JOURNAL OF LEARNING ANALYTICS

SOLAR SOCIETY for LEARNING ANALYTICS RESEARCH

Volume 5(3), 222–234. http://dx.doi.org/10.18608/jla.2018.53.14

A Method for Automatically Analyzing Intelligent Tutoring System Dialogues with Coh-Metrix

Christopher R. Wolfe[1], Colin L. Widmer[2], Christine V. Torrese [3], Mitchell Dandignac[4]

Abstract

We developed a method for using Coh-Metrix to automatically analyze tutorial dialogues. Coh-Metrix, a web-based tool for automatically evaluating text, is freely available to researchers. We applied the method to 190 tutorial dialogues between women and *BRCA Gist* from two experiments. *BRCA Gist* is an intelligent tutoring system (ITS) to help women make decisions about genetic testing for breast cancer risk. Tutorial dialogues scored high on measures of textual cohesion (deep cohesion, referential cohesion, and the composite variable formality). They also scored high on measures of the situation model (LSA verb overlap and causal verb and causal particle). However, there was mixed support for the hypothesis that higher scores on Coh-Metrix variables would predict subsequent comprehension. A Coh-Metrix principle is that the observable cohesion of a text is a reliable guide to the coherence of the reader's mental representation of that text. Thus it appears that interacting with *BRCA Gist* helped people form coherent mental representations of complex medical materials. We conclude that Coh-Metrix can be used to reliably assess tutorial dialogues and make inferences about the mental representations of people engaged in conversation with an ITS based on observable characteristics of the statements people make.

Notes for Practice

- Previous research suggests that learning is facilitated by tutorial dialogues; specifically, in the intelligent tutoring system, *BRCA Gist* tutorial dialogues contribute to learning.
- Coh-Metrix 3.0 is a web-based tool for the automatic evaluation of text and discourse that we applied to analyzing *BRCA Gist* tutorial dialogues about genetic breast cancer risk.
- In two studies, we found that the dialogues scored high on deep cohesion, referential cohesion, and the situation model variable LSA verb overlap. This indicates that participants' mental representations were coherent and conceptually well integrated at the level of actions.
- It appears that this method of using Coh-Metrix allows one to make reliable inferences about the mental representations of people engaged in conversation with an intelligent tutoring system.

Keywords

Discourse technology, tutoring, breast cancer education, textual assessment, simulation model

Submitted: 29.01.2018 — **Accepted:** 08.06.2018 — **Published:** 11.12.2018

Corresponding author [1] Email: WolfeCR@MiamiOH.edu Address: Department of Psychology, Miami University, Oxford, OH 45056 ORCID 0000-0001-6095-9457

[2] Email: colin.widmer@wright.edu Address: Wright State Research Institute, 4035 Colonel Glenn Highway, Beavercreek, PH 45431

[2] Email: torrescv@miamioh.edu Address: Department of Psychology, Miami University, Oxford, OH 45056

[4] Email: dandigm@miamioh.edu Address: Department of Psychology, Miami University, Oxford, OH 45056

1. Introduction

Individual one-on-one human tutoring is arguably the "gold standard" for facilitating deep conceptual understanding (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001). Expert tutors reportedly produce gains up to two standard deviations over typical teaching practice, sometimes labelled the "2 sigma effect" (Bloom, 1984), and ordinary tutors typically yielding effect sizes of about .79 compared to standard instructional methods (VanLehn, 2011). Intelligent tutoring systems (ITSs; Graesser, 2011; Wolfe, Fisher, Reyna, & Hu, 2012) are computer-based systems that mimic essential characteristics of one-on-one tutoring. Over a quarter of a century of research suggests that the best ITS yield gains comparable to many average human tutors (VanLehn, 2011). Given their importance, it would be practically and theoretically desirable to have a method to automatically evaluate tutorial dialogues using advanced discourse technologies such as Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014).

The purpose of this research is to develop a method for using Coh-Metrix to automatically evaluate tutorial dialogues, producing reliable quantitative measures of psycholinguistic variables. Our aim is to develop a broadly applicable method to better understand what makes for an effective tutorial dialogue and address basic research questions about discourse psychology. Although psycholinguistic concepts such as the situation model may be less familiar to some learning researchers, ultimately ours is the familiar psychological problem of making inferences about a person's internal state based on observable behaviour. In this case, the behaviour is the observable characteristics of the statements people make in tutorial dialogues, and Coh-Metrix provides analytic tools for assessing those statements. Our ultimate goal is to investigate the mental representation of tutorial content and the relationship between learning outcomes and the cognitive and psycholinguistic variables captured in tutorial dialogues. In addition, we seek to develop a method that could potentially be used to improve how an ITS (or human tutor) interacts with people, and perhaps in the future be used to guide real-time adjustments in ITS feedback to optimize learning.

1.1. The BRCA Gist Intelligent Tutoring System

BRCA Gist (BReast CAncer and Genetics Intelligent Semantic Tutoring; Wolfe, Reyna, et al., 2015) is a web-based ITS developed to help women understand and make decisions about genetic testing for breast cancer risk. *BRCA Gist* is grounded in Fuzzy-Trace Theory (Reyna, 2008; 2012) and was the first ITS applied to lay people's medical decision making (Wolfe, Widmer, et al., 2013). *BRCA Gist* was developed using AutoTutor Lite (Hu, Han, & Cai, 2008) and uses talking animated agents to present images, and video clips to teach complex concepts to everyday people. What makes it an "intelligent" tutoring system is the ability to engage in dialogues with people and respond differently depending on what they say by typing in a text box (Wolfe, Widmer, et al., 2013).

BRCA Gist is composed of modules that are punctuated by tutorial dialogues. After receiving didactic information about breast cancer and genetic risk in the form of speech, text, images, and video clips, participants engaged in five tutorial dialogues in which they had to generate brief "gist" explanations (Widmer et al., 2015) on the following topics: What is breast cancer? How does breast cancer spread? How do genes affect breast cancer? What should someone do with a negative test result? and What should someone do with a positive test result?

Gist explanations are much briefer than the more elaborate explanations produced by AutoTutor in teaching conceptual physics (Graesser et al., 2004). *BRCA Gist* primarily uses pumps and prompts (Wolfe, Reyna, et al., 2016) to help people develop good succinct answers to each of those questions.

