and what they would do in the event of a positive test result (Wolfe et al., 2015). As is the case in academic learning, dialogues supporting decision making are most likely to be fruitful when decision makers actively generate explanations. We argue that discourse technologies for decision making should focus on the consequences of decision alternatives, that pumps and prompts should focus on decision-relevant dimensions, and that systems should encourage “gist explanations” in the users own words rather than precise verbatim information. Below, we elaborate on these ideas with respect to decisions about genetic testing for breast cancer risk.

Decisions about genetic testing for breast cancer risk

About one in eight American women will develop breast cancer sometime during their lifetimes (National Cancer Institute, 2012), and many healthy women contemplate genetic testing for breast cancer risk. Yet, decisions about whether to be tested for genetic risk of breast cancer are challenging.

Unfortunately, patients and healthcare providers exhibit systematic biases in estimating such risks (Offit, 2006; Reyna, Lloyd, & Whalen, 2001; Reyna, Nelson, Han, & Dieckmann, 2009). Genetic tests for mutations in the *BRCA1* and *BRCA2* genes (as well as newer tests for genes accounting for a smaller portion of breast cancer cases) are expensive and often not covered by insurance unless there is a clear family history of ovarian and breast (Agus, 2013; Andrews, 2013). There are only about 3000 genetic counselors in the United States, (Karow, 2013). Thus, there are simply not enough genetic counselors to talk with every woman considering genetic testing for breast cancer risk.

Genetic testing for breast cancer risk can save lives. However, from a medical perspective, most women are not good candidates for testing due to the low base rate, expense, and relatively high rate of ambiguous results. Moreover, interest in genetic testing does not always correspond with assessed medical risk, and low-risk women are unlikely to consider potential negative implications of testing. Many patients are unsure what they would do if they received positive, negative, or ambiguous results. Those receiving positive results, however, must decide about measures such as tamoxifen treatments, more frequent mammograms, screening for ovarian cancer, and prophylactic mastectomy (Armstrong, Eisen, & Weber, 2000; Chao et al., 2003; Stefanek, Hartmann, & Nelson, 2001). Additionally, negative results do not guarantee freedom from cancer.

“Traditional approaches to risk communication in health care and medical decision making emphasize telling the facts in order to facilitate informed decisions when risk is involved” (Brust-Renck, Royer, & Reyna, 2013, p. 244). There have been thoughtful efforts to include patients in shared decision making (Col, 2011). Efforts to improve patient understanding include storytelling (Shaffer, Hulsey, & Zikmund-Fisher, 2013; Shaffer & Zikmund-Fisher, 2013), improving the visual display of quantitative medical information (Ancker, Senathirajah, Kukafka, & Starren, 2006), and even patient assistance with diagnosis (Graedon & Graedon, 2014). The typical approach centers on simplifying information for patients, with admonitions such as “use plain language to make written and verbal materials more understandable” and “consider presenting only the information that is most critical to the patients’ decision making, even at the expense of completeness” (Fagerlin, Zikmund-Fisher, & Ubel, 2011). Unfortunately, patients and physicians typically have little time for adequate discussion. There are also a number of excellent resources for cancer patients including web pages by the National Cancer Institute (2016), the American Cancer Society (2016), and the Cleveland Clinic (2016). However, they were apparently not designed following recent advances in cognitive psychology to help everyday people form useful mental representations of complex content.

BRCA Gist

To address these issues, we developed an ITS called *BRCA Gist* (BREast Cancer and Genetics Intelligent Semantic Tutoring; Wolfe et al., 2015). *BRCA Gist* engages women in a dialogue about issues related to genetic testing for breast cancer risk (Armstrong, Eisen, & Weber, 2000; Berliner & Fay, 2007; Stefanek, Hartmann, & Nelson, 2001) and is the first ITS created to help patients make medical decisions. *BRCA Gist* was designed with input from medical experts using information on the National Cancer Institute website. The goal is to complement the clinical encounter between a patient and physician or genetic counselor, rather than replace it, or to be used as a freestanding public health resource for women (Wolfe et al., 2015).

BRCA Gist was created using a platform called AutoTutor Lite (Institute for Intelligent Systems, Memphis, TN; Ney, Graesser, & Hu, 2014; Wolfe, Fisher, Reyna, & Hu, 2012; Wolfe et al., 2016), a web-based version of AutoTutor (Institute for Intelligent Systems, Memphis, TN; Graesser, 2011; Graesser & McNamara, 2010). AutoTutor Lite is used to create ITSs that allows people to interact with it online using a web browser. AutoTutor Lite uses an artificial intelligence technique called semantic decomposition for natural language processing (Hu, Han, & Cai, 2008) to mimic one-on-one human tutoring (Graesser, 2011). Through dialogue with animated talking pedagogical agents, AutoTutor Lite guides the learner toward target expectations by soliciting verbal responses from the learner. AutoTutor Lite compares the text entered by a user with a set of “expectation texts” we developed using qualitative verbal data and refined through an iterative series of successive refinements (Wolfe et al., 2013). AutoTutor Lite uses

Latent Semantic Analysis (Graesser, Wiemer-Hastings, Wiemer-Hastings, & Harter, 2000) to compare sentences entered by users with an expectation text that embodies good answers so it can respond appropriately to different users depending on what they say. Latent Semantic Analysis permits *BRCA Gist* to assess how much progress participants are making in meeting those expectations for a good answer—even when people explain the gist of key concepts in their own words, deviating from the exact phrases in the expectations texts. AutoTutor Lite allowed us to create effective tutorial dialogues without a team of programmers needed to develop dialogues in other ITSs (Wolfe et al., 2015, under review). For a fuller technical description of *BRCA Gist* and AutoTutor Lite, see Wolfe and colleagues (2013).

Fuzzy-Trace Theory

BRCA Gist is guided by Fuzzy-Trace Theory (FTT), Reyna's (2008a) evidence-based theory of medical decision making. *BRCA Gist* uses three key ideas rooted in FTT to help people make good medical decisions. First, as explained below, gist-based interventions (Reyna, 2008b) improve knowledge, understanding, and decision-making in medical contexts. Second, helping people explain the gist of complex medical information *in their own words* fosters learning and comprehension (Lloyd & Reyna, 2009). Third, at least some tutorial dialogues should focus on understanding the consequences of decision alternatives (Wolfe et al., under review).

FTT is a dual-processes approach to judgment and decision-making (Barbey & Sloman, 2007; Evans, 2008; Reyna, 2004; Sloman, 1996). Research has indicated that adults encode multiple representations but primarily rely on gist in judgment and decision making. "Gist" refers to representations of the essential bottom-line meaning of information. At the time of encoding, people form parallel memory representations along a continuum from verbatim representations of specific superficial details to gist representations capturing the bottom-line meaning but shedding much irrelevant detail (Reyna, 2012; Reyna & Brainerd, 2011; Reyna, Nelson, Han, & Pignone, 2015). In other words, a key FTT concept is that when making decisions (including medical decisions), people encode gist and verbatim representations but prefer to reason with the vaguest bottom-line gist that can be used to decide among options (Reyna, Chick, Corbin, & Hsia, 2014; Wilhelms & Reyna, 2013). This fuzzy processing preference *increases* rather than decreases with expertise, age, and experience (Reyna, 2008a; Reyna & Lloyd, 2006) because gist captures meaning (rather than rote memorization as in verbatim representations), and it is often more useful to rely on such simple meaningful representations (Reyna & Mills, 2014) that encapsulate decision-relevant information. Patients who make better health and medical decisions appear to distill their experience into flexible gist representations that yield better decisions (Brewer et al., 2012; Fraenkel et al., 2012; Mills, Reyna, & Estrada, 2008; Reyna et al., 2011; Reyna & Mills, 2014).

