Wolfe, C. R., Dandignac, M., & Reyna, V. F. (2019). A theoretically motivated method for automatically evaluating texts for gist inferences. *Behavior Research Methods, 51*(6), 2419-2437. https://doi.org/10.3758/s13428-019-01284-4

Behavior Research Methods (2019) 51:2419–2437
https://doi.org/10.3758/s13428-019-01284-4

A theoretically motivated method for automatically evaluating texts for gist inferences

Christopher R. Wolfe[1] · Mitchell Dandignac[1] · Valerie F. Reyna[2]

Published online: 24 July 2019

Abstract

We developed a method to automatically assess texts for features that help readers produce gist inferences. Following fuzzy-trace theory, we used a procedure in which participants recalled events under gist or verbatim instructions. Applying Coh-Metrix, we analyzed written responses in order to create gist inference scores (GISs), or seven variables converted to Z scores and averaged, which assess the potential for readers to form gist inferences from observable text characteristics. Coh-Metrix measures reflect referential cohesion and deep cohesion, which increase GIS because they facilitate coherent mental representations. Conversely, word concreteness, hypernymy for nouns and verbs (specificity), and imageability decrease GIS, because they promote verbatim representations. Also, the difference between abstract verb overlap among sentences (using latent semantic analysis) and more concrete verb overlap (using WordNet) should enhance coherent gist inferences, rather than verbatim memory for specific verbs. In the first study, gist condition responses scored nearly two standard deviations higher on GIS than did the verbatim condition responses. Predictions based on GIS were confirmed in two text analysis studies of 50 scientific journal article texts and 50 news articles and editorials. Texts from the Discussion sections of psychology journal articles scored significantly higher on GIS than did texts from the Method sections of the same journal articles. News reports also scored significantly lower than editorials on the same topics from the same news outlets. GIS proved better at discriminating among texts than did alternative formulae. In a behavioral experiment with closely matched text pairs, people randomly assigned to high-GIS versions scored significantly higher on knowledge and comprehension.

Keywords Discourse technology · Readability · Patient education · Understanding

In many important contexts, essential meaning is conveyed in printed words. Consider the words in messages about medications, such as "Harmful or fatal if swallowed." For many such texts, even short ones, understanding the meaning of the text requires the reader to make a number of inferences beyond the stated words on a page. Even the familiar phrase "harmful or fatal if swallowed" requires one to infer that it is the contents of the bottle, rather than swallowing the bottle itself, that are dangerous. However, research suggests that texts differ in the ease with which they facilitate such inferences (Allen, Jacovina, & McNamara, 2016). Thus, it is practically useful and theoretically important to have methods for assessing texts for their capacity to allow readers to make appropriate inferences.

Inferring cognitive processes in reading based solely on observable characteristics of texts is an audacious undertaking with clear limitations. One obvious issue is the lack of information about the reader. For example, when focusing exclusively on text we know nothing about the reader's reading proficiency, domain knowledge, goals, interests, or motivation. Yet the approach is not unprecedented. Perhaps the best known example is Flesch–Kincaid Reading Grade Level (FKGL), which serves as a metric for government communications and is embedded in modern word-processing software. Researchers have assessed the "readability" of texts using the basic approach of Flesch–Kincaid since Rudolf Flesch published "A New Readability Yardstick" in 1948 (Flesch, 1948). Flesch–Kincaid is a "data lean" approach to assessing readability in that it relies solely on the number of words per sentence and the number of syllables per word. The specific formula for FKGL is 0.39 × (words/sentences) + 11.8 × (syllables/words) – 15.59.

✉ Christopher R. Wolfe
WolfeCR@MiamiOH.edu

[1] Miami University, Oxford, OH, USA

[2] Cornell University, Ithaca, NY, USA

Flesch–Kincaid has proved useful for many purposes, including helping to make texts that convey complex medical information more understandable to laypeople. However, it is not without shortcomings. Two issues that have received little attention are reification and reverse engineering. The problem of reification of proximal variables is not so much a shortcoming of Flesch–Kincaid as it is of the ways people use it. Rather than acknowledging that Flesch–Kincaid is a proximal index to the ease with which readers with different levels of formal education will read a text, too many users treat it as synonymous with readability—that Flesch–Kincaid *is* readability. This is problematic because it suggests that if FKGL is at an appropriate level, authors need not take any additional actions to make their texts understandable.

A second problem of FKGL is one of reverse engineering or editing text to grade level. Just because the book *Go Dog Go!* has shorter words and shorter sentences than *War and Peace* does not mean that revising one's texts to short, choppy sentences automatically means that they are understood better by everyone. To illustrate, consider the importance of the three-syllable word "however." The following text has an FKGL of 10.33: "Tamoxifen is a selective estrogen receptor modulator that can reduce the risk of the cancer recurring by 50% in premenopausal women. However, it can cause serious side effects, including blood clots and stroke." Removing the word "however" reduces the average number of words per sentence and the average number of syllables per word, thus reducing Flesch–Kincaid to 9.76, over half a grade level lower than the unedited text. Yet the word "however" is helpful, because it signals the reader in real time of a switch from discussing positive to negative aspects of the drug. Unfortunately, it is easy to edit texts to reduce FKGL in ways that actually undermine their understandability.

In contrast to FKGL's emphasis on surface "verbatim" features of text, the goal of our research has been to develop a proximal index to the likelihood with which readers will develop useful and appropriate "gist inferences" from a given text. This work is guided by fuzzy-trace theory (FTT), an approach supported by over two decades of research (see Reyna, 2008; Reyna & Brainerd, 1995; Wolfe et al., 2015). FTT is a dual-process theory emphasizing meaning making and mental representation. A key tenet of FTT is that people simultaneously encode information into multiple representations that range from verbatim surface details to the bottom line meaning or gist. Here the term "gist" is used much as it is in everyday usage to reflect simple core meaning. FTT suggests that people independently process both gist and verbatim representations of experiences.

Thus, gist representations capture the bottom-line meaning, whereas verbatim representations emphasize surface details. FTT indicates that gist and verbatim processing occurs in parallel, rather than gist being derived from verbatim representations as in earlier theories. Moreover, people prefer to reason with the most gist-like mental representation available for a given task. Empirical evidence suggests that the reasoning of domain experts is even more gist- like than novices (Reyna & Lloyd, 2006) and that the process of cognitive development is one of migrating from relying on verbatim to increasingly relying on gist representations (Brainerd, Reyna, & Holliday, 2018). Thus, reasoning and decision making are advanced by helping people develop appropriate gist representations, which emphasizes meaning including inferences that go beyond the literal text (e.g., Reyna & Kiernan, 1994) rather than emphasizing verbatim details.