BRCA Gist has been shown to effectively teach women about breast cancer in multiple randomized, controlled experiments conducted in laboratory, web, and field settings (Wolfe, Reyna, et al., 2012; Wolfe, Widmer, et al., 2013; Wolfe, Reyna, et al., 2015; Widmer et al., 2015). Women who interacted with *BRCA Gist* outperformed women who read comparable information on the National Cancer Institute website or interacted with an equally effortful ITS teaching an irrelevant topic (nutrition). Results consistently showed that those interacting with *BRCA Gist* performed significantly better (Wolfe, Reyna, et al., 2015; Widmer et al., 2015), with large effect sizes, on measures of knowledge ($\eta^2 = 0.23$), gist comprehension ($\eta^2 = 0.27$), and risk assessment ($\eta^2 = 0.23$). *BRCA Gist* participants were also more likely to appropriately recommend genetic testing on scenarios where people were at high genetic risk than for those with no genetic risk factors (Widmer et al., 2015; Wolfe, Reyna, et al., 2012; 2013; 2015).

In another controlled experiment, participants interacted with several versions of the tutor, including the full version of *BRCA Gist* and a version that was identical except for excluding the gist explanation dialogues (Wolfe, Reyna, et al., 2016). Participants who received the full version of *BRCA Gist* had a mean score of 76.6% correct on a declarative knowledge test, significantly higher than participants who received no gist explanation dialogues (M = 69.8% correct). These data provide direct evidence that the gist explanation dialogues add value to the outcomes above and beyond the exact same didactic presentation of relevant information (Wolfe, Widmer, et al., 2016).

Fine-grained analyses of the verbal interactions between women and *BRCA Gist* were also conducted, assessing the relationship between those verbal interactions and outcome measures (Wolfe, Reyna, et al., 2013; Widmer et al., 2015; Wolfe, Widmer, et al., 2016). *BRCA Gist* produces coverage scores (CO scores) that compare participants' verbal inputs to brief ideal answers using Latent Semantic Analysis (LSA; Wolfe, Reyna, et al., 2013). Researchers also scored the dialogues for coverage of content from the tutorial (for example, metastasis and genetic mutations) using a set of reliable rubrics. CO scores and rubric scores for the gist explanation dialogues were highly correlated ($p < .001$), indicating that *BRCA Gist* is successfully able to use CO scores to distinguish between good and poor responses (Widmer et al., 2015; Wolfe, Widmer, et al., 2016). CO scores were also good predictors of declarative knowledge in the majority of cases (Widmer et al., 2015; Wolfe, Widmer, et al., 2016).

Thus, we have good evidence that *BRCA Gist* yields good learning and decision making outcomes, that part of the locus of that effectiveness is engaging in gist explanation tutorial dialogues, and that the more people cover the content in their tutorial dialogues the higher they score on subsequent knowledge outcomes. However, these dialogues are an untapped resource holding a wealth of psycholinguistic information that could potentially shed light on key characteristics of effective

tutorial dialogues. Three psycholinguistic constructs with direct relevance to learning from tutorial dialogues are cohesion, the situation model, and formality. "Cohesion generally refers to the presence or absence of explicit cues in the text that allow the reader to make connections between the ideas in the text" (Crossley, Kyle, & McNamara, 2016, p. 1228). "Situation models are mental representations that model a reader's understanding of the meaning of a text" (Zwaan & Radvansky, 1998). Formality of text refers to a "linguistic system based on logic and/or mathematics that is distinguished by its clarity, explicitness, and simple verifiability" (Bussmann, 1996, p. 169). "Formal discourse tends to occur in well-prepared spoken or written texts in which there is a need to be precise, coherent, accurate, and persuasive to an educated audience" (Li & Graesser, 2016, p. 922). Until recently, analyzing verbal data of this kind required many subjective judgments or was a matter of qualitative interpretation. However, recent advances in discourse technologies make it possible to assess tutorial dialogues using quantitative measures. One tool for automatically assessing text and discourse is Coh-Metrix.

1.2. Coh-Metrix

Coh-Metrix 3.0 is a web-based tool for the automatic evaluation of text and discourse. Coh-Metrix is freely available to the research community for research purposes (see http://cohmetrix.com/). The "Coh" in Coh-Metrix stands for cohesion, and a basic principle underlying the computational approach is that the observable cohesion of a text, as measured by Coh-Metrix, is a reliable guide to the coherence of the reader's mental representation of that text. Coh-Metrix assesses text samples on 106 normed variables organized into the categories text "easability" principle components, situation model, referential cohesion, lexical diversity, connectives, syntactic complexity, syntactic pattern density, word information, readability, and descriptive statistics. In addition, Coh-Metrix can be used to calculate an over-arching cohesion variable formality. Of most significance for assessing tutorial dialogues are components of easability assessing cohesion, particularly deep cohesion and referential cohesion, and aspects of the situation model, especially Latent Semantic Analysis verb overlap, causal verb and causal particle incidence, and causal verb incidence. The components of easability assessing cohesion and the situation model are the most important for assessing tutorial dialogues because they measure the causal and functional relationships at the heart of understanding genetic breast cancer risk. Other variables such as lexical diversity, connectives, and syntactic complexity may tell us something about the learner, but they are unlikely to shed light on what and how people learn.

The "easability" of a text refers to its difficulty and complexity, and thereby how easy it is to read and comprehend (McNamara et al., 2014). Coh-Metrix includes the traditional measures of readability: Flesch-Kincaid Grade Level and Flesch reading ease. However, these traditional measures only calculate difficulty by looking at word length and sentence length within the text (McNamara et al., 2014). These superficial measures are too simplistic to measure the full nuance of how texts differ in difficulty and create different representations within the reader. For example, low knowledge readers benefit from more explanatory text that is more coherent while high knowledge learners benefit from text that is less coherent (McNamara, Kintsch, Butler Songer, & Kintsch, 1996).

Coh-Metrix includes eight principle components of text easability that examine the difficulty or easability of texts, with more precision. A percentile score and z-score is calculated for each component. These eight components provide information at multiple levels of linguistic analysis. They examine both the superficial aspects of words and sentences as well as the relationships between ideas in the text (McNamara et al., 2014). One component of text easability particularly relevant to assessing tutorial dialogues is deep cohesion (PCDCz), which measures the degree that the text contains causal and intentional connectives to highlight causal and logical relationships in the text (McNamara et al., 2014). Examples of these connectives include starting sentences with "as a result," "however," or "in contrast." These connectives help make the relationships between sentences more explicit so the reader does not have to make as many inferences (McNamara et al., 2014). Consider the following pair of sentences: "Drug A has been shown to treat breast cancer. The side effects could be fatal." This might be confusing because the two sentences seem to generate opposite ideas about the viability of Drug A. A connective would make the intended relationship more explicit and increase deep cohesion. Consider the amended sentences: "Drug A has been shown to treat breast cancer. However, the side effects could be fatal." Inserting the term "however" explicitly highlights that the second sentence is an exception, or condition, to what was stated in the first sentence, thus increasing the deep cohesion between the two ideas.