FTT has been successfully used in a number of domains including argumentation (Britt, Kurby, Dandotkar, & Wolfe, 2008), health (Reyna & Adam, 2003), learning (Wolfe, Reyna, & Brainerd, 2005), legal reasoning (Reyna et al., 2015), numeracy (Reyna, Nelson, Han, & Dieckmann, 2009), and probabilistic reasoning (Wolfe, Fisher, & Reyna, 2012; Wolfe & Reyna, 2010). FTT stands in contrast to traditional rational choice models, such as Prospect Theory, that assume people consistently weigh perceived probabilities against stable well-defined preferences to arrive at optimal decisions (Kahneman, 2003; Tversky & Kahneman, 1992). Moreover, the FTT concept that understanding and experience are distilled into useful gist representations should be distinguished from perspectives emphasizing fast and frugal heuristics (Gigerenzer & Edwards, 2003; Wegwarth, Gaissmaier, & Gigerenzer, 2009). The "fast and frugal" approach (Wegwarth, Gaissmaier, & Gigerenzer, 2009) emphasizes "strategies that ignore information to make decisions faster, more frugally, and/or more accurately than more complex methods" (Gigerenzer & Gaissmaier, 2011, p. 453). The essence of gist is not processing smaller quantities of information. Rather, gist entails understanding *meaning*, integrating dimensions of information to distill its essence (i.e., "insight"; Reyna et al., 2015). Gist is more than simply processing fewer dimensions of information that are "good enough" (Reyna, 2013). The extent to which a gist representation is "insightful" depends on the knowledge, experience, and education of the individual

(Reyna & Lloyd, 2006) as well as other individual differences. FTT makes a number of sometimes counterintuitive predictions not supported by other theories (Reyna, 2008b; Wolfe, 1995).

After presenting didactic information, *BRCA Gist* asks people questions and helps them form good gist explanations of key decision-relevant information. Below is a sample tutorial dialogue from an earlier experiment between a woman and *BRCA Gist* (for details of that experiment, see Widmer et al., 2015). The participant is answering a question about what someone should do in the event of a positive test result for a mutation in the *BRCA1* or *BRCA2* gene:

1. *BRCA Gist*: What should someone do if she tests positive?
2. Person: If someone has tested positive for *BRCA* mutations, the first step is active surveillance, so that if the woman does have cancer, it can be detected early.
3. *BRCA Gist*: Good. Can you say more about active surveillance?
4. Person: Active surveillance includes clinical breast exams, MRIs, mammograms, and routine breast exams at home. If a questionable lump is found via active surveillance, it can be investigated further through a biopsy, to see if there is any cancer.
5. *BRCA Gist*: See if you can say something about drugs.
6. Person: Woman can also take the chemo drug tamoxifen to decrease their chances to getting breast cancer. Tamoxifen might work for people with *BRCA* genes, but there have been few studies on this specific population.
7. *BRCA Gist*: Please write a little bit more.
8. Person: It is approved for woman who are both pre- and postmenopausal. It can also be used by women who have already had breast cancer and are in remission, as a prevention strategy.
9. *BRCA Gist*: Ok, let's move on. Please click on the finish button.

Whereas in academic domains verbal exchanges between a student and AutoTutor may be 100 turns long (Graesser, McNamara, & VanLehn, 2005), self-explanations in *BRCA Gist* are briefer, with users typically making about seven dialogue moves to form a gist explanation. In this example we can see that *BRCA Gist* uses the *pump* “Please write a little bit more” and several *prompts* for additional information on specific topics, including “See if you can say something about drugs.”

Several interventions that test FTT have been developed, which differ from standard interventions by translating health communications into bottom-line gist (Fraenkel et al., 2012, 2015; Reyna, Weldon, & McCormick, 2015). The efficacy of *BRCA Gist* has been tested in randomized, controlled experiments in the laboratory, online, and in a local community (Widmer et al., 2015; Wolfe et al., 2015). Participants who interacted with *BRCA Gist* scored higher on measures of knowledge and gist comprehension than those who received comparable information in a more verbatim format from the National Cancer Institute website (Wolfe et al., 2015). In previous research, Wolfe and colleagues (under review) empirically demonstrated the added value of gist explanation dialogues. Participants who were randomly assigned to *BRCA Gist* performed significantly better on tests of declarative knowledge and the ability to apply knowledge to assess genetic breast cancer risk than those assigned to a version of *BRCA Gist* that was identical in every respect (avatars, didactic materials, graphics, etc.) without the gist explanation dialogues. Moreover, in a systematic analysis of gist explanation dialogues in several *BRCA Gist* experiments, Widmer and colleagues (2015) found a strong positive relationship between the quality of gist explanation dialogues and subsequent outcomes including measures of knowledge, comprehension, and risk assessment. One purpose of the current experiment was to learn how to optimize pumps and prompts to help people form effective gist explanations leading to these better outcomes.

According to FTT, both gist and verbatim levels of representation are encoded but have different effects on decision making (e.g., Reyna, 2008b). For example, Mills, Reyna, and Estrada (2008) found negative correlations between risky behavior and gist-based risk perceptions (more perceived risk was associated with less risky behavior) yet positive correlations between specific verbatim perceptions of risk and behavior. We exploit that difference by varying general and specific retrieval cues to reinforce different kinds of learning from discourse. The basic logic of this experiment was to alter the existing

successful version of *BRCA Gist* to push two ideas about dialogue pumps and prompts to greater extremes. FTT postulates that people are gist processors who prefer to reason with the essential bottom meaning of information. Thus, we changed the pumps and prompts to more fully ask the user for either the bottom-line gist meaning of information in the participants' own words or specific verbatim information from the tutorial. Crossed with that, we changed *BRCA Gist* dialogues to either pumps for general information or prompts for specific information. Drawing on research on retrieval cues in memory (Mills, Reyna, & Estrada, 2008; Reyna et al., 2011), FTT predicts that specific questions act as cues to different mental representations and associated modes of processing (Reyna et al., 2011). More general (but meaningful) cues tend to elicit gist memories, whereas more specific cues that recap learned material tend to elicit verbatim memories (e.g., Mills et al., 2008; Reyna & Brainerd, 1995; Reyna et al., 2011). The relationship between specificity (pump or prompt in an ITS) and requested precision of expression (gist or verbatim) has not been explored in previous research. Reyna and colleagues (2008) note that verbatim questions are more specific than gist questions in two ways: their content is more specific and they require more precise responses. Following FTT and the concept of cue specificity, we predict that consistent gist general pumps and verbatim specific prompts should elicit better recall than inconsistent verbatim general pumps and gist specific prompts. We predict that general gist-based dialogue pumps will yield better performance on categorical risk assessment and increase knowledge by cueing appropriate gist representations. By evoking verbatim representations, specific dialogue prompts are predicted to increase declarative knowledge without affecting categorical risk assessment. Predictions are provided in Table 1.

Methods

Participants

Participants were 434 undergraduate women recruited at a university in the Midwest and a university in the Northeast who received course credit for participating. We experienced technical difficulties with the system freezing at various times throughout the experimental period. Data from 33 participants were missing or excluded because interactions with *BRCA Gist* were disrupted or they did not complete the experiment, leaving a total of 401 participants providing usable data. Problems affected each condition randomly and in about the same proportion, with number's per group ranging from 75 in the Verbatim Specific condition to 84 in the Gist Specific condition (described below). Recruitment criteria were that participants had to be women over the age of 18 who did not have breast cancer. According to self-reports, the mean age of participants was 19.2 years ($SD = 1.4$), with 19.6% Asian or Asian American, 5.1% Black or African American, 7.5% Latina, 68.5% White, 3.9% mixed ethnicity, and 2.2%

Table 1. Hypotheses with supporting principles of FTT.