The act of reading requires people to make an astonishing array of inferences (Magliano & Graesser, 1991). In a classic review, Graesser, Singer, and Trabasso (1994) identified 13 different classes of inferences, ranging from referential and case-structure role assignment to the author's intent. Psycholinguistic researchers differ theoretically in whether they posit that inferences pertaining to superordinate goals, themes, and causal consequences are made online, "in real time," during the act of reading, or shortly afterward, upon reflection (Singer & Spear, 2015). FTT is largely silent on whether different kinds of inferences are made online or offline shortly after reading (see Abadie & Camos, 2018). The phrase "gist inference" does not refer to any one particular class of inference in the literature such as thematic or instrumental inferences (Graesser et al., 1994). Research on different kinds of inferences has been conducted in the context of FTT (e.g., linear syllogistic inferences and pragmatic inferences; see Reyna & Kiernan, 1994). Prior work also indicates that there are other semantic and pragmatic interpretations of verbal and numerical information (e.g., metaphorical interpretations; representations of numerical magnitude, etc.) that contribute to gist representations. Thus, gist inferences go beyond surface form, connect propositions in texts, and, in so doing, capture core meaning but with less precision than surface-form sentences. Gist representations will require readers to go beyond surface form to extract meaning and inferences about the text's bottom line meaning. We use the term "gist inferences" to capture this notion of interpretive, essential (as in "essence," or less precise) meaning at the level of both sentences and connections among sentences.

FTT's gist–verbatim distinction is informed by classic psycholinguistic findings (e.g., Bransford & Franks, 1971; Clark & Clark, 1977; Kintsch, 1974). However, FTT makes different assumptions about verbatim and gist representations. "Bransford and Franks claimed that verbatim representations of the surface form were processed to extract gist representations of meaning and then the verbatim surface form was discarded" (Wilhelms, Fraenkel, & Reyna, 2018, p. 715). However, a good deal of research has contradicted such claims about semantic abstraction (see Reyna, 2012, for a review). FTT suggests that authors who wish to create expository texts that facilitate good reasoning and decision-making should

write in ways that help readers form appropriate inferences about the bottom line meaning of those texts.

To assess texts for their ability to produce gist inferences, we used the powerful discourse technology Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014) developed at the University of Memphis. In contrast to Flesch–Kincaid, Coh-Metrix is a "data rich" tool drawing upon hundreds of studies in different labs with meaningful texts and human research participants. For example, one variable, concreteness, is based on the MRC Psycholinguistic database with human ratings of more than 150,000 words (Coltheart, 1981; McNamara et al., 2014).

Coh-Metrix provides a multilevel framework (Graesser & McNamara, 2011) that measures texts across sentences, paragraphs, and the entire text. It computes over 100 linguistic variables at the level of descriptive statistics, including text "easability" (i.e., ease of reading) principal components, referential cohesion, latent semantic analysis (LSA), lexical diversity, connectives, situation model, syntactic complexity, syntactic pattern density, word information, and traditional measures of readability including FKGL. The "Coh" in Coh-Metrix stands for "Cohesion" and a key concept behind the development of Coh-Metrix is that the cohesion observable within a text predicts the coherence of cognitive representations within the reader (McNamara et al., 2014). Coh-Metrix has been applied to study a wide variety of issues including educational materials and learning environments (Dowell, Graesser, & Cai, 2016), the discourse characteristics of good undergraduate essays (McNamara, Crossley, & McCarthy, 2010), and textual dimensions of effective tutorial dialogues with Intelligent Tutoring Systems (Wolfe, Widmer, Torrese, & Dandignac, 2018). Coh-Metrix 3.0 is available to researchers via a web interface at http://cohmetrix.com/.

We developed a method to use Coh-Metrix to create a gist inference score (GIS). The GIS is designed to predict the extent to which people will make meaningful inferences from a text that can be used in subsequent decision making. Ours is a novel use of Coh-Metrix based on FTT that builds upon the work of McNamara et al. (2014) in ways that may not have been anticipated by the creators of Coh-Metrix. For example, although Coh-Metrix is designed to analyze text within a multilevel theoretical framework of comprehension (Graesser, McNamara, & Kulikowich, 2011), FTT is not necessarily aligned or opposed to the multilevel approach. A good deal of evidence suggests that gist memory traces are encoded independently from verbatim representations rather than derived from them (Reyna & Brainerd, 1995). However, some might suggest that a multilevel framework implies that the former arise from the later. This assumption of older psycholinguistic theories was thoroughly tested, and rejected, in research on FTT (e.g., Reyna & Kiernan, 1994, 1995). Nevertheless, resolving fundamental theoretical disputes is clearly beyond the scope of this article. Suffice to say that we are using Coh-Metrix as a tool to gain insights into text properties that facilitate decision making using a proximal measure designed to predict inference making based solely on observable text characteristics, without definitively deciding central issues in discourse psychology. We also recognize that there is more to comprehension than making gist inferences. For example, familiarity with the individual words one is reading clearly affects comprehension. At the same time, a large body of research suggests that making appropriate inferences is a key aspect of comprehension (Graesser et al., 1994; Reyna, Corbin, Weldon, & Brainerd, 2016).

There is a kinship between our concept of gist inference and the construction–integration concept of "macropropositions," which was a pivotal influence on FTT (Kintsch, 1988; Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983). According to Kintsch and van Dijk (1978, p. 372), "Macro-operators transform the propositions of a text base into a set of macropropositions that represent the gist of the text. They do so by deleting or generalizing all propositions that are either irrelevant or redundant and by constructing new inferred propositions." Thus, this work informs how hierarchies of gist representations are formed when text is encoded. However, macropropositions are not the same as gist inferences. FTT differs in core respects from prior models of narrative representation. In contrast to prior views, and grounded in extensive evidence, FTT suggests that gist representations are formed at the time of encoding independently from verbatim representations rather than through associations between verbatim and gist representations of the same stimulus (Reyna & Brainerd, 1995). The construction–integration model provides a detailed account of how macropropositions could be formed from micropropositions, which is roughly analogous to levels of gist in FTT, but that is distinct from the relationship between verbatim and gist representations (see Reyna & Brainerd, 1995; Reyna et al., 2016). In this work, the purpose of GISs is to provide a proximal assessment of texts for their potential to yield inferences that are likely to be useful in subsequent decision making.