Another component of easability particularly relevant to assessing tutorial dialogues is referential cohesion (PCREFz), which measures the degree that words and ideas overlap across sentences, paragraphs, and the entire text. A text with high referential cohesion explicitly connects the sentences and paragraphs together by repeating the same words and ideas. A text low in referential cohesion means the reader has fewer explicit connections available to connect the main ideas of the text, typically making the text more challenging (McNamara et al., 2014). Consider the following passage: "Breast cancer can be a threat to life. It is formed in the tissue of the ducts and lobules." In this passage, "it" in the second sentences is a reference to "breast cancer" in the first sentence, providing some referential cohesion between the two sentences. However, referential cohesion would increase even more if the second sentence started with "breast cancer" instead of "it" because the connection

between the sentences would be made even more explicit. Conversely, a text that flits from theme to theme without continuity would be quite low in referential cohesion.

Text comprehension can differ across the local and global levels of the text. Local, or micro-level, processing examines the textbase of the actual words used but comprehending the textbase does not necessarily mean there is global or macro-level processing of the text (Kintsch & Van Dijk, 1978; Miller & Kintsch, 1980; Kintsch & Yarbrough, 1982). As people read, they generate inferences in developing a situation model (Graesser, Singer, & Trabasso, 1994). The situation model is a mental representation that integrates the meaning of words with the given context or situation. Situation models help people understand the underlying meaning of a text that cannot be generated by the explicit words alone. The situation model is important to consider because discerning the meaning of text requires inferences beyond the explicit textbase (McNamara et al., 2014). According to the Construction-Integration Model, situation models help promote global processing, which leads to a deeper understanding of the text (Kintsch, 1988; 1994). In Coh-Metrix, the situation model is assessed using variables that measure the causal and connective words that produce inferences about the deeper meaning of a text.

One variable that Coh-Metrix uses to capture the situation model is Latent Semantic Analysis Verb Overlap (SMCAUSlsa), which measures the extent to which verbs repeat across the text (McNamara et al., 2014). Using latent semantic analysis, words such as "spread" and "divide" have a degree of similarity (in the cancer context) so would increase the verb overlap if they occurred in the same text. Thus, LSA assesses conceptual overlap of verbs without relying on the use of the exact same words. Coh-Metrix also assesses the situation model through variables that examine words that reference causality within the text using incidence scores. Incidence scores mean the raw number of times a particular category of word shows up in the text. Causal verb incidence (SMCAUSv) includes words like "spread," "divide," and "move," which describe a change in the state of something mentioned in the text. Causal verb and causal particle incidence (SMCAUSvp) includes the previously mentioned causal verbs as well counting the number of connectives like "because" and "in order to," that connect a causal verb to the rest of the sentence. A more cohesive text is one with a higher proportion of connectives, or particles i.e., words, that relate a verb to the text (McNamara et al., 2014). Without these connectives, the reader will have to infer the relation between the actions and events within a text, decreasing cohesion and potentially creating a text that is more difficult to read.

Formality is a comprehensive measure of text easability that was created to add breadth to other unidimensional measures of text easability by establishing a formula combining several Coh-Metrix text easability components. Nelson, Perfetti, Liben, and Liben (2011) showed that five variables within Coh-Metrix related to text difficulty in four different samples of text so these were used to create formality. Formal text is high in cohesion and low in narrativity (storytelling text, typically using familiar language), syntactic simplicity (using fewer words and simpler sentence structure), and word concreteness (concrete words such as "blood" as opposed to abstract words such as "metastasis"). Formality is highly correlated with Flesch-Kincaid Grade Level (Graesser et al., 2014). Additionally, formality improves upon previous unidimensional measures because it offers a more sensitive scale to measure higher grade level text and narrative text. The formality formula using z-scores is:

(Deep Cohesion+Referential Cohesion-Narrativity-Syntactic Simplicity-Word Concreteness)/5

There are some linguistic analyses of tutorial dialogues. Veel (1999) studied student–teacher dialogues about mathematics, assessing variables such as lexical density from a sociolinguistic perspective. In a study of conversational agents teaching ninth grade biology, Dyke, Adamson, Howley, and Rosé (2013) found that re-voicing support (e.g., having the agent say, "So let me see if I've got your thinking right. You're saying…") significantly increased learning. Graesser and colleagues used Coh-Metrix to compare the tutorial dialogues of novice physics students to those who had already completed a college course interacting with the ITS AutoTutor (Graesser, Jeon, Yang, & Cai, 2007). They found differences in conceptual overlap as measured by Latent Semantic Analysis with the more advanced students demonstrating greater global knowledge. They also found that dialogues between an ITS and students were similar to those between human tutors and students, but very different than an analysis of textbooks on the same topics.

2. Theory and Hypothesis

A basic principle underlying Coh-Metrix is that the observable cohesion of a text is a reliable guide to the coherence of the reader's mental representation of that text. Our overarching hypothesis is corollary: that we can use Coh-Metrix to make inferences about the mental representation of interlocutors in tutorial dialogues (i.e., learners) based on the observable characteristics of the statements they make in those dialogues.

BRCA Gist is grounded in Fuzzy-Trace Theory (Reyna 2004; Wolfe, Reyna, & Brainerd, 2005). A key Fuzzy-Trace Theory concept is that information is encoded and represented through dual processes in parallel along a continuum with vague gist representations extracting the bottom line meaning at one end of the continuum and precise verbatim

representations capturing the superficial characteristics at the other. Here the word gist is used much as in everyday speech to express essential meaning. Converging evidence from published research suggests that although gist representations are sometimes the source of fallacious reasoning and suboptimal performance, forming appropriate gist representations, rather than precise verbatim recall, generally leads to superior judgment and decision making (Blalock & Reyna, 2016). Thus a key goal of *BRCA Gist* is to help women form appropriate and useful gist representations of concepts related to genetic breast cancer risk.