Hypotheses	FTT Principle/Source of Prediction
1: All four <i>BRCA Gist</i> versions will produce superior performance on declarative knowledge, gist comprehension, and categorical risk assessment than the control group.	Replication and Extension Manipulation Check
2: Gist general pumps and verbatim specific prompts will produce higher declarative knowledge test scores than inconsistent verbatim general pumps and gist specific prompts (interaction between manipulations).	Cue Specificity
3: Inducing gist responses will produce better performance on the categorical risk assessment task than inducing verbatim responses.	Superiority of Appropriate Gist Representations for Decision Making
4: Specific dialogue prompts will increase declarative knowledge without improving categorical risk assessment.	Limitations of Verbatim Representations for Decision Making
5: Greater coverage of content in participant explanations in the tutorial dialogues will be positively correlated with subsequent declarative knowledge and gist comprehension scores.	Efficacy of Gist Explanation

Table 2. Sample clauses used to form gist and verbatim pumps and prompts for tutorial dialogues.

Sample Clauses	
Verbatim focused	"Can you remember more exact information about . . . " "Tell me more specific information about . . . " "Can you provide more precise information about . . . " "Please write a bit more using the exact words from the lesson about . . . "
Gist focused	"Give me the gist of the information . . . " "In your own words, tell me the gist of . . . " "Can you provide something else about the essence of . . . " "Try to add more of the gist of . . . "
Specific prompts	"Let's see if you can add something else in your own words about . . . " " . . . about genes and breast cancer risk." " . . . about the importance of the age at which a relative gets breast cancer." " . . . about whether people in some specific countries and ethnic groups are at higher risk for BRCA mutations." " . . . about genetic risk factors."
General pumps	" . . . from this lesson." " . . . from this exercise." " . . . information I gave you earlier." " . . . of what you learned in this exercise." " . . . "about what I said before."

selecting “other ethnicity” in nonmutually exclusive categories (i.e., Hispanic, Latina, or Spanish was asked separately).

Experimental Design

Participants were randomly assigned to one of four abbreviated versions of the *BRCA Gist* tutor or a control group that received an equally time consuming and effortful tutor about an unrelated topic (nutrition). All the abbreviated versions of *BRCA Gist* consisted of the second and third of five *BRCA Gist* modules and the corresponding dialogues on “how do genes affect breast cancer risk” and “what should someone do if she receives a positive test result for genetic risk of breast cancer.” All didactic information including text and images were the same in each condition and the corresponding modules of the full *BRCA Gist* used in previous experiments. Also identical in each condition were the dialogue parameter settings and “back end” expectation texts (for a technical discussion of *BRCA Gist* tutorial dialogues, see Widmer et al., 2015; Wolfe et al., 2013). Thus, *BRCA Gist* assessed dialogues in all conditions in an identical manner with participants receiving the exact same feedback regarding performance (e.g., “good job” or “you seem to be off track”).

The dialogue conditions crossed the factors of requested precision of expression and specificity of cues. In addition to targeted overall instructions, half of the dialogues asked for precise verbatim information from participants and the other half of the dialogues asked for the gist of the information in the participants’ own words. Crossed with the requested precision of expression was the specificity dimension of cues. Half of the dialogues had prompts about specific information from the lesson and the other half of the dialogues had general or global pumps (for sample clauses see Table 2).

Thus, an example of a Verbatim Specific *BRCA Gist* dialogue turn (prompt) is “Can you provide more precise information about the importance of the age at which a relative gets breast cancer?” An example of a Gist General *BRCA Gist* dialogue turn (pump) is “Let’s see if you can add something else in your own words about what you learned in this exercise.”

Measures¹

A medical expert vetted research instruments assessing declarative knowledge, gist comprehension, and risk assessment as described below.

¹Copies of these instruments can be found online at <http://mdm.sagepub.com/content/35/1/46/suppl/DC1>

Declarative knowledge. We developed 52 four-alternative multiple-choice items on breast cancer, genetic risk, and genetic testing (see Wolfe et al., 2015). Items were created corresponding to modules on breast cancer and how it spreads; quantitative concepts and genetic risk; mutations, genetic testing, and genetic risk; and consequences of genetic testing. To illustrate, one item reads, “What is the goal of surveillance? (answer: to find cancer early when it is most treatable).” Unanswered items were scored as incorrect. Cronbach’s alpha for Declarative Knowledge is .88.

Gist comprehension of genetic breast cancer risk. We developed a 40 item, 1–7 Likert-scale instrument measuring gist comprehension of important information about breast cancer and genetic testing (see Wolfe et al., 2015). Gist comprehension items such as, “the greatest danger of dying from breast cancer is when it spreads to other parts of the body” express the gist of that knowledge—the essential bottom-line meaning. People can strongly endorse statements such as these without remembering the precise verbatim details. The item stem is stated at a general level such that verbatim information is not needed to answer the question. The response format permits degrees of agreement, with some items reversed scored. Cronbach’s alpha for Gist Comprehension is .85.

Risk assessment scenarios. These scenarios are further described in Wolfe et al. (2015). Participants received 12 scenarios describing a woman with no risk factors or medium or high genetic breast cancer risk based on Pedigree Assessment Tool scores of 0, 3–5, and 8–10, respectively. Each description includes a name, age, ethnicity, hometown, family health facts, and personal health facts. Scenarios were equated for age, range of words (between 56 and 60), range of Flesch Reading Ease Scores (between 56.9 and 62.9), and range of Flesch-Kincaid Grade Level Scores (between 7.3 and 7.9). The scenarios are challenging for laypeople to categorize. To illustrate, one high-risk scenario read as follows:

Claire is an unattached 35-year-old New Yorker. She has a vegan diet and is an avid jogger. Her family is of Scottish-Irish heritage. Recently, her 51-year-old uncle Sean was diagnosed with cancer of the breast. Claire has several siblings and to the best of her knowledge, her uncle Sean is the only family member with breast cancer. Participants evaluated risk by categorizing degree of genetic breast cancer risk for each woman as low, medium, or high.

Procedure

Participants were recruited online, and the experiment took place in the laboratory at each university. Participants were run individually or in small groups of two or three at separate workstations. After providing informed consent, participants interacted with the tutor to which they were randomly assigned and then completed the dependent measures, which were administered online at the same workstations. In all conditions, interacting with the avatar and completing the dependent measures took approximately 60 minutes. After the experiment, participants were thanked and debriefed.

Results

The general plan of analysis was first to examine the effects of the experimental conditions relative to the control group with a series of 1×5 ANOVAs followed by Tukey’s HSD tests (Hypothesis 1); then to compare the tutorial dialogue conditions to one another with respect to outcome measures with a series of 2×2 ANOVAs examining requested precision of expression and specificity excluding the control group; and finally to assess the effects of the dialogue conditions on the participant dialogues themselves with a series of 2×2 ANOVAs, again excluding the control group (which did not involve these tutorial dialogues). For each analysis we found a significant main effect for location with participants at the Northeastern university scoring better than those at the Midwestern university, perhaps because the former is more academically selective. However, in no case did we find a significant interaction between location and experimental condition. Thus, location was excluded from subsequent analyses.