Our GIS formula consists of seven variables with some positively weighted and others weighted negatively. These variables span three overarching areas: text cohesion, verb overlap, and word concreteness. Three of these Coh-Metrix variables are themselves composites of other variables, as a result of a principal components analysis of a large text corpus (McNamara et al., 2014). The other four are individual variables, which in some cases are also represented in the composite variables. That is to say, we argue that some individual variables warrant additional weight in predicting gist inferences. Each of these will be spelled out in greater detail below.

We argue that referential and deep cohesion help facilitate comprehension and gist representations. The Coh-Metrix composite variable **Referential Cohesion** (PCREFz) assesses the extent to which words overlap across sentences and the

entire text, forming meaningful threads. The classic psycholinguistic example is, "George got some beer out of the trunk. The beer was warm." in which the repetition of the word "beer" significantly improves processing (Haviland & Clark, 1974; McNamara et al., 2014). The principal components analysis of this composite variable positively weighs content word overlap, argument overlap, noun overlap, stem overlap, LSA given versus new, and LSA overlap; and negatively weighs type–token ratio, lexical diversity, dissimilarity of parts of speech between sentences, and dissimilarity of words between sentences (Graesser et al., 2011). Another positively weighted composite variable is **Deep Cohesion** (PCDCz), which is an index of the degree to which the text contains logical and intentional connectives that reflect causal and logical relationships within the text. Words such as "but, however, because, resulting in, and additionally" help readers make connections among different text passages. The principal components analysis of this composite variable positively weighs connectives, causal connectives, temporal connectives, logical connectives, causal cohesion, and intentional cohesion (Graesser et al., 2011).

The situation model is the representation derived from a text integrating the given text base with existing knowledge (Zwaan & Radvansky, 1998). A central dimension of forming a coherent situation model is the extent to which actions, as represented by verbs, are related to one another across a text. FTT suggests that abstract, rather than concrete verb overlap might help active readers construct gist situation models. Coh-Metrix uses two variables to assess the extent to which verbs (actions) are interconnected across a text. **Verb Overlap LSA** (SMCAUSlsa) uses LSA to assess the connection between each pair of verbs (technically the cosign of two vectors; McNamara et al., 2014) as one approach to measuring verb overlap. Because this approach is more abstract, it is weighted positively in the GIS formula. Coh-Metrix also uses **Verb Overlap WordNet** (SMCAUSwn) to assess verb overlap. This approach mainly counts identical verbs or those in the closely associated synonym set (McNamara et al., 2014) and because it is highly specific, FTT suggests that is more likely to lead readers to form detailed representations (closer to the verbatim level of representation), and thus it is negatively weighted in the GIS formula.

To illustrate, consider the following two texts of three sentences each. The first text is "Tumors spread to adjacent cells. Cancer metastasizes through blood vessels. Cancer travels through the lymph system." This text encourages readers to form a gist representation of the way cancer moves from one part of the body to the other. By way of contrast, "Tumors metastasize to adjacent cells. Cancer metastasizes through blood vessels. Cancer metastasizes through the lymph system." repeats the same verb "metastasize," which draws the reader's attention to the surface-level features of the text, thus encouraging a verbatim representation. Although readers can infer that the text is about the movement of cancer, many readers—especially those less familiar with the term "metastasize"—are less likely to form the appropriate gist inference.

FTT suggests that concrete, imaginable words are more likely to yield accessible verbatim rather than gist representations (Brainerd, Yang, Reyna, Howe, & Mills, 2008).

Three variables at the level of individual words indicate that verbatim representations are likely to be enhanced, and are thus scored negatively for gist inferences. **Word Concreteness** (PCCNCz) is a composite variable that measures the extent to which words that are concrete (rather than abstract) and evoke mental images. Example of words high on word concreteness are "table, chair, and street." The principal component analysis of this composite variable positively weighs meaningfulness, concreteness, and imagery, and negatively weighs age of acquisition (Graesser et al., 2011). **Imageability for Content Words** (WRDIMGc) is an index of how easy it is to construct a mental image for 4,825 words (McNamara et al., 2014). For example, the word "hammer" has been identified as high on imageability whereas the word "reason" is low. Because high imageability is likely to be associated with highly specific representations, including this variable in addition to the principal component is warranted. Finally, **Hypernymy for Nouns and Verbs** (WRDHYPnv) represents the specificity of a word within a hierarchy. Words with many hypernymy levels are generally more tangible whereas words with few hypernymy levels tend to be less so. For example, the words "Manx–Cat–Feline–Mammal–Vertebrates" can be arrayed in a hierarchy. Texts with many words such as "Manx" are less likely to help readers develop gist inferences than are texts with words with fewer levels of hierarchy, such as "mammal."

Each of the seven variables in the GIS formula is expressed on a different numeric scale so the first step was to put them on common footing by converting them to Z scores. Coh-Metrix reports some of the variables as Z scores already, and for others we created estimated Z scores from norms provided by McNamara et al. (2014) for Social Studies texts, 11th grade to adult. Table 1 provides the means and standard deviations used to create the estimated Z scores. Because the Coh-Metrix convention is that variables ending in the letter "Z" are presented in Z scores, for the variable names used in the GIS formula we have appended the letter "z" to the end of variable names to indicate that they are converted to estimated Z score units, by subtracting the estimated population mean from the variable and dividing by the estimated population standard deviation using the norms for Social Studies texts from 11th grade to adult from McNamara et al.'s (2014) Appendix B: Coh-Metrix Indices Norms.

Once the seven variables were all converted to Z scores, they were combined by finding the unweighted mean, with some variables being counted positively and others negatively, as outlined above (see Fig. 1). It should be noted that our

Table 1 Means and standard deviations used to create estimated Z scores for gist inference scores (variable names ending in Z are already expressed in Z-score units)

Coh-Metrix variable number	Coh-Metrix variable name	Description	Estimated mean	Estimated standard deviation	Estimated Z score variable name
16	PCCNCz	Word Concreteness Z Score			PCCNCz
18	PCREFz	Referential Cohesion Z Score	–	–	PCREFz
20	PCDCz	Deep Cohesion Z Score	–	–	PCDCz
64	SMCAUSlsa	LSA Verb Overlap	0.097	0.04	zSMCAUSlsa
65	SMCAUSwn	WordNet Verb Overlap	0.553	0.096	zSMCAUSwn
98	WRDIMGc	Word Imageability	410.346	24.994	zWRDIMGc
103	WRDHYPnv	Hypernymy Nouns & Verbs	1.843	0.26	zWRDHYPnv

claims are limited to expository texts. We are not prepared to make claims about GIS for narratives, poetry, allegory, or other kinds of texts.