Fuzzy-Trace Theory suggests that what people say in these tutorial dialogues emphasizing gist should be reflected in measures of their gist comprehension of breast cancer and genetic risk as assessed by subsequent tests. We hypothesize that Coh-Metrix scores on relevant dimensions associated with the coherence of mental representations should correspond with measures of participants' gist comprehension of breast cancer and genetic risk as assessed by subsequent tests. Specifically, coherence variables such as referential cohesion and deep cohesion, as well as those associated with the situation model, such as Latent Semantic Analysis verb overlap, should correspond to participants' gist comprehension.

Thus the basic logic underlying this investigation is that learning about breast cancer and genetic risk through tutorials changes the mental representation of covered content, and that engaging in tutorial dialogues about those topics increases knowledge and further refines the coherence and integration of participants' situation models and gist representations. An analysis of what people say in those dialogues provides insights into their mental representations, and their performance of subsequent gist comprehension tasks is also a function of the quality of those gist representations.

In many studies of texts using Coh-Metrix, researchers assess and make predictions about the mental representations readers will likely develop of texts based on the observable characteristics of those texts. In particular (for readers with relatively little domain knowledge), the greater the proximal measures of the cohesion of texts, the greater the coherence of readers' mental representations. And the higher texts score on proximal variables related to the situation model, the more fully developed are the actual (cognitive) situation models developed by readers. We extend these constructs to tutorial dialogues. The speech produced by learners provides information about their mental representations. We theorize that the scores their speech yields on proximal Coh-Metrix variables assessing coherence and the situation model provide insights into their cognitive representations. This leads us to the following four hypotheses:

Hypothesis 1. Tutorial dialogues will score high on easability measures of cohesion, specifically deep cohesion (PCDCz) and referential cohesion (PCREFz).

Hypothesis 2. Tutorial dialogues will score high on measures of the situation model, specifically those assessing the representation of causal relationships, LSA verb overlap (SMCAUSlas), causal verb and causal particle (SMCAUSvp), and causal verb (SMCAUSv).

Hypothesis 3. Tutorial dialogues will score high on the measure formality based on the formula (Deep Cohesion+Referential Cohesion-Narrativity-Syntactic Simplicity-Word Concreteness)/5.

If these three hypotheses are supported, then a fourth hypothesis can be made pertaining to individual differences among participants in this successful group of participants who interacted with the *BRCA Gist* tutor — that Coh-Metrix variables assessing coherence and the situation model will predict learning outcomes, particularly measures of gist comprehension.

Hypothesis 4. Higher scores on easability (cohesion), the situation model, and formality will predict subsequent performance on measures of gist comprehension.

3. Material and Methods

3.1. Composition of the text corpus and data collection

The texts used in the Coh-Metrix analysis were taken from dialogues between *BRCA Gist* and women learning about genetic testing for breast cancer risk from two studies. Study 1 used dialogues taken from research conducted with undergraduate women in laboratory conditions. Study 2 used dialogues taken from field and web research conducted with a more diverse population. (See Wolfe, Reyna, et al. (2015) and Widmer et al. (2015) respectively for the methodological details of each *BRCA Gist* study.) The participants who generated the texts used in both studies were randomly assigned to engage in five tutorial dialogues in *BRCA Gist* on the following topics: What is breast cancer? How does breast cancer spread? How do genes affect breast cancer? What should someone do with a negative test result? and What should someone do with a positive test result? Participants who were randomly assigned to other conditions or who skipped one or more dialogues were

not included in these analyses. There were 18 participants who completed all five dialogues in Study 1 and 20 in Study 2, which provided a total of 190 tutorial dialogue texts for the Coh-Metrix analyses ([18+20] x 5 = 190). The dialogues were extracted from AutoTutor Lite web logs that captured every keystroke.

3.2. Preparing Raw Text Data for Coh-Metrix Analysis

Raw text data from the web logs was first "cleaned up" manually following best practices for pre-processing (Dowell, Graesser, & Cai, 2016), meaning that extraneous characters were removed and the data were put into the format of a dialogue. Below is a sample *BRCA Gist* dialogue with a woman on the topic, "What is breast cancer?"

Participant Input 1: Breast *cancer* is a when cancerous cells abnormally divide, grow, and do not die properly.
Response 1: You are off to a great start. Try to think about the specific characteristics of breast cancer.
Participant Input 2: It can be a threat to life.
Response 2: Good job. Can you say something about the growth of normal cells compared to cancer cells?
Participant Input 3: It is formed in the tissue of the ducts and lobules.
Response 3: Okay. Can you add something new about normal BRCA genes and BRCA mutations?
Participant Input 4: It is formed in the tissue of the ducts and lobules.
Response 4:
Participant Input 5: *BRACA* mutations make women more at risk for breast cancer.
Response 5: Alright. Can you say something else about different kinds of tumors?
Participant Input 6: Benign tumors are not usually life threatening, but malignant tumors can spread and threaten life.
Response 6:
Participant Input 7: Women are at risk if family members have had ovarian cancer.
Response 7: Great job. Please click on the finish button to continue.

As illustrated above, participants made some typos and spelling mistakes indicated above with italics. *BRCA Gist* proved robust to these errors but for the purposes of Coh-Metrix analysis we spell checked the texts with MS Word. In addition, occasionally a participant would repeat the same sentence twice by hitting "return" on the keyboard twice in rapid succession (underlined in the example above). This did not affect *BRCA Gist*'s internal coverage scores, but would affect Coh-Metrix variables. Thus, repeated sentences were removed manually. Finally, because the goal of *BRCA Gist* dialogues was to help people form good answers to the five questions, we removed everything *BRCA Gist* said, leaving the words of the participant preserved as a single paragraph for each question. This is justified for a system that uses pumps and prompts in this way (Wolfe, Widmer, et al., 2016) but may be less so for tutorial dialogues in which the tutor integrates the presentation of didactic materials into back-and-forth conversations with the learner. Below is an example of the dialogue above as prepared for Coh-Metrix analysis:

Breast cancer is a when cancerous cells abnormally divide, grow, and do not die properly. It can be a threat to life. It is formed in the tissue of the ducts and lobules. BRCA mutations make women more at risk for breast cancer. Benign tumors are not usually life threatening, but malignant tumors can spread and threaten life. Women are at risk if family members have had ovarian cancer.