Control group participants had a mean of 61% correct ($SD = 12.5$) on Declarative Knowledge, which was significantly lower than the Gist General and Verbatim Specific conditions, $F(4, 396) = 4.37$, $p = .002$, $\eta^2 = .0423$, Tukey's HSD = 5.75, and slightly but not significantly lower than the Gist Specific and Verbatim General conditions (Table 1). Control participants had a mean Gist Comprehension score of 4.35 ($SD = .74$) which was significantly lower than the Gist General and Verbatim Specific conditions, $F(4, 398) = 2.92$, $p = .021$, $\eta^2 = .0285$, Tukey's HSD = .29, and again slightly but not significantly lower than the Gist Specific and Verbatim General conditions (Table 3). For Risk Assessment, control participants had a mean of 50.4% correct ($SD = 13\%$), which was significantly lower than all other groups, $F(4, 399) = 12.67$, $p < .0001$, $\eta^2 = .1127$, Tukey's HSD = 6.88.

Turning to the 2×2 comparisons among dialogue conditions, for Declarative Knowledge percent correct there was a significant interaction between requested precision of expression and specificity conditions (Hypothesis 2), $F(3, 316) = 2.65$, $p = .006$, $\eta^2 = .0240$. As seen in Figure 1 and Table 3, participants in the Gist General and Verbatim Specific conditions scored significantly higher than those in the Gist Specific and Verbatim General conditions. Before the experiment we identified a subset of 24 Declarative Knowledge items specifically covered in the two tutorial modules and we found a comparable interaction for that subset of items, $F(3, 316) = 2.29$, $p = .027$, $\eta^2 = .0163$ (Table 1). Participants in the Gist General and Verbatim Specific conditions scored significantly higher than those in the Gist Specific and Verbatim General conditions. With respect to Gist Comprehension, although the Gist General and Verbatim Specific conditions scored significantly higher than the control group, there were no significant differences among dialogue conditions, $F < 1$. Before the experiment we identified 20 Gist Comprehension items covered in the two modules. There were no significant differences among dialogue conditions for that subset of items, $F < 1$.

With respect to overall Risk Assessment, there were no significant differences among dialogue conditions, $F < 1$. However, for the four scenarios representing women at high genetic risk for breast cancer, there was a significant main effect for the requested precision of expression condition (Hypothesis 3), $F(3, 318) = 1.77$, $p = .024$, $\eta^2 = .0153$. Participants in the gist dialogue conditions scored significantly higher than those in the verbatim conditions (Figure 2 and Table 3). The effects of specificity and the requested precision of expression by specificity interactions were not significant (Hypothesis 4), and there were no significant differences among groups for low-risk and medium-risk scenarios. These results suggest that asking people to provide gist explanations in their own words helps them to better apply their knowledge to identify situations in which a women is at high genetic risk of breast cancer.

The final set of analyses address the effects of the dialogue manipulations on how participants verbally respond in the dialogues as they developed self-explanations. The first approach is to examine final coverage (CO) scores for each dialogue generated by the *BRCA Gist* dialogue engine. CO score is a variable generated by AutoTutor Lite using Latent Semantic Analysis that represents the cumulative degree to which the expectations have been met by a person's total responses combined across conversational turns (i.e., the semantic similarity between a person's self-explanation text and the expectation text we created for each dialogue question. CO scores range from 0 to less than 1, with higher numbers corresponding to better coverage of the expectations text. In our previous research, mean CO scores for all our tutorial dialogues were less than .5. However, maximum scores differ greatly from one expectation text to the other, and comparisons are only meaningful within dialogue (for a more detailed treatment see Wolfe et al., 2013). Previous research indicates that final CO scores are highly correlated with the judgments of trained researchers using reliable scoring rubrics (Wolfe et al., 2015, under review) to assess how well participants covered a set of expectations about a good answer to each question. Other discourse measures include number of conversational turns, number of words per dialogue, and mean number of words per turn.

Table 4 shows mean final CO score, turns, words, and words per turns for both tutorial dialogues. There were no significant differences among groups for final CO score on the dialogue about genes and breast cancer risk, $F(3, 287) = 1.85$, $p = .14$, or for the dialogue about what to do if someone tests positive for *BRCA* mutations, $F < 1$, and CO scores were comparable with those found in previous

Table 3. Declarative knowledge, gist comprehension, risk assessment, and dialogue coverage means, and SD by condition.

	Verbatim Specific Prompts	Verbatim General Pumps	Gist Specific Prompts	Gist General Pumps	Control
Declarative knowledge percent correct (all 52 items)	68.3% ^a (11.0)	64.6% (16.5)	64.0% (13.9)	68.7% ^a (10.6)	61% ^b (12.5)
Declarative knowledge 24 items on covered modules	69.2% ^c (12.5)	67.5% (17.5)	64.4% (16.7)	70.1% ^c (12.1)	57.3% ^d (12.9)
Gist comprehension	4.66 (.66)	4.54 (.80)	4.58 (.62)	4.67 (.39)	4.35 ^e (.74)
Gist comprehension for covered modules	4.90 (.728)	4.78 (.881)	4.80 (.706)	4.93 (.510)	4.55 ^f (.795)
Total risk assessment percent correct	65.6% (14.4)	62.9% (18.7)	62.3% (16.9)	65.5% (14.0)	50.4% ^d (13.0)
High risk assessment percent correct	58.8% (20.2)	58.8% (22.1)	63.7% ^g (21.4)	64.4% ^g (20.6)	44.6% ^e (18.3)
Final CO score genes and BC risk	.401 (.10)	.392 (.12)	.400 (.13)	.362 (.11)	
Final CO score positive test result	.258 (.08)	.250 (.10)	.248 (.15)	.236 (.10)	

Values in parentheses are SD.

^aInteraction: Verbatim Specific and Gist General significantly higher than Verbatim General and Gist Specific, $p = .006$.

^bVerbatim Specific and Gist General significantly higher than control, $p = .002$.

^cInteraction: Verbatim Specific and Gist General significantly higher than Verbatim General and Gist Specific, $p = .027$.

^dControl significantly lower than all experimental conditions, $p < .0001$.

^eVerbatim Specific and Gist General significantly higher than control, $p = .021$.

^fVerbatim Specific and Gist General significantly higher than control, $p = .012$.

^gGist significantly higher than Verbatim, $p = .024$.

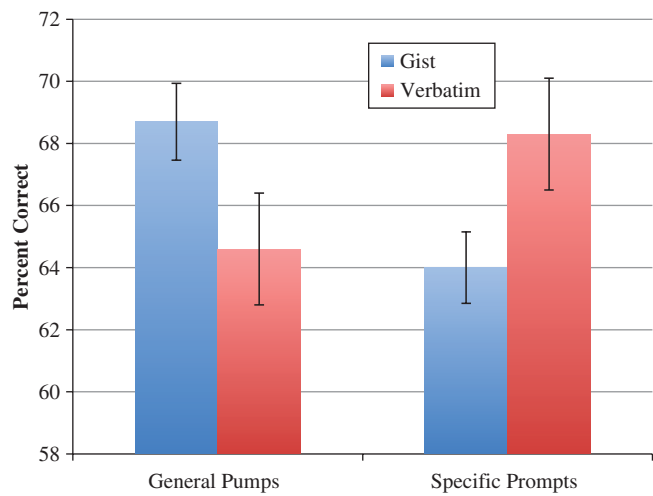


Figure 1. Declarative knowledge percent correct by condition.

research with the complete *BRCA Gist*. There were no differences among conditions for number of conversational turns, $F < 1$. There were no significant differences among conditions for number of words on either the dialogue about genes, $F < 1$, or the dialogue about positive test results, $F(3, 281) = 1.29, p = .28$. With respect to words per turn, there were no significant differences among conditions for the dialogue about genes, $F < 1$. However, for the dialogue about positive test results, participants in the gist conditions generated significantly more words per turn (16.6) than those in the verbatim conditions (14.6), $F(3, 281) = 6.9, p = .009, \eta^2 = .0245$. The effects of specificity and the interaction were not significant, $F < 1$.