To assess reliability and validity, GIS was applied to three text corpora. The basic approach was to find texts that are equated for a number of characteristics, and that were identified a priori as being relatively high or low in helping readers form gist representations. We then computed GIS scores for those texts and tested the hypothesis that texts identified as higher on gist will yield significantly higher GIS scores than those identified as more verbatim. This was followed by a behavioral experiment about knowledge and comprehension, to further test these hypotheses.

Study 1: Verbal data from a memory study

The GIS formula was developed using verbal data from Smith's (2017) unpublished Master's thesis. Smith replicated Schooler and Engstler-Schooler's (1990) study of verbal overshadowing, in which participants watched a video of a bank robbery and identified the robber from a lineup with some describing him verbally beforehand. Smith argued that the original instructions would prompt a highly verbatim description and added a condition in which participants were encouraged to produce a gist-like description following FTT. He found that the verbal overshadowing effect was limited to the original verbatim instructions. However, for our present purposes the important result is that the study produced texts that were confirmed by a reliable rubric as being gist or verbatim descriptions, blind to condition.

In the verbatim condition, Smith (2017, p. 11) used the instructions, "Please describe the appearance of the bank robber in as much detail as possible. It is important that you attempt to describe all of his different facial features. Please write down everything that you can think of regarding the bank robber's appearance. It is important that you try to describe him for the full 5 min." In the gist condition, Smith (2017, p. 11) used the instructions, "Tell me about the bank robber. What type of person is he? If he was a character in a story, what kind of character would he be? It is important that you write about him for a full 5 min, but talk about what is important without going into unnecessary detail." An example of a verbatim response is, "The bank robber was a white male with brown hair. He had dark eyes, seemed like dark brown eyes. He had mustache and he had a fair amount of hair on his hair. He looked straight at the person and never avoided eye contact. He asked for something and he gave the person an envelope of some sort. He was wearing a dark colored jacket. He didn't make much facial movements, but he looked over to the left a few times." (Smith, 2017, p. 12). An example of a gist description is

> The robber from the video was a quiet but assertive and shady figure. He wore dark clothing and had a strange mustache. He would definitely be the villain in the story, but he was not exactly cold and forceful, so it's possible he is battling some inner conflicts. Maybe he is rethinking his plan to steal the money, or maybe he is stealing the money to be able to pay for something dire, like hospital bills for a sick family member. He could also be a calculated bank robber, speaking quietly so as not to draw attention to himself. (Smith, 2017, p. 13).

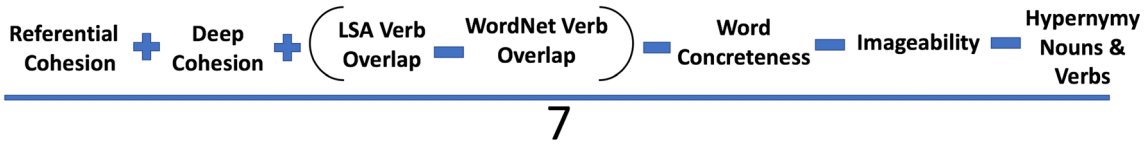

Fig. 1 Gist inference score formula (all variables converted to Z scores)

Results

The memory study yielded 66 texts total; 29 verbatim and 37 gist. The texts were closely matched for content in that they describe the same scenes. The texts averaged 104 words in length (SD = 46.5). Each text was subjected to a Coh-Metrix analysis with the GIS calculated for each text using the formula and procedure described above. The gist condition produced a mean GIS of 0.514 that was significantly higher than the one in the verbatim condition, which yielded a mean GIS of – 0.864, $F(1, 64) = 62.63$, $p < .0001$, $d = 1.38$. Table 2 presents the means and standard deviations for gist and verbatim condition texts and the raw constituent Coh-Metrix variables (i.e., variables such as WRDHYPnv before they were converted to estimates Z scores).

To provide further evidence that each GIS variable contributes to predicting differences among texts (despite the fact that some are composite variables and others individual variables) we conducted post-hoc hierarchical logistical regression analyses starting with the three composite variables word concreteness, referential cohesion, and deep cohesion predicting whether texts were gist or verbatim. Adding verb overlap LSA, we found that the effects likelihood ratio test for SMCAUSlsa was significant, $\chi^2(1) = 7.79$, $p = .005$. Adding verb overlap WordNet, we found that the effects likelihood ratio test for SMCAUSwn was significant, $\chi^2(1) = 13.32$, $p = .0003$. Adding imageability for content words, we found that the effects likelihood ratio test for WRDIMGc was significant, $\chi^2(1) = 20.80$, $p < .000$. Adding hypermymy for nouns and verbs, we found that the effects likelihood ratio test for WRDHYPnv was significant, $\chi^2(1) = 20.20$, $p = .0001$. With all seven GIS variables predicting whether texts were gist or verbatim, we found that $\chi^2(7) = 90.52$, $p < .0001$, with a log likelihood ratio of 45.26 for the entire model. The effects likely ratio test for each variable was significant at $p < .002$ for each variable, with Log Worth ranging from 10.60 for verb overlap WordNet, to 2.37 for word concreteness. These results further support the utility of the seven GIS variables.

Discussion

Each gist text had a higher GIS score than any of the verbatim texts, which is indicated by the very large effect size. The GIS formula indicates that the gist texts were close to the middle of the scale whereas the verbatim texts scored very low. Thus, Coh-Metrix and the GIS formula are sensitive to the verbatim descriptions of the bank robber as illustrated in the example above. It is difficult to make inferences about a physical description such as "white male with brown hair" but easier to do so when the text mentions characteristics such as he was a "quiet but assertive and shady figure." Having found preliminary evidence for the validity of the GIS formula our next task was to test it on a very different set of texts.

Study 2: Journal article methods versus discussion

We reasoned that in reading a peer-reviewed scientific psychology journal article, it should be easier to make gist inferences when reading the General Discussion than when reading the Method section. Ideally, the Method section provides enough detail about a study so that it could be replicated by others. The General Discussion, by way of contrast, emphasizes interpretation of results, often in theoretical terms. Thus, we hypothesized that text samples taken from the General Discussion should produce higher GIS scores than those from the Method section of the same articles by the same authors.