Next we combined all five gist explanation paragraphs into a single text for each person with blank lines separating the paragraphs and the subject number on top. The set of five paragraphs were copied and pasted into the textbox at http://cohmetrix.com for analysis by Coh-Metrix 3.0 (the subject number was excluded from the analysis). For each one, after clicking on Submit, we clicked on Save Data to create a spreadsheet that was then imported for statistical analysis, as described below.

3.3. Analytic Approach

Given that Coh-Metrix yields 106 variables for each text, it is important to limit the problem space in a principled way. We argue that replication is part of an effective strategy to assess reliability and minimize the likelihood of over-interpreting chance findings. Thus we analyzed tutorial dialogues from two studies with different populations. Starting with two data sets reduces the chance of falling victim to idiosyncratic characteristics of a particular set of texts. Descriptive statistics were benchmarked against norms for social studies texts written at an 11th grade to adult level according to common core educational standards from the Touchstone Applied Science Associates (TASA) text corpus, found in McNamara et al. (2014), Appendix B.

As a measure of learning and understanding, we used Gist Comprehension of Genetic Breast Cancer Risk (Wolfe, Reyna, et al., 2013; Fisher et al., 2013). This 40-item, 1–7 Likert-scale instrument assesses gist comprehension of key information on breast cancer and genetic testing. Items such as "the greatest danger of dying from breast cancer is when it spreads to other parts of the body" and "Genetic risk factors account for the majority of breast cancer cases" (reversed scored) express

the gist of that knowledge — the essential bottom-line meaning. People can express different degrees of agreement with such statements without remembering the precise verbatim details. Cronbach alpha for Gist Comprehension is 0.85 (see Wolfe, Reyna, et al., 2015, Appendix B for the Gist Comprehension instrument).

We conducted regression analysis predicting gist comprehension at the level of superordinate categories text easability principal component z-scores and situation model, including all variables in a particular set. McNamara et al. (2014, p. 86) note that "percentile and z-scores have a monotonic but not a linear relationship to each other. Generally the z-scores are the preferred scores for research and statistical purposes." Including all variables within a category (e.g., situation model) prevents "cherry picking." We did not analyze the data on other dimensions such as lexical diversity choosing *a priori* only those variables that make sense theoretically. We calculated formality scores using the formula (Deep Cohesion+Referential Cohesion-Narrativity-Syntactic Simplicity-Word Concreteness)/5.

4. Results

4.1. Describing the Dialogues with Coh-Metrix

Text easability principle components were highly consistent across both replications. Table 1 presents the z-score and percentile score for narrativity, syntactic simplicity, word concreteness, referential cohesion, deep cohesion, verb cohesion, connectivity, and temporality. For purposes of comparison, as an index to typical texts, Table 1 presents norms from McNamara and colleagues (2014) for social studies texts written at an 11th grade to adult level according to common core educational standards.

Table 1. Easability z-score means and standard deviations with percentile by study (with norms)

Variable (Z-score)	Study 1	Study 2	Norms (TASA Common Core Grade 11+ Social Studies)*
PCNARz Narrativity	–.179 (Percentile 44, SD = .44)	–.222 (Percentile 42, SD = 43)	–.742 (SD = .572)
PCSYNz Syntactic Simplicity	.677 (Percentile 72, SD = .61)	.768 (Percentile 74, SD = .62)	–.101 (SD = .746)
PCCNCz Word Concreteness	.290 (Percentile 60, SD = .54)	.211 (Percentile 58, SD = .66)	.034 (SD = .964)
PCREFz Referential Cohesion	1.892 (Percentile 94, SD = .70)	2.124 (Percentile 96, SD = .74)	–.310 (SD = .855)
PCDCz Deep Cohesion	.915 (Percentile 76, SD = .82)	1.086 (Percentile 80, SD = .81)	.366 (SD = .366)
PCVERBz Verb Cohesion	–.257 (Percentile 41, SD = .67)	–.372 (Percentile 37, SD = .48)	–.289 (SD = .977)
PCCONNz Connectivity	–2.722 (Percentile 4, SD = 1.18)	–2.907 (Percentile 1, SD = .76)	–2.254 (SD = 1.276)
PCTEMPz Temporality	–.111 (Percentile 46, SD = .46)	–.323 (Percentile 39, SD = .55)	–.154 (SD = 1.085)

* Norms from Touchstone Applied Science Associates (TASA) 11th grade to adult social studies Common Core text corpus, McNamara et al. (2014) Appendix B.

As predicted in Hypothesis 1, the dialogues scored quite high on deep cohesion and referential cohesion. For referential cohesion the tutorial dialogues consistently scored above the 94th percentile, and for deep cohesion they consistently scored above the 76th percentile. This suggests that interacting with *BRCA Gist* helped people make highly cohesive statements on issues pertaining to genetic breast cancer risk. They also scored high on syntactic simplicity, which may be an artifact of the dialogue format given that speech is typically less syntactically complex than prose. The dialogues scored quite low on connectivity. This component examines the degree to which text contains connectives that explicitly express relations in the text and the number of logical relations explicitly stated. Many participants scored zero on this dimension, which may also be an artifact of the dialogue format.

Situation model scores were highly consistent across both replications. Table 2 presents the situation model variables causal verb, causal verb and causal particle, intentional verb, ratio causal particles to causal verbs, ratio intentional particles to partial verbs, LSA verb overlap, WordNet verb overlap, and temporal cohesion. In Table 2, we created estimated z-scores by taking the mean and subtracting the normative mean divided by the normative standard deviation. We distinguish these calculated z-scores from those provided by Coh-Metrix by calling them estimated z-scores. Again we provide norms for 11th grade to adult social studies texts for the purposes of comparison.