We analyzed the relationship between discourse-level variables and declarative knowledge. As can be seen in Table 5, for both dialogues there were positive correlations between declarative knowledge outcomes and both final CO score and number of words per turn. For the dialogue about positive test results there was also a positive correlation between declarative knowledge score and number of words generated by participants. However, when multiple variables are subjected to multiple regression analysis, only the CO score significantly predicts declarative knowledge for the dialogue about genes and cancer, $F(4, 281) = 4.0, p < .05, \eta^2 = .0551$, and for the dialogue about positive test results,

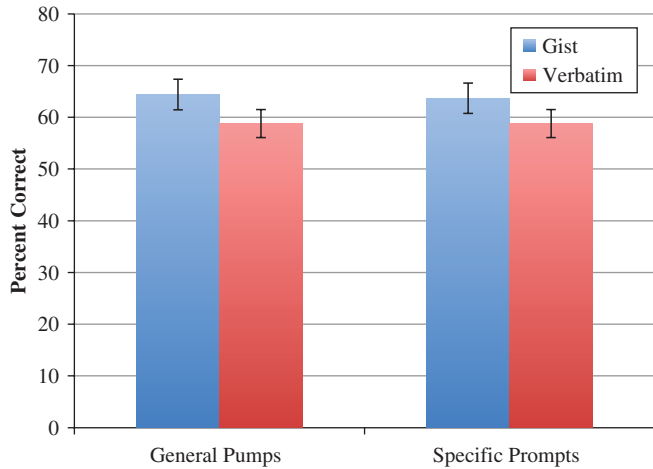


Figure 2. High-risk assessment percent correct by condition.

Table 4. Mean final CO score, turns, words, and words per turns for both tutorial dialogues by condition.

	Gist General	Gist Specific	Verbatim General	Verbatim Specific
<i>Genes and cancer dialogue</i>				
CO score	.362 (.11)	.402 (.12)	.394 (.12)	.404 (.10)
Conversational turns	7.84 (3.80)	8.40 (3.48)	8.03 (3.59)	7.56 (2.93)
Number of words	136 (71)	140 (62)	130 (66)	126 (61)
Number of words per turn	17.3 (5.3)	16.9 (4.2)	16.4 (4.7)	16.8 (5.2)
<i>Positive test result dialogue</i>				
CO	.236 (.10)	.249 (.15)	.252 (.10)	.261 (.08)
Conversational turns	6.81 (2.41)	6.96 (2.19)	6.92 (2.60)	7.39 (2.14)
Number of words	117 (70)	115 (54)	99 (59)	112 (53)
Number of words per turn	16.8 (7.5)	16.4 (6.9)	14.1 (5.7)	15.1 (5.6)

Values in parentheses are SD.

Table 5. Correlations between declarative knowledge and CO scores, conversational turns, words, and words per turn by dialogue.

	Declarative Knowledge Score	Statistical Significance
<i>Genes and cancer dialogue</i>		
CO score	.23 ^a	$p < .0001$
Conversational turns	-.10	$p = .08$
Number of words	.06	$p = .30$
Number of words per turn	.25 ^b	$p < .0001$
<i>Positive test result dialogue</i>		
CO score	.30 ^a	$p < .0001$
Conversational turns	.06	$p = .33$
Number of words	.20 ^b	$p < .0006$
Number of words per turn	.24 ^b	$p < .0001$

^a $p < .5$ in multiple regression analysis with other discourse variables.

^b $p < .5$ in paired comparisons.

$F(4, 281) = 10.2, p = .0015, \eta^2 = .0870$. In addition, there were positive correlations between CO score and gist comprehension for both the dialogue about genes, $r(284) = .14, p = .016$, and the dialogue about positive test results, $r(284) = .25, p < .0001$. This suggests that providing gist explanations that more fully answer the question is associated with greater knowledge and comprehension irrespective of experimental condition (Hypothesis 5).

In reading over the dialogues, one is struck by how gist-like they are regardless of dialogue condition. To illustrate, here is a dialogue between participant 1516 in the Verbatim Specific condition and *BRCA Gist* in answer to the question what should someone do if she receives a positive test result for genetic risk of breast cancer. Here, the final CO score was .33 (typos in original):

1. Participant: If someone has tested positive for a *BRCA* mutation, they can have a mastectomy to remove all the healthy breast tissue that is at risk.
2. *BRCA Gist*: Good. Can you remember other specific information about active surveillance?
3. Participant: If a woman has tested positive for *BRCA* mutations, she may want to use surveillance to catch any cancer formations early on to reduce the risk of dying from breast cancer.
4. Participant: Active surveillance can catch cancer early in, that that it may be treated effectively.
5. Participant: Having a mastectomy or removing the fallopian tubes will lower the chances of developing breast or ovarian cancer.
6. *BRCA Gist*: Please write a bit more using the exact words from the lesson about what someone should do if she tests positive for *BRCA* mutations.
7. Participant: There is also a drug which can be taken to better fight off the risk of cancer which has been approved by the FDA
8. *BRCA Gist*: See if you can say something else using the exact words that I said before about drugs.

9. Participant: If a woman tests positive for *BRCA* mutations, she should continue surveillance with MRI's.
10. Participant: Tamoxalin is a drug which can be used to help prevent breast cancer.
11. Participant: tamoxalin does decrease the risk of breast cancer, but it is still possible to develop breast cancer.
12. *BRCA Gist*: Ok. You are finished. Please tell the experimenter that you are ready to continue.

Here it is evident that the participant is using her own words even when specifically asked to use "the exact words from the lesson." Below is a dialogue from participant 1559 in the Gist General condition on answer to the same question (typos and spelling mistakes in the original). Her final CO score was also .33:

1. Participant: If someone tested positive for *BRCA* mutations, she should use surveillance so she can catch the breast cancer early.
2. *BRCA Gist*: Thinking about the gist of the question, exactly what should someone do?
3. Participant: She should get mammograms or an MRI regularly to catch the breast cancer.
4. Participant: She can also get prophylactic surgery.
5. Participant: She can get the at risk breast tissue or fallopian tubes removed.
6. *BRCA Gist*: How about this, try to add more of the gist of the tutorial.
7. Participant: By removing some healthy breast or ovarian tissue, she is reducing her risk of developing breast or ovarian cancer.
8. *BRCA Gist*: Nice job. Let's see if you can add something else in your own words about what I said before.
9. Participant: She can also take Tamoxifen.
10. Participant: Tamoxifen is a chemo drug that reduces the risk of developing breast cancer.
11. Participant: Tamoxifen is approved by the FDA.
12. *BRCA Gist*: Ok. You are finished. Please tell the experimenter that you are ready to continue.

Below are two sample dialogues in answer to the question how do genes affect breast cancer risk. In each case, typos and spelling mistakes are retained from the original. Participant 1506 is in the Verbatim Specific condition and had a final CO score of .55:

Participant: Genes affect breast cancer risk because each gene has a specific function in the body.