Articles were selected from PsycINFO, filtering on "Open Access," "Peer-Reviewed," and dates "2010 through 2018" using search terms "Psychology" and sorting by "Relevance" and selecting every 5th article. For 25 articles, we selected the first four paragraphs from the Method section of the first experiment, and selected the last four paragraphs from General Discussion section, making 50 texts total. In preparing the texts for analysis, we deleted headings and subheadings, and rounded down the longer text (typically the Discussion section) to a paragraph that contained the same number of words as the shorter text, to control for any effects of text length.

Table 2 Means and standard deviations for GIS and raw constituent Coh-Metrix variables for the gist and verbatim condition texts

Text	GIS	PCCNCz	PCREFz	PCDCz	SMCAUSlsa	SMCAUSwn	WRDIMGc	WRDHYPnv
Gist texts	0.514 (SD = 0.492)	– 0.155 (SD = 1.081)	0.524 (SD = 1.282)	0.278 (SD = 1.124)	0.128 (SD = 0.044)	0.590 (SD = 0.110)	394.680 (SD = 28.162)	1.426 (SD = 0.260)
Verbatim texts	– 0.864 (SD = 0.903)	2.310 (SD = 2.051)	0.337 (SD = 1.055)	– 2.089 (SD = X1.289)	0.157 (SD = 0.117)	0.679 (SD = 0.269)	465.445 (SD = 36.552)	1.834 (SD = 0.370)

PCCNCz is Word Concreteness Z score; PCREFz is Referential Cohesion Z score; PCDCz is Deep Cohesion Z score; SMCAUSlsa is LSA Verb Overlap; SMCAUSwn is WordNet Verb Overlap; WRDIMGc is Word Imageability; and WRDHYPnv is Hypernymy Nouns and Verbs

To illustrate, below is a paragraph of text from the Method section of Morelli, Bianchi, Baiocco, Pezzuti, and Chirumbolo (2017, p. 115).

> Alcohol consumption was assessed with the Alcohol Use Disorders Identification Test (AUDIT) (Babor et al. 2001). This scale was developed by the World Health Organization to evaluate alcohol-related problems and the possible risk for individual health. The scale assesses the amount and frequency of drinking, the alcohol addiction, and the problems related to alcohol abuse. In our scale, the participants had to rate eight items on a five-point scale, ranging from 0 (*never*) to 4 (*frequently* or *daily*). A total score for alcohol consumption based on these items was used in this study. In this sample, the scale reached a Cronbach alpha of 0.76.

By way of contrast, below is a paragraph from the Discussion section of the same article (Morelli et al., 2017, p. 119).

> As we hypothesized, gender differences were found for the three sexting subdimensions (i.e., receiving, sending, and posting sexts): Males were more likely to send, receive, and post sexts. These results may be explained by referring to the Italian cultural context. Previous studies found that Italian male adolescents are more likely than females to report that they find erotic materials enjoyable and arousing, and reported stronger positive expectancies about receiving sexts (Eurispes & Telefono Azzurro, 2012).

Results

The Method section text samples ($N = 25$) had a mean of 506 words, and Discussion text samples ($N = 25$) had a mean of 578 words. Each text sample was subjected to a Coh-Metrix analysis of GIS score as outlined above. Text samples from the General Discussion produced a mean GIS score of 0.443 that was significantly higher than the Method section text samples, with a mean GIS score of – 0.297, paired $t = 7.88$, $p < .0001$, $d = 1.75$. Table 3 presents the means and standard deviations for the Method and Discussion section texts and the raw constituent Coh-Metrix variables.

To further illustrate the utility of GIS scores, the Morelli et al. (2017) article presented in the sample paragraphs above produced GIS scores of – 0.795 for the Method section and + 0.802 for the Discussion. Figures 2 and 3 represent each of the seven elements of GIS for both of these texts.

In Figs. 2 and 3, the green bars correspond to elements that are weighted positively in the GIS formula, and the red bars correspond to elements that are weighted negatively in the formula, with the bars from left to right corresponding to Referential Cohesion, Deep Cohesion, Verb Overlap LSA, Verb Overlap WordNet, Word Concreteness, Imageability for Content Words, and Hypernymy for Nouns and Verbs. The Method section (Fig. 2) earned a very low GIS because all of the elements that are positively weighted are negative numbers, and all of the elements that are negatively weighted are positive numbers. This suggests that it should be difficult to make inferences about the Method section. The Discussion section (Fig. 3) is low on Referential Cohesion, which would discourage gist inferences. Nonetheless, it is high on Deep Cohesion, the difference between Verb Overlap LSA and Verb Overlap Word Net is high, and it is low on each of the variables associated with concreteness and imageability. Thus, the Discussion section earned a high GIS score, suggesting that readers would readily make gist inferences when reading this text.

Discussion

Discussion section texts had higher GIS scores than Method section texts, which is captured in the large effect size. Figures 2 and 3 provide a more detailed breakdown of the scores for each kind of text. Figures such as these can be used for several purposes, including revising texts to increase GIS by concentrating on specific elements, and conducting basic research on GIS and specific psycholinguistic variables (Dandignac & Wolfe, 2018). These results should not be construed as implying that the Discussion sections are written better than the Method sections. Rather, the Method section of a psychology journal article serves specific functions in which explicit detail is paramount and one should predict relatively fewer inferences, whereas the Discussion should yield more gist inferences, due to the role it plays in conveying the meaning of an investigation. Naturally, expert readers might "connect the dots" and infer gist from Method sections, but that gist is due more to the contribution of a reader than to the contribution of a text.