Table 2. Situation Model means and standard deviations with estimated z-scores by study (with norms)

Variable	Study 1	Study 2	Norms (TASA Common Core Grade 11+ Social Studies)*
SMCAUSv Causal Verb	33.761 (SD = 11.66, Estimated Z = .43)	32.062 (SD = 10.35, Estimated Z = .28)	29.043 (SD = 10.936)
SMCAUSvp Causal Verb & Causal Particle	48.566 (SD = 11.82, Estimated Z = .78)	46.787 (SD = 12.34, Estimated Z = .64)	38.772 (SD = 12.597)
SMINTEp Intentional Verb	16.141 (SD = 6.48, Estimated Z = –.21)	15.571 (SD = 6.56, Estimated Z = –.27)	18.227 (SD = 9.953)
SMCAUSr Ratio Causal Particles to Causal Verbs	.480 (SD = .30, Estimated Z = .46)	.484 (SD = .32, Estimated Z = .47)	.346 (SD = .291)
SMINTEr Ratio Intentional Particles to Partial Verbs	.851 (SD = .71, Estimated Z = –.33)	.938 (SD = .63, Estimated Z = –.31)	1.268 (SD = 1.061)
SMCAUSlas LSA Verb Overlap	.154 (SD = .03, Estimated Z = 3.61)	.151 (SD = .03, Estimated Z = 3.53)	.0097 (SD = .040)
SMCAUSwn WordNet Verb Overlap	.556 (SD = .07, Estimated Z = .03)	.569 (SD = 0.7, Estimated Z = .17)	.553 (SD = .096)
SNTEMP Temporal Cohesion	.826 (SD = .04, Estimated Z = .08)	.802 (SD = .05, Estimated Z = –.15)	.818 (SD = .105)

* Norms (also used to calculate estimated z-scores) from Touchstone Applied Science Associates (TASA) 11th grade to adult social studies common core text corpus, McNamara et al. (2014) Appendix B.

Consistent with Hypothesis 2, dialogues scored very high on situation model scores for LSA verb overlap. This means that participant utterances were conceptually linked together quite strongly even though people did not always use the exact same words. As predicted, the dialogues also scored high on causal verb and causal particle, suggesting a good understanding of the causal relationships associated with breast cancer and genetic testing. However, contrary to expectations, the dialogues did not score high on causal verbs.

With respect to formality, Study 1 produced a mean of .404 (SD = .31) and Study 2 yielded a mean of .491 (SD = .26). Examining the formality of texts for different ages and genres, Graesser et al. (2014) report mean formality sores of less than .20 for 11th grade to adult social studies texts, and less than .25 for science texts. Thus, as predicted by Hypothesis 3, these tutorial dialogues were consistently quite high on formality.

4.2. Predicting Gist Comprehension of Breast Cancer with Coh-Metrix:

Hypothesis 4, that Coh-Metrix variables would predict gist comprehension, was not consistently supported by the data. First, we conducted a regression analysis predicting gist comprehension by all of the easability variables. The overall regression equation model was not a significant predictor for Study 1: $F(8, 18)=1.99$, $p=0.11$ with $R^2=.47$, and adjusted $R^2=.23$. However, it was significant for Study 2: $F(8, 20)=3.66$, $p<0.009$ with $R^2=.59$, and adjusted $R^2=.43$. Table 3 presents the regression formula (intercept and Beta weights) for gist comprehension by the easability variables. We expected a negative relationship between word concreteness and gist comprehension such that the more participants used concrete words, the lower they scored on gist comprehension of breast cancer. This is precisely what we found in Study 2, but although word concreteness was a significant predictor of gist comprehension in Study 1 as well, the Beta weight is negative rather than positive. In Study 1 the greater the degree of connectivity the higher the gist comprehension score, though this effect was not replicated in Study 2. Contrary to expectations, deep cohesion and referential cohesion were not significant predictors of gist comprehension.

Table 3. Regression Beta weights predicting Gist Comprehension for easability z-score variables by study

Variable (Z-score)	Study 1	Study 2
Intercept	4.4469	5.1503
PCNARz Narrativity	–0.5332	–0.1776
PCSYNz Syntactic Simplicity	–0.1165	–0.0929
PCCNCz Word Concreteness	–1.0446*	0.6376**
PCREFz Referential Cohesion	0.1992	0.1213
PCDCzDeep Cohesion	–0.2113	–0.1691
PCVERBz Verb Cohesion	–0.3080	0.3162
PCCONNz Connectivity	0.3619ŧ	–0.0644
PCTEMPz Temporality	–0.3021	0.3366

* $p = 0.018$, ŧ $p = 0.009$, ** $p = 0.010$

Next we conducted a regression analysis predicting gist comprehension with all of the situation model variables. The overall regression equation model was not significant for Study 1: $F(8, 18) = 1.75$, R^2=.44, adjusted R^2=.19. However, the overall regression model was a significant predictor of gist comprehension of breast cancer for Study 2: $F(2, 20) = 4.78$, p=.002, R^2= .66, adjusted R^2=.52. Intentional verbs were a significant predictor of gist comprehension in both studies, but the Beta weight was in the opposite direction — negative in Study 1 but positive in Study 2. Contrary to expectations, LSA verb overlap was a negative predictor in Study 2 but the related variable WordNet Verb Overlap was a positive predictor of gist comprehension in Study 2 (neither was significant in Study 1 though the non-significant trends were in the same direction; see Table 4). Unlike LSA verb overlap, WordNet represents verb overlap in binary yes or no terms, so because "divide" and "spread" are different words (albeit with similar meanings in the context of cancer) they would not increase WordNet verb overlap if they appeared together (McNamara et al., 2014). A text that repeats just the word "spread" would have higher verb overlap within the WordNet verb overlap.

Table 4. Regression Beta weights predicting Gist Comprehension for situation model variables by study

Variable	Study 1	Study 2
Intercept	7.8958	6.0544
SMCAUSv Causal Verb	–0.0209	0.0885
SMCAUSvp Causal Verb & Causal Particle	0.0157	–0.0916
SMINTEp Intentional Verb	–0.0751*	0.0193**
SMCAUSr Ratio Causal Particles to Causal Verbs	–0.6395	1.4816
SMINTEr Ratio Intentional Particles to Partial Verbs	–0.0745	–0.1334
SMCAUSlas LSA Verb Overlap	–6.1723	–8.656Ŧ
SMCAUSwn WordNet Verb Overlap	1.2442	3.5600ŧ
SNTEMP Temporal Cohesion	–0.8279	–1.1566

* p = 0.009, ** p = 0.009, Ŧ p = 0.012, ŧ p = 0.022

Finally, we did not find an association between gist comprehension and formality in either study: for Study 1 $F<1$, $p>.05$; for Study 2 $F<1$, $p>.05$.