1. Participant: Specifically *BRCA1* and *BRCA2*, which when not mutated work to repress abnormal cell growth.
2. *BRCA Gist*: Tell me more specific information about genes and breast cancer risk.
3. Participant: If the two genes *BRCA1* and *BRCA2* are mutated they cannot function properly and people with these mutations have a higher risk of getting breast cancer.
4. Participant: You inherit half of your genes from your mother and half from your father, so if either parent has one of these mutated genes or has breast cancer your risk will have greatly increased.
5. *BRCA Gist*: Well done. Tell me more specific information about the importance of the age at which a relative gets breast cancer?
6. Participant: If your father or a male relative in your family has acquired breast cancer, there is a higher probability that you will also get breast cancer.
7. *BRCA Gist*: Nice job. Can you provide more precise information about whether people in some specific countries and ethnic groups are at higher risk for *BRCA* mutations?
8. Participant: Sometimes race can play an affect, the Achsnezik Jews are at higher risk for *BRCA* mutatuions as shown by scientific studies and therefore also at higher risk for getting breast cancer.
9. Participant: Another indicator that one's risk for breast cancer is increased is if a relative has or they have ovarian cancer, which also leads to higher risk of breast cancer.
10. *BRCA Gist*: Good job.

Participant 50 is answering the same question in the Gist General condition and had a final CO score of .56:

1. Participant: Genes affect breast cancer risk because the history of people that have had breast cancer in your family in the past affects the chances that you will get it.
2. Participant: The chances are especially increased if your mother or grandmother had breast cancer are even more increased if your ancestor who had breast cancer developed it before menopause.
3. *BRCA Gist*: You are doing well. Give me the gist of the information I gave you earlier.
4. Participant: The more immediate your family is who developed breast cancer and the younger they developed it, the higher risk you have of developing breast cancer.
5. *BRCA Gist*: Good job. In your own words, tell me the gist of what you learned in this exercise?
6. Participant: That there are many factors that go into people getting breast cancer, some that can be avoided and others that cannot.
7. *BRCA Gist*: Nice job. Can you provide something else about the essence of what I told you before?
8. Participant: Also if you have any family that is Ashkenazi Jewish this greatly increases your chances of getting breast cancer because people of this religious domain already have medium risk
9. Participant: Also, if you drink alcohol it increases your chances of getting breast cancer and your chances decrease as you get older, especially after menopause.
10. *BRCA Gist*: Good job.

It is evident that for the most part participants answered with the meaning of the information rather than stating the precise working from the tutorial. Participants generally took what they learned in the tutorial and restated in their own words regardless of experimental condition. The dialogue above also contains an error that was not caught by *BRCA Gist*. Participant 50 said your chances of getting breast cancer decrease as you get older, whereas they actually increase. It is unclear whether the participant simply misspoke, or whether this is evidence of a deeper misunderstanding. These dialogues also indicate that the AutoTutor Lite platform used to develop *BRCA Gist* is robust to typos and spelling mistakes.

Discussion

FTT principles of cue specificity, superiority of appropriate gist representations for decision making, and the effectiveness of generating gist explanations all received empirical support in this experiment. Hypothesis 1, that all *BRCA Gist* groups would outperform the control group, was generally supported with the exception of the verbatim general condition (see Table 3). This suggests that the findings are best understood in the context of a successful tutoring intervention. We found the predicted interaction based on the notion of cue specificity that gist general and verbatim specific groups did significantly better than the gist specific and verbatim general groups (Hypothesis 2, see Figure 1). Practically speaking, it is easier to create an ITS with general pumps than specific prompts so the relatively strong performance of the Gist General group is good news for developers (for related findings see Kopp, Britt, Mills, & Graesser, 2012). Hypothesis 3 was supported only for the high-risk scenarios (see Table 3). There was a significant main effect for the gist instructions but no significant differences on the other scenarios. However, the finding that the verbatim instructions did not improve decision making on the categorical risk assessment task, despite the predicted significantly higher scores on declarative knowledge for the verbatim specific condition supports Hypothesis 4. We found that higher CO scores indicating coverage of the material in the tutorial dialogues are correlated with better subsequent performance on the gist comprehension and declarative knowledge tests (Hypothesis 5, see Table 5). This replicates earlier studies and although the results of the current investigation are correlational, they are consistent with experimental evidence that providing

responses in one's own words in tutorial dialogues produces higher learning gains (Wolfe et al., under review).

In an ITS *pumps* are dialogue moves that encourage users to say more, for example, “What else?” (Graesser, 2011). We have ample evidence that pumps for information stated at a general level that are designed to evoke a gist representation by asking users to respond in their own words lead to better outcomes. Participants in the Gist General (own words pump) condition scored significantly higher than the control on declarative knowledge, gist comprehension, and overall risk assessment. People in the Gist General (own words pump) condition also scored significantly higher than the Verbatim General (exact tutorial words pump) on declarative knowledge and risk assessment for scenarios involving high genetic risk for breast cancer. Indeed the Verbatim General (exact words pump) condition was not significantly better than the control group on declarative knowledge and gist comprehension. These results are of interest for both practical and theoretical reasons. Theoretically, this result supports the FTT principle that people are mainly gist processors and that interventions designed to requested appropriate gist explanations are more likely to produce better decision making. Manipulations that help people develop appropriate gist representations increase both knowledge and risk assessment.

Prompts are dialogue moves that guide users to fill in missing information (Graesser, 2011). As predicted, verbatim specific prompts significantly increased declarative knowledge without affecting risk assessment. The Gist Specific group did significantly better than the Verbatim Specific group on risk assessment for scenarios involving high genetic risk for breast cancer. However, the Verbatim Specific group outperformed the Gist Specific group on declarative knowledge. Consistent gist general pumps and verbatim specific prompts yielded better performance on tests of declarative knowledge than inconsistent verbatim general pumps or gist specific prompts, echoing previous research on cue specificity (Mills, Reyna, & Estrada, 2008; Reyna et al., 2011).

Asking people to respond “using the exact words from the tutorial” to specific prompts on topics such as “the importance of the age at which a relative gets breast cancer” was useful for testing FTT hypotheses about the relationships among gist and verbatim representations, declarative knowledge, and risk assessment. However, at a practical level perhaps the best prompts for specific information in ITS or by human healthcare providers will be those that do not mention “your own words” or “exact wording” at all. This was the approach taken in the original full version of *BRCA Gist* exemplified by prompts such as “what should someone do if she tests positive” and “can you say more about active surveillance” and “see if you can say something about drugs.”

These results also have implications for physicians and other healthcare providers aspiring to improve patient education and shared decision making (Col, 2011). Evidence for our approach focusing on the patient's gist understanding rather than precise verbatim information or simple information reduction has now been found in a number of studies (e.g., Widmer et al., 2015; Wolfe et al., 2015). The current study suggests that providers are wise to also consider the value of encouraging patients to share their understanding of complex medical information. General pumps coupled with the judicious use of specific prompts are good strategies for guiding the discourse to improve understanding. Perhaps the biggest lesson for physicians and other healthcare providers is the wisdom of simply listening, of letting the patients speak in their own words, both to provide opportunities to address misconceptions and to allow patients to more fully develop their own mental representations.

We used subtle experimental manipulations with all didactic wording, avatars, images, and feedback on response quality the same in each *BRCA Gist* condition, which consisted of only a portion of the full *BRCA Gist* tutor. Performance on the Declarative Knowledge and Gist Comprehension tasks was somewhat lower than has been found in previous research with the full *BRCA Gist* tutor (Wolfe et al., 2015, under review) with mean Declarative Knowledge percent correct 76.6% ($SD = 12.6\%$) and mean Gist Comprehension 5.22 ($SD = .58$). Performance was somewhat better in the current experiment on the Risk Assessment scenarios compared with a mean of 62.7% percent correct ($SD = 14.1\%$) in previous research with the complete *BRCA Gist* (Wolfe et al., under review).