Study 3: News reports versus editorials

News reports (i.e., newspaper articles) are focused on facts. "Journalism students are taught about the five Ws: who, what, when, where and why" (Cole, 2008), with an emphasis on what happened. (Naturally, "why" encourages gist, but why receives less emphasis in news than in, for example, books about history.) Thus, we predicted that news reports would score significantly lower on GIS than editorials (op-ed pieces) on the same topics from the same news outlets. We predict that editorials will have higher gist scores than news reports because news articles emphasize facts whereas editorials provide a more coherent narrative even though they

Table 3 Means and standard deviations for GIS and raw constituent Coh-Metrix variables for Discussion section and Method section texts

Text	GIS	PCCNCz	PCREFz	PCDCz	SMCAUSlsa	SMCAUSwn	WRDIMGc	WRDHYPnv
Discussion section texts	0.45 (*SD* = 0.42)	– 1.12 (*SD* = 0.76)	– 0.29 (*SD* = 0.56)	0.36 (*SD* = 0.97)	0.10 (*SD* = 0.02)	0.46 (*SD* = 0.06)	377.54 (*SD* = 16.46)	1.97 (*SD* = 0.20)
Method section texts	– 0.297 (*SD* = 0.39)	0.19 (*SD* = 0.89)	– 0.52 (*SD* = 0.46)	– 0.06 (*SD* = 0.74)	0.10 (*SD* = 0.04)	0.56 (*SD* = 0.08)	414.57 (*SD* = 18.49)	2.13 (*SD* = 0.28)

PCCNCz is Word Concreteness Z score; PCREFz is Referential Cohesion Z score; PCDCz is Deep Cohesion Z score; SMCAUSlsa is LSA Verb Overlap; SMCAUSwn is WordNet Verb Overlap; WRDIMGc is Word Imageability; and WRDHYPnv is Hypernymy Nouns and Verbs

typically recount facts in a journalistic style to support an argument.

We compared 25 news reports to 25 editorials matched on topics, news source, and week published. The texts were collected from the open access news sites CNN, MSNBC, Fox News, Politico, and the Washington Post. Sample topics include Stephen Hawking, Stormy Daniels, Rex Tillerson, and gun control. We removed headlines, headings, images, and the like following the same procedure, as in Study 2.

To illustrate, consider sample paragraphs from two texts on the topic of Jared Kushner from the news outlet CNN. In a news article Merica (2018) wrote,

> CNN reported earlier on Tuesday that Kushner has been stripped of his access to the nation's top secrets after chief of staff John Kelly mandated changes to the security clearance system. Kushner had been working on a temporary clearance, but, under the new system, aides who previously had 'top secret' interim clearances saw their access downgrade to the less sensitive 'secret' designation.

On the same day in the same news outlet (CNN), Jen Psaki (2018) opined,

> The reliance on Jared Kushner as the primary negotiator for everything from Middle East peace to trade deals to the United States' relationships in Asia has been dying a slow death over the last year due to a combination of his lack of experience, lack of respect from world leaders and the actions of his boss and father-in-law, President Donald Trump. The official loss of his interim top security clearance should come as no surprise. A valid question the White House should have to answer is why it took so long.

Results

All of the news articles and editorials were between 350–550 words in length. Each text sample was subjected to a Coh-Metrix analysis of GIS score as outlined above. The news reports produced a mean GIS score of – 0.620 (*SD* = 0.35) that, as predicted, was significantly lower than the mean of – 0.252 (*SD* = 0.31) for editorials, paired $t = 3.92$, $p = .0006$. Table 4 presents the means and standard deviations for editorials and news report texts and the raw constituent Coh-Metrix variables.