5. Discussion

5.1. Discussion of Results

We found comparable results in both replicates for our first three hypotheses. Theoretically, greater cohesion in tutorial dialogues indicates greater coherence in mental representations. As predicted in Hypothesis 1, the dialogues were high on deep cohesion and referential cohesion as measured in z-scores. This suggests that interacting with *BRCA Gist* helped people form coherent mental representations of issues pertaining to genetic breast cancer risk. The dialogues scored very low on connectivity, i.e., the degree to which the text contains connectives that explicitly express relations and the number of explicitly stated logical relations. It may be that the dialogue format leads people to make connections implicit rather than explicit because of the statements made by the *BRCA Gist* animated agent. Perhaps participants were following Grice's (1975) given-new contract at a finer level of discourse. As predicted in Hypothesis 3, the tutorial dialogues scored uniformly high on the composite cohesiveness variable formality. This suggests that interacting with *BRCA Gist* helped people form mental representations that are clear, coherent, and logical.

As predicted in Hypothesis 2, dialogues earned very high scores on the situation model variable LSA verb overlap. This indicates that participants' representations were conceptually well integrated at the level of actions even though people did not always use the exact same words. The dialogues also scored high on causal verb and causal particle, further suggesting a good understanding of the causal relationships associated with genetic testing and breast cancer risk. However, the dialogues did not score high on causal verbs. It may be that in the context of genetic testing, which neither causes nor is caused by breast cancer, the combination of causal particles and verbs is indicative of understanding.

Hypothesis 4 received mixed support. We predicted that higher scores on easability, the situation model, and formality would be associated with subsequent gist comprehension. There is some evidence that high text cohesion and a well-developed situation model in tutorial dialogues leads to greater gist comprehension. On some variables in one study we found this — yet the other study sometimes produced the opposite result. Coh-Metrix permits researchers to perform many *post hoc* tests, and we were able to rule out some explanations. For example, we get the same pattern of results controlling for the length of texts in words. Analyzing the variables word concreteness and intentional verb alone did not change the direction of the beta weights in the regression formulae, suggesting that the contradictory results are not an artifact of co-linearity.

It is possible that some of the differences between studies represent knowledge differences between the undergraduate sample used in Study 1 and the more diverse sample in Study 2 that included college graduates and people with advanced degrees. McNamara and colleagues (1996) found that low knowledge readers benefit from more coherent texts whereas high knowledge learners benefited from text that were less coherent. Participants in both studies benefitted from their interactions with Coh-Metrix (Wolfe, Reyna, et al., 2015; Widmer et al., 2015). Moreover, their responses were quite similar (see Tables 1 and 2). However, it is not known whether pre-existing knowledge on genetic breast cancer risk affected the relationship between participants' responses on dimensions such as deep cohesion and subsequent gist comprehension.

It could be that some underlying variable with a complex relationship to coherence and the situation model explains the contradictory results from the two studies. Fuzzy-Trace Theory suggests that the key to interventions such as *BRCA Gist* is helping people form a useful gist representation (Reyna, 2004; 2008; 2012). Exploring whether Coh-Metrix variables can be used to assess the extent to which readers and speakers have more gist-like or verbatim-like mental representations is a promising avenue for further research. Other factors not measured such as learning strategies and metacognitive reasoning may also be important.

One shortcoming of this research is that there is not a "gold standard" for assessing mental representations to compare with the results of our Coh-Metrix analyses. Our use of replication provides a high degree of confidence in the reliability of these measures because both replicates produced comparable Coh-Metrix scores on the variables of interest. Of course, we frequently make inferences about mental representations based on proximal variables, for example, when we infer that someone does not know much about a topic because they receive a low test score. Psycholinguistic variables such as the situation model are generally assessed by asking inferential questions — which would be difficult, if not impossible to do with "texts" generated by participants themselves.

A second potential weakness of this investigation is that we did not assess knowledge or gist comprehension of genetic breast cancer risk prior to participants' interactions with the *BRCA Gist* ITS. We have experimental evidence that people who interact with *BRCA Gist* perform better than controls (Wolfe, Reyna, et al., 2015), and that the version of *BRCA Gist* with tutorial dialogues yields significantly better performance than an identical version without tutorial dialogues (Wolfe, Widmer, et al., 2016). However, in this investigation we cannot determine the extent to which differences in performance between Study 1 and Study 2 with respect to the hypothesized relationship between gist comprehension and specified Coh-Metrix variables are a result of differences in pre-existing knowledge, and which are the result of differences in the quality of those interactions, or noise in the data.

Another potential shortcoming is determining the optimal level or unit of analysis for analyzing the tutorial dialogues. In this work, we treated each of the five dialogues engaged in by each participant as a single text of five paragraphs. However, it might be that a more fine-grained analysis would yield more nuanced results, for example by analyzing conversations about "what is breast cancer?" and "how do genes affect breast cancer risk?" separately, and then correlating Coh-Metrix with a smaller set of gist comprehension questions. We believe our approach is defensible, and that it avoids the possibility of "cherry picking" results. However, it may also gloss over important subtleties in the verbal data. Ultimately, the best level of analysis should be dictated by the research questions.

5.2. Discussion of Methodology

The decision to build in replication from the start proved to be astute. To illustrate, had we only conducted our analysis on the Study 1 data, we might have been drawn to the conclusion that word concreteness and intentional verbs both decrease gist comprehension, and if we had only examined the data from Study 2 we might have concluded that word concreteness and intentional verbs both increase gist comprehension. Conversely, and perhaps more importantly, the consistency of findings across studies give us greater confidence that high levels of cohesion, formality, and causal dimensions of the situation model are essential characteristics of these successful tutorial dialogues.

Entering clustered variables (e.g., all easability, situation model) together is theoretically sensible and reduces the problem space. However, there may be better approaches. A reasonable alternative to the built-in replications we employed is to develop a model with one data set, and then test it on another. For example, a variable such as temporality has few implications for the *BRCA Gist* tutorial. Rather than automatically including it, researchers might develop their best model with one data set (in this case excluding temporality) and then test that best model with a second data set. This is an antidote to cherry picking. However, it forces the researcher to make potentially dozens or even hundreds of decisions about what to include and exclude from the model. Thus having *a priori* hypotheses is of paramount importance even if the guiding theory is indifferent with respect to many modelling decisions.

Treating all participant input as a single text is justified give the *BRCA Gist* goal of helping people form good answers. This method produced reliable results that were predicted and theoretically meaningful. However, for tutors engaged in question answering or mixed initiative dialogues, this procedure may not be warranted. Under circumstances such as this, researchers may wish to analyze the data both ways — with and without the verbal contributions of the tutor to assess how this changes the scores for relevant variables. Analyzing dialogues in this way also makes it difficult to assess how specific

tutor statements correspond to changes in coherence and the situation model. Yet restricting the texts to those produced by the learner make it conceptually more straightforward to make inferences about the mental representations of the learner.