These differences are not surprising given that participants in the current study interacted with a truncated version of *BRCA Gist*.

The effect sizes reported here are generally modest (i.e., small to medium, with η^2 ranging from .025 to .113). However, it is worth noting again that the experimental manipulations are quite subtle with all conditions sharing the identical didactic text and graphics, with the same avatars providing feedback using identical wording with the same criteria for responding to participants in different ways. Thus, differences between pumps and prompts in the gist and verbatim conditions are the result of very small changes in wording. Other sources of variance probably include differences in knowledge of and interest in breast cancer; differences in motivation to participate in research; and other social, cultural, and cognitive differences.

There are a number of shortcomings and limitations to this investigation. We did not include a manipulation check to make sure participants understood the gist and verbatim requested precision of expression instructions. In future work, it may be wise to include a post-test survey assessing whether participants understood these instructions. Another strategy would be to start with a validation trial task in which participants demonstrate they are able to follow the instructions to “use the exact words” or use “your own” words in recounting a brief passage with uncommon expressions such as “strolled” or “sauntered” instead of “walked” or “went.” The ability to follow such instructions would be taken as evidence that any deviations from exact wording could not be attributed to misunderstanding the instructions.

Another issue that makes interpretation more difficult is that in the *BRCA Gist* tutorials, the didactic text itself was designed to promote gist comprehension and emphasized the bottom-line meaning rather than encouraging rote memorization of verbatim facts (Wolfe et al., 2015; under review). This makes it more difficult to distinguish gist from verbatim responses than, for example, the National Cancer Institute website. It may be that the differences between the gist and verbatim instructions would have been more pronounced if we had used materials such as those more commonly used in patient education. Nonetheless, the evidence suggests that people prefer to talk about the essential bottom-line meaning of what they are learning in their own words rather than precisely reconstruct specific text from the lessons—even when prompted to do so. We also did not include a condition with unaltered sections of *BRCA Gist* corresponding to the content covered in the abbreviated tutorials, making it difficult to know how much better (or worse) participants might have scored under those conditions.

There is a kinship between general pumps for the bottom-line meaning of information (exemplified in our Gist General condition) and free recall (retrieval without the aid of memory cues) and a corresponding relationship between prompts for specific information and cued recall (retrieval with the aid of memory cues), and both can be thought of as embodying “testing effects” (Brainerd & Reyna, 1996; Carpenter, Pashler, & Vul, 2006; Roediger & Karpicke, 2006). In each case, actively generating responses at the time of learning improves subsequent performance. General cues tend to elicit gist memories through free recall, whereas more specific cues that recap learned material tend to elicit verbatim memories (Mills et al., 2008; Reyna & Brainerd, 1995; Reyna, 2011). It is well established that retrieval practice strengthens learning (Reyna, 2012). However, the interactions between people and the ITS are better understood as dialogues rather than tests. What appears to be the key is forming self-explanations of the bottom-line meaning or gist of decision-relevant dimensions of a decision domain and the consequences of different decision alternatives. This can best be accomplished through a mix of gist-evoking pumps to discuss the bottom line meaning of key issues and specific prompts without reference to answering in one’s own words.

A new generation of web-based discourse technologies is beginning to change the landscape of e-commerce and online education. There is an acute need for tools to help everyday people make complex medical decisions. The early evidence portends that emerging discourse technologies such as *BRCA Gist* will play an important role in helping everyday people understand a bewildering array of health information and engage in effective shared medical decision making with physicians, genetic counselors, and other healthcare professionals.

Acknowledgment

We thank Hannah Abner, Kate Baker, Andrew Circelli, Cecelia Favede, Colleen Gillard, Jennifer Kalkman, Mitch McDaniel, Madyson Murray, Vani Rajkumar, Margaret Schaller, Grant Schmidt, and Michael Thomas for capable assistance with data collection.

References

- Agus, D. B. (2013). The outrageous cost of a gene test. *New York Times*. Available at: http://www.nytimes.com/2013/05/21/opinion/the-outrageous-cost-of-a-gene-test.html?_r=. Updated May 20, 2013. Accessed July 10, 2015.
- American Cancer Society. (2016). What is breast cancer? Available at: <http://www.cancer.org/cancer/breastcancer/detailguide/breast-cancer-what-is-breast-cancer>. Accessed January 28, 2016.
- Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: a systematic review. *Journal of the American Medical Informatics Association*, 13, 608–618.
- Andrews, M. (2013). Coverage gaps can hamper access to some breast cancer screening, care. *Kaiser Health News*. Available at: <http://www.kaiserhealthnews.org/Features/Insuring-Your-Health/2013/052813-Michelle-Andrews-on-breast-cancer-care.aspx>. Updated May 27, 2013. Accessed July 10, 2015.
- Armstrong, K., Eisen, A., & Weber, B. (2000). Assessing the risk of breast cancer. *New England Journal of Medicine*, 342, 564–571.
- Arnott, E., Hastings, P., & Allbritton, D. (2008). Research methods tutor: evaluation of a dialogue-based tutoring system in the classroom. *Behavior Research Methods*, 40, 694–698.
- Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, 30, 241–254.
- 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.
- Brainerd, C. J., & Reyna, V. F. (1996). Mere memory testing creates false memories in children. *Developmental Psychology*, 32, 467–478.
- 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.
- Britt, M. A., Kurby, C. A., Dandotkar, S., & Wolfe, C. R. (2008). I agreed with what? Memory for simple argument claims. *Discourse Processes*, 45, 52–84.
- Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, 13, 826–830.
- Chao, C., Studts, J. L., Abell, T., Hadley, T., Roetzer, L., Dineen, S., Lorenz, D., . . . McMasters, K. M. (2003). Adjuvant chemotherapy for breast cancer: How presentation of recurrence risk influences decision-making. *Journal of Clinical Oncology*, 21, 4299–4305.
- Chi, M., VanLehn, K., & Litman, D. (2010). Do micro-level tutorial decisions matter: Applying reinforcement learning to induce pedagogical tutorial tactics. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent Tutoring Systems* (pp. 224–234). Berlin Heidelberg: Springer.
- Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011). An evaluation of pedagogical tutorial tactics for a natural language tutoring system: A reinforcement learning approach. *International Journal of Artificial Intelligence in Education*, 21, 83–113.
- Chi, M. T. (2000). Self-explaining expository texts: The dual 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.
- Cleveland Clinic. (2016). Diseases and conditions: breast. Available at: https://my.clevelandclinic.org/health/diseases_conditions/hic_Breast_Cancer_An_Overview. Accessed January 28, 2016.
- Cohn, J., Olde, B., Bolton, A., Schmorrow, D., & Freeman, H. (2015). Adaptive and generative agents for training content development. In R. Sottolare, A. Graesser, X. Hu, & L. Brawner (Eds.), *Design recommendations for intelligent tutoring systems: authoring tools* (Volume 3; pp. 161–168). Orlando, FL: U.S. Army Research Laboratory.
- Col, N. (2011). Interprofessional education about shared decision making for patients in primary care settings. *Journal of Interprofessional Care*, 25, 409–415.
- 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.