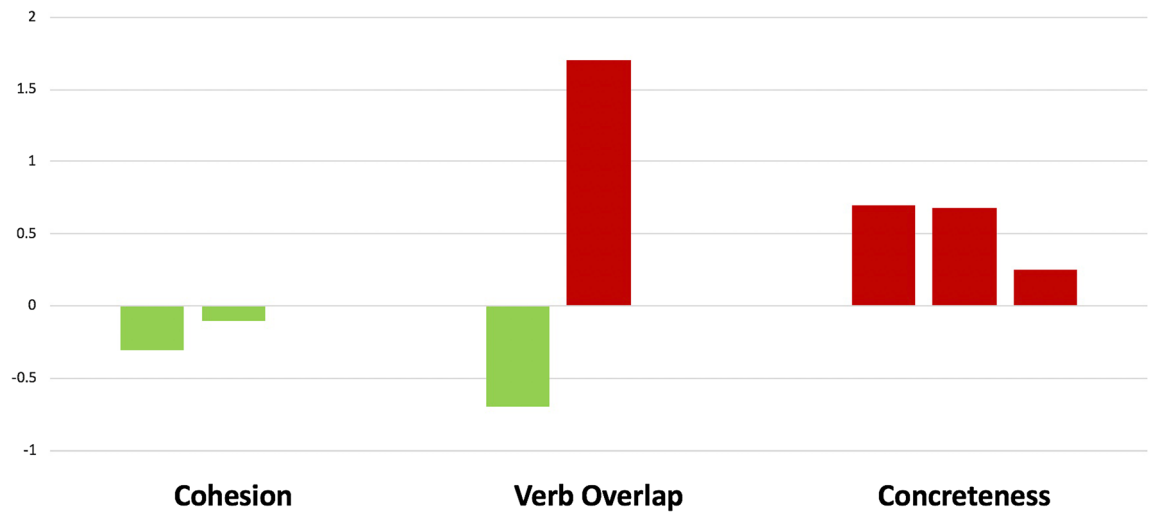

Fig. 2 GIS element Z scores for the Morelli et al. (2017) Method section

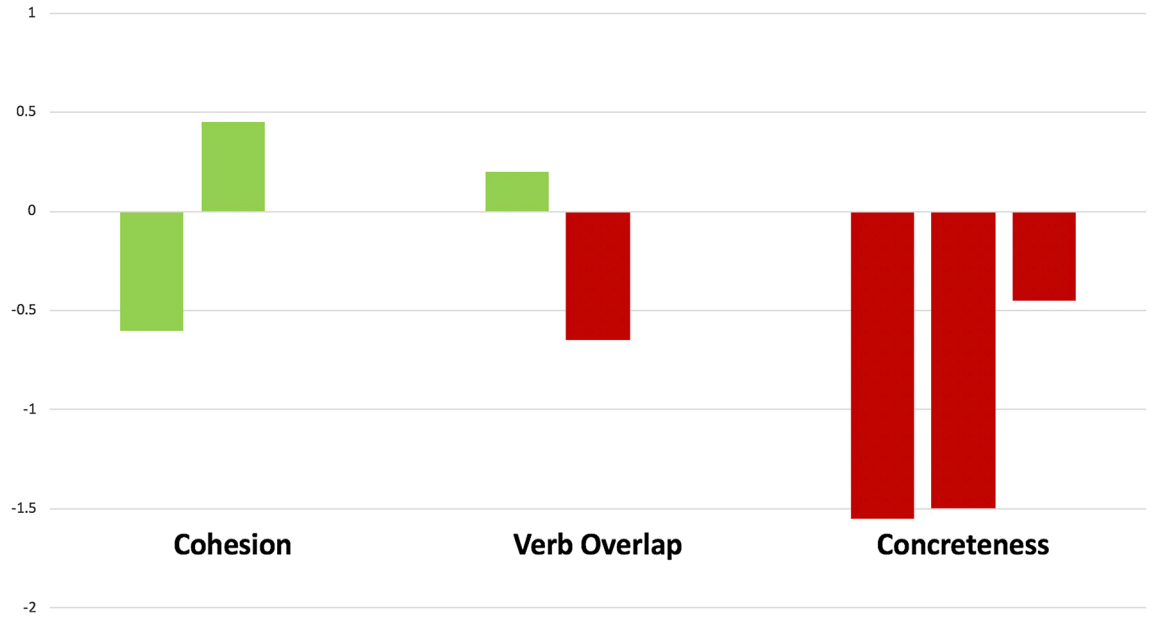

Fig. 3 GIS element Z scores for the Morelli et al. (2017) Discussion section

To exemplify GISs in greater detail, the news article by Merica (2018) on Jared Kushner quoted above produced GIS = – 0.750, and the paired editorial on the same topic by Psaki (2018), also quoted above, had GIS = – 0.074. Figures 4 and 5 represent each of the seven elements of GIS for both of these texts.

In Figs. 4 and 5, the green bars correspond to elements that are weighted positively in the GIS formula, and the red bars correspond to elements that are weighted negatively in the formula. The bars from left to right correspond to Referential Cohesion, Deep Cohesion, Verb Overlap LSA, Verb Overlap WordNet, Word Concreteness, Imageability for Content Words, and Hypernymy for Nouns and Verbs. As can be seen in Fig. 4, this news report has a positive score for Referential Cohesion, but all the rest of the GIS variables are the opposite of what would produce a positive GIS score. This suggests that it should be difficult to make many gist inferences about the topic at hand. By way of contrast, the editorial exemplified in Fig. 5 is very low on Referential Cohesion but is high on Deep Cohesion. The difference between the two measures of verb overlap contributes toward a higher GIS score, as do appropriately low scores for Word Concreteness and Imageability. Overall, these variables produce a score for this editorial in the middle of the GIS scale.

To assess the GIS formula empirically relative to alternatives, we conducted a test of whether GIS with and without each of the constituent variables was better able to discriminate between texts in all three studies relative to pooled standard deviations. Table 5 presents the Study 1 mean difference between the gist condition and verbatim condition texts divided by their pooled standard deviations; the Study 2 mean difference between the Discussion and Method sections texts divided by their pooled standard deviations; and the Study 3 mean difference between editorial and news articles divided by their pooled standard deviations. Here we compare the full seven-variable GIS formula to versions removing each of the seven variables, a version without both verb overlap variables, and Graesser's Formality measure (Dowell et al., 2016; Graesser et al., 2014).

It can be seen that GIS reveals large predicted differences between texts, ranging from 1.12 to 1.96 standard deviations. In some studies, a version of the GIS formula without one of

Table 4 Means and standard deviations for GIS and raw constituent Coh-Metrix variables for editorials and news report texts

Text	GIS	PCCNCz	PCREFz	PCDCz	SMCAUSlsa	SMCAUSwn	WRDIMGc	WRDHYPnv
Editorial texts	– 0.252 (SD = 0.305)	0.116 (SD = 0.714)	– 1.202 (SD = 0.686)	0.177 (SD = 0.757)	– 0.574 (SD = 0.849)	– 0.424 (SD = 0.973)	0.124 (SD = 0.734)	– 0.320 (*SD* = 0.781)
News report texts	– 0.620 (SD = 0.351)	0.412 (SD = 0.714)	– 1.099 (SD = 0.740)	– 0.091 (SD = 0.635)	– 1.068 (SD = 0.686)	0.195 (SD = 0.806)	0.416 (SD = 0.709)	– 0.487 (*SD* = 0.655)

PCCNCz is Word Concreteness Z score; PCREFz is Referential Cohesion Z score; PCDCz is Deep Cohesion Z score; SMCAUSlsa is LSA Verb Overlap; SMCAUSwn is WordNet Verb Overlap; WRDIMGc is Word Imageability; and WRDHYPnv is Hypernymy Nouns and Verbs

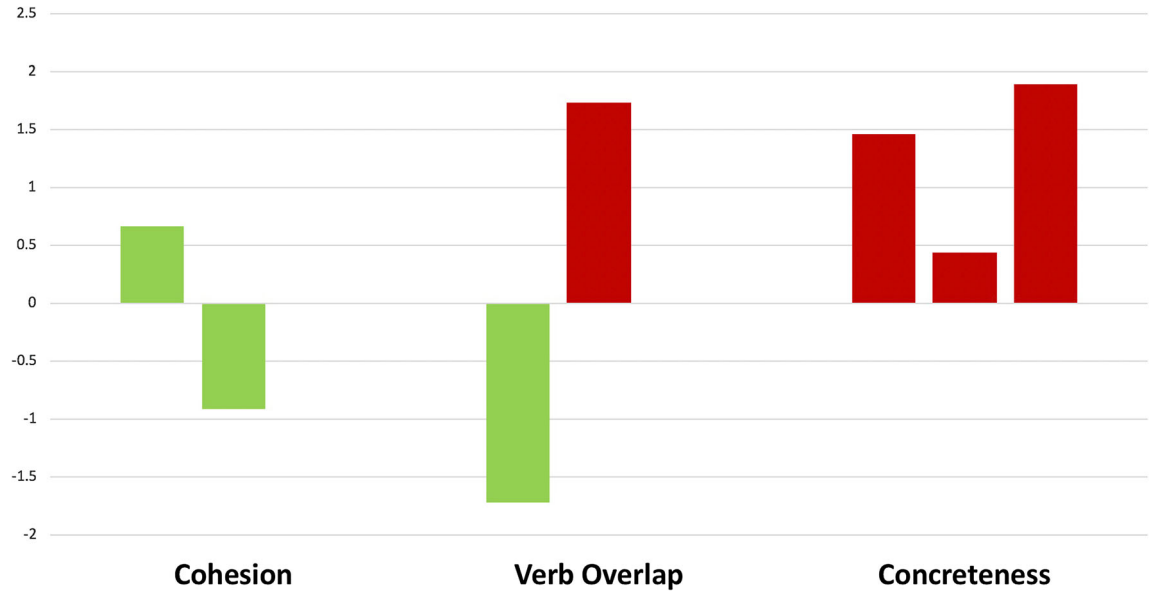

Fig. 4 GIS element Z scores for the news report by Merica (2018) on Jared Kushner

the constituent variables produced even larger differences. However, no alternative version consistently discriminated better between texts than GIS did across all three studies.