Coh-Metrix does not offer advice on how to write texts to raise or lower specific variable values, and provides even less guidance on how to structure tutorial dialogues to maximize deep cohesion and other variables. Future work on authoring texts and engineering ITS dialogues to score high on desirable Coh-Metrix dimensions such as cohesion and causal aspects of the situation model would make for a significant contribution. Further work on Coh-Metrix and gist formation following Fuzzy-Trace Theory may shed light on these relationships.

6. Conclusion

The results of this investigation suggest that interacting with *BRCA Gist* helped people form mental representations that are clear, coherent, logical, and conceptually well integrated at the level of actions. These results are encouraging about the prospect of following the method presented here using Coh-Metrix to make reliable inferences about the mental representations of people engaged in conversation with an intelligent tutoring system.

Declaration of Conflict of Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The publication of this article received financial support from the National Cancer Institute, award number R21CA149796 for the original data collection. 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.

Acknowledgments

We thank the National Cancer Institute for their support and Savannah J. Kreiner and Kennedy R. Braum for their valuable assistance with analyzing the data.

References

Blalock, S. J., & Reyna, V. F. (2016). Using Fuzzy-Trace Theory to understand and improve health judgments, decisions, and behaviors: A literature review. *Health Psychology, 35*, 781–792. http://dx.doi.org/10.1037/hea0000384

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*(6), 4–16. http://dx.doi.org/10.3102/0013189X013006004

Bussmann, H. (1996). *Routledge dictionary of language and linguistics*. London: Routledge.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., &, Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, *25*(4), 471–533. http://dx.doi.org/10.1207/s15516709cog2504_1

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*, 1227–1237. http://dx.doi.org/10.3758/s13428-015-0651-7

Dowell, N. M. M., Graesser, A. C., & Cai, Z. (2016). Language and discourse analysis with Coh-Metrix: Applications from educational material to learning environments at scale. *Journal of Learning Analytics, 3*, 72–95. http://dx.doi.org/10.18608/jla.2016.33.5

Dyke, G., Adamson, D., Howley, I., & Rosé, C. P. (2013). Enhancing scientific reasoning and discussion with conversational agents. *IEEE Transactions on Learning Technologies, 6*, 240–247. http://dx.doi.org/10.1109/TLT.2013.25

Fisher, C. R., Wolfe, C. R., Reyna, V. F., Widmer, C. L., Cedillos, E. M., & Brust-Renck, P. G. (2013). A signal detection analysis of gist-based discrimination of genetic breast cancer risk. *Behavior Research Methods*, *45*, 613–622. http://dx.doi.org/10.3758/s13428-013-0364-8)

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

Graesser, A. C., Jeon, M., Yang, Y., & Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. *Information Design Journal, 15*, 199–213. http://dx.doi.org/10.1075/idj.15.3.02gra

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. http://dx.doi.org/10.3758/BF03195563

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text

characteristics at multiple levels of language and discourse. *The Elementary School Journal, 115(2)*, 210–229. http://dx.doi.org/10.1086/678293

Graesser, A., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371–395.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech Acts* (pp. 41–58). New York: Academic Press.

Hu, X., Han, L., & Cai, Z. (2008). Semantic decomposition of student's contributions: An implementation of LCC in AutoTutor Lite. Paper presented at the *Society for Computers in Psychology*, 13 November 2008, Chicago, IL.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163–182.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49*(4), 294–303.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394. http://dx.doi.org/10.1037/0033-295X.85.5.363

Kintsch, W., & Yarbrough, C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology, 74*(6), 828–834. http://dx.doi.org/10.1037/0022-0663.74.6.828

Li, H. Y., & Graesser, A. C. (2016). Formality of the Chinese collective leadership. *Behavior Research Methods, 48*, 922–935. http://dx.doi.org/10.3758/s13428-016-0775-4

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press.

McNamara, D. S., Kintsch, E., Butler Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1–43. http://dx.doi.org/10.1207/s1532690xci1401_1

Miller, J., & Kintsch, W. (1980). Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory, 6*(4), 335–354. http://dx.doi.org/10.1037/0278-7393.6.4.335

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2011). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. New York: Student Achievement Partners.

Reyna, V. F. (2004). How people make decisions that involve risk: A dual-processes approach. *Current Directions in Psychological Science, 13*, 60–66. http://dx.doi.org/10.1111/j.0963-7214.2004.00275.x

Reyna, V. F. (2008). Theories of medical decision making and health: An evidence-based approach. *Medical Decision Making*, *28*(6), 829–833. http://dx.doi.org/10.1177/0272989X08327069

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in Fuzzy-Trace Theory. *Judgment and Decision Making*, *7*(3), 332–359.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. http://dx.doi.org/10.1080/00461520.2011.611369

Veel, R. (1999), Language, knowledge and authority in school mathematics. In F. Christie (Ed.), *Pedagogy and the shaping of consciousness: Linguistic and social processes* (pp. 185–216). London: Cassell.

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. http://dx.doi.org/10.3758/s13428-015-0592-1

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., & Brainerd, C. J. (2005). Fuzzy-Trace Theory: Implications for transfer in teaching and learning. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 53–88). Greenwich, CT: Information Age Press.

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

Wolfe, C. R., Reyna, V. F., Widmer, C. L., Cedillos, E. M., Fisher, C. R., Brust-Renck, P. G., Chaudhry, S., & Damas Vannucchi, I. (2013). Efficacy of a web-based intelligent tutoring system on genetic testing for breast cancer risk. Paper presented at 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. http://dx.doi.org/0272989X14535983

Wolfe, C. R., Reyna, V. F., Widmer, C. L., Cedillos, E., Weil, A. M., & Brust-Renck, P. G. (2016). Pumps and prompts for

gist explanations in tutorial dialogues about breast cancer. *Discourse Processes, 55*(1). http://dx.doi.org/10.1080/0163853X.2016.1199626

Wolfe, C. R., Widmer, C. L., Reyna, V. F., Hu, X., Cedillos, E. M., Fisher, C. R., Brust-Renck, P. G., Williams, T. C., Damas Vannucchi, I., & Weil, A. M. (2013). The development and analysis of tutorial dialogues in AutoTutor Lite. *Behavior Research Methods*, *45*, 623–636. http://dx.doi.org/10.3758/s13428-013-0352-z

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

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162–185.