- Fagerlin, A., Zikmund-Fisher, B. J., & Ubel, P. A. (2011). Helping patients decide: Ten steps to better risk communication. *JNCI Journal of the National Cancer Institute*, 103, 1–8.
- Fraenkel, L., Matzko, C. K., Webb, D. E., Oppermann, B., Charpentier, P., Peters, E., . . . Newman, D. (2015). Use of decision support for improved knowledge, values clarification, and informed choice in patients with rheumatoid arthritis. *Arthritis Care & Research*, 67, 1496–1502. doi:10.1002/acr.22659
- Fraenkel, L., Peters, E., Charpentier, P., Olsen, B., Errante, L., Schoen, R., & Reyna, V. F. (2012). A decision tool to improve the quality of care in Rheumatoid Arthritis. *Arthritis Care & Research*, 64, 977–985.
- Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: From innumeracy to insight. *British Medical Journal*, 327, 741–744.
- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62, 451–482.
- Graedon, T., & Graedon, J. (2014). Let patients help with diagnosis. *Diagnosis*, 1, 49–51.
- Graesser, A. C. (2011). Learning, thinking, and emoting with discourse technologies. *American Psychologist*, 66, 746–757.
- Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48, 612–618.
- Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36, 180–192.
- 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., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query. *AutoTutor, and iSTART. Educational Psychologist*, 40, 225–234.
- Graesser, A., Person, N., & Magliano, J. (1995). Collaborative dialog patterns in naturalistic one-on-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.
- Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58, 697–720.
- 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.
- Lloyd, F. J., & Reyna, V. F. (2009). Clinical gist and medical education: connecting the dots. *Journal of American Medical Association*, 302, 1332–1333.
- McNamara, D. S., Jacovina, M. E., Snow, E. L., & Allen, L. K. (2015). From generating in the lab to tutoring systems in classrooms. *American Journal of Psychology*, 128, 159–172.
- Mills, B., Reyna, V. F., & Estrada, S. (2008). Explaining contradictory relations between risk perception and risk taking. *Psychological Science*, 19, 429–433.
- National Cancer Institute. (2016). Breast cancer—patient version. Available at: <http://www.cancer.gov/types/breast>. Accessed January 28, 2016.
- Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, 24, 427–469.
- Offit, K. (2006). BRCA mutation frequency and penetrance: new data, old debate. *Journal of National Cancer Institute*, 98, 1675–1677.
- 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. (2008a). A theory of medical decision-making and health: Fuzzy trace theory. *Medical Decision-making*, 28, 850–865.
- Reyna, V. F. (2008b). Theories of medical decision-making and health: An evidence-based approach. *Medical Decision-making*, 28, 829–833.
- Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision-making*, 7, 332–359.
- Reyna, V. F. (2013). Intuition, reasoning, and development: A fuzzy-trace theory approach. In P. Barrouillet & C. Gauffroy (Eds.), *The development of thinking and reasoning* (pp. 193–220). Hove, England: Psychology Press.
- Reyna, V. F., & Adam, M. B. (2003). Fuzzy-trace theory, risk communication, and product labeling in sexually transmitted diseases. *Risk Analysis*, 23, 325–342.
- 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., Chick, C. F., Corbin, J. C., & Hsia, A. N. (2014). Developmental reversals in risky decision-making: Intelligence agents show larger decision biases than college students. *Psychological Science*, 25, 76–84.
- Reyna, V. F., Estrada, S. M., DeMarinis, J. A., Myers, R. M., Stanis, 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.

- Reyna, V. F., Hans, V. P., Corbin, J. C., Yeh, R., Lin, K., & Royer, C. (2015). The gist of juries: Testing a model of damage award decision making. *Psychology, Public Policy, and Law*, 21, 280–294.
- Reyna, V. F., & Kiernan, B. (1994). Development of gist versus verbatim memory in sentence recognition: Effects of lexical familiarity, semantic content, encoding instructions, and retention interval. *Developmental Psychology*, 30, 178–191.
- 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.
- Reyna, V. F., Lloyd, F., & Whalen, P. (2001). Genetic testing and medical decision-making. *Archives of Internal Medicine*, 161, 2406–2408.
- Reyna, V. F., & Mills, B. A. (2014). Theoretically motivated interventions for reducing sexual risk taking in adolescence: A randomized controlled experiment using fuzzy-trace theory. *Journal of Experimental Psychology: General*, 143, 1627–1648.
- Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision-making. *Psychological Bulletin*, 135, 943–973.
- Reyna, V. F., Nelson, W. L., Han, P. K., & Pignone, M. P. (2015). Decision making and cancer. *American Psychologist*, 70, 105–118.
- Reyna, V. F., Weldon, R. B., & McCormick, M. (2015). Educating intuition: Reducing risky decisions using Fuzzy-Trace Theory. *Current Directions in Psychological Science*, 24, 392–398.
- Roediger, H. L. III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17, 249–255.
- Roscoe, R. D., & Chi, M. T. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science*, 36, 321–350.
- Shaffer, V. A., Hulse, L., & Zikmund-Fisher, B. J. (2013). The effects of process-focused versus experience-focused narratives in a breast cancer treatment decision task. *Patient Education & Counseling*, 93, 255–264.
- Shaffer, V. A., & Zikmund-Fisher, B. J. (2013). All stories are not alike: A purpose-, content-, and valence-based taxonomy of patient narratives in decision aids. *Medical Decision Making*, 33, 4–13.
- Singer, M., & Remillard, G. (2008). Veridical and false memory for text: A multiprocess analysis. *Journal of Memory and Language*, 59, 18–35.
- Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3–22.
- Snow, E. L., Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2015). Does agency matter? Exploring the impact of controlled behaviors within a game-based environment. *Computers & Education*, 82, 378–392.
- Sottile, R., Graesser, A., Hu, X., & Brawner, L. (2015). *Design recommendations for intelligent tutoring systems—Volume 3: authoring tools and expert modeling Techniques*. Orlando, FL: U.S. Army Research Laboratory.
- Stefanek, M., Hartmann, L., & Nelson, W. (2001). Risk-reduction mastectomy: Clinical issues and research needs. *Journal of the National Cancer Institute*, 93, 1297–1306.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5, 297–323.
- van de Sande, C., & Greeno, J. G. (2010). A framing of instructional explanations: Let us explain with you. In M. K. Stein & L. Kucan (Eds.), *Instructional explanations in the disciplines* (pp. 69–82). New York, NY: Springer.
- 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.
- Wegwarth, O., Gaissmaier, W., & Gigerenzer, G. (2009). Smart strategies for doctors and doctors-in-training: heuristics in medicine. *Medical Education*, 43, 721–728.
- 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.
- Wilhelms, E. A., & Reyna, V. F. (2013). Fuzzy trace theory and medical decisions by minors: Differences in reasoning between adolescents and adults. *Journal of Medicine and Philosophy*, 38, 268–282.
- Wolfe, C. R. (1995). Information seeking on Bayesian conditional probability problems: A fuzzy-trace theory account. *Journal of Behavioral Decision Making*, 8, 85–108.
- Wolfe, C. R., Fisher, C. R., & Reyna, V. F. (2013). Semantic coherence and inconsistency in estimating conditional probabilities. *Journal of Behavioral Decision Making*, 26, 237–246.
- 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. (2010). Semantic coherence and fallacies in estimating joint probabilities. *Journal of Behavioral Decision Making*, 23, 203–223.
- Wolfe, C. R., Reyna, V. F., & Brainerd, C. J. (2005). Fuzzy-trace theory: implications for transfer in teaching and learning. In J. P. Mestre (Eds.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 53–88). Greenwich, CT: Information Age Press.

- Wolfe, C. R., Reyna, V. F., Widmer, C. L., Brust-Renck, P. G., Cedillos, E. M., Hu, X., & Weil, A. M. (2016). Understanding genetic breast cancer risk: Processing loci of the BRCA Gist intelligent tutoring system. *Learning & Individual Differences*, 49, 178–189. doi:[10.1016/j.lindif.2016.06.009](https://doi.org/10.1016/j.lindif.2016.06.009)
- 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., 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.