Discussion

News reports scored significantly lower on GIS than editorials, suggesting that it would be more difficult to make gist inferences after reading a news article than an op-ed piece. Note that our claim is not that people will completely fail to form gist representations or make gist inferences from reading a news article. Indeed, FTT (Reyna, 2008) is a dual-process theory suggesting that people continually form gist representations in parallel with encoding verbatim details. Rather, those inferences should be fewer and less richly interconnected than for comparable texts with high GIS scores. Thus, in the case of the news article about Jared Kushner, readers with low domain knowledge may only infer that Kushner was in trouble. Of course, readers with high knowledge of current events are more likely to integrate the Merica (2018) article with their preexisting knowledge and beliefs about diplomacy, the Trump administration, Jared Kushner, and so forth, but an article

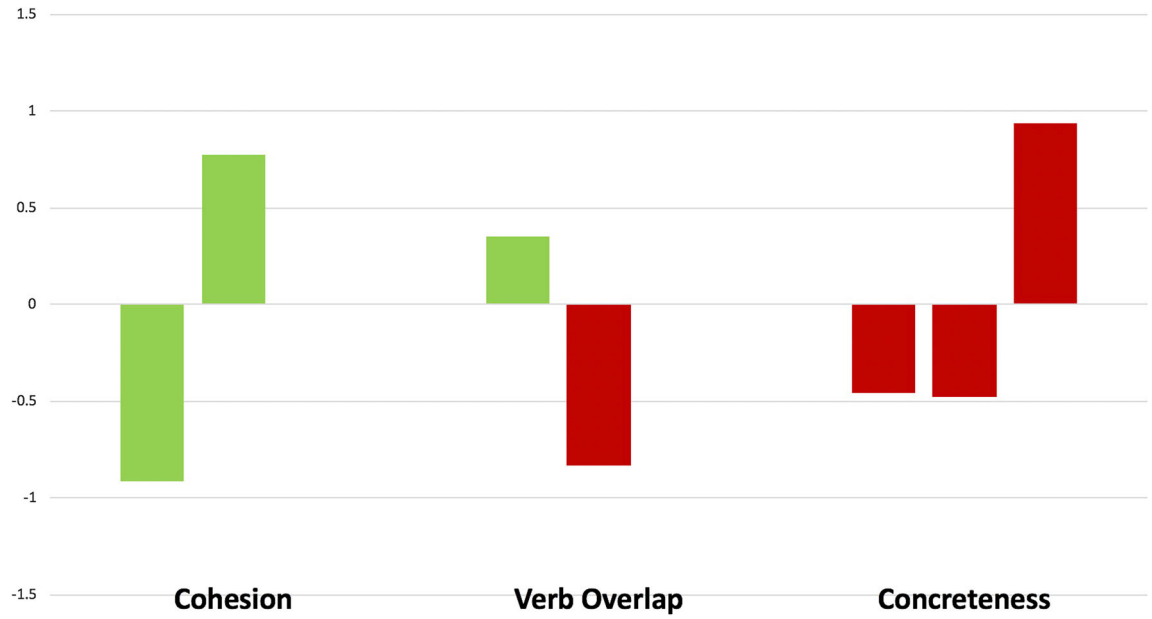

Fig. 5 GIS element Z scores for the editorial by Psaki (2018) on Jared Kushner

Table 5 Mean differences between texts over pooled standard deviations of gist inference scores (GISs) and alternative formulae for the Study 1, Study 2, and Study 3 texts

GIS and alternative configurations	Description of removed variables	Study 1 mean Gist: verbatim condition GIS over pooled standard deviation	Study 2 mean discussion: methods GIS over pooled standard deviation	Study 3 mean editorial: news GIS over pooled standard deviation
GIS	Seven-Variable GIS Formula	1.963	1.825	1.120
GIS without PCCNCz	Word Concreteness	1.941	1.755	1.226
GIS without PCREFz	Referential Cohesion	2.005	1.774	1.096
GIS without PCDCz	Deep Cohesion	1.588	1.832	1.226
GIS without SMCAUSlsa	LSA Verb Overlap	2.752	1.913	0.959
GIS without SMCAUSwn	WordNet Verb Overlap	1.498	1.581	0.854
GIS without WRDIMGc	Word Imageability	1.683	1.577	1.222
GIS without WRDHYPnv	Hypernymy Nouns & Verbs	2.008	1.936	1.037
GIS without SMCAUSlsa and SMCAUSwn	Verb Overlap LSA and WordNet	2.147	0.741	0.702
Formality	Five-Variable	1.777	1.437	0.536

Formulae: GIS = (0 – PCCNCz + PCREFz + PCDCz + (SMCAUSlsa – SMCAUSwn) – WRDIMGc – WRDHYPnv)/7. Formality = (PCREFz + PCDCz – PCNARz – PCSYNz – PCCNCz)/5, where PCNARz is narrativity and PCSYNz is syntactic simplicity

with a low GIS score will make those cognitive tasks more difficult than a comparable text with a higher score.

Editorials scored higher than new reports, but still produced low GIS scores in absolute terms. Figure 5 suggests that the author of this editorial could have facilitated the reader's ability to form gist inferences by increasing Referential Cohesion and by decreasing Hypernymy for Nouns and Verbs. The approach developed by Dandignac and Wolfe (2018) suggests that Referential Cohesion can be increased by making more explicit referential links among sentences. For example, the Psaki (2018) editorial says, "Jared Kushner's security clearance downgrade is not the end of the story, unless you are talking about his diplomatic career. That is over." Here Referential Cohesion would be increased by replacing "That is over" with "Kushner's diplomatic career is over," though changes of this kind might increase Verb Overlap WordNet. Hypernymy could be changed by replacing very specific words such as "downgrade" with less specific terms, such as "reduction" or "lowering." Of course, authors have a number of competing demands and might be satisfied with texts that produce few gist inferences. Nonetheless, GIS, and the kind of information presented in Fig. 5, are useful in helping authors make decisions about how to increase the likelihood that readers will make appropriate inferences in the context of other goals and constraints.

No alternative version of the GIS formula consistently discriminated better among texts than did GIS. Thus, the seven variable GIS formula appears to be superior to the tested alternatives. The eyewitness descriptions used in Study 1 were selected as our starting place because we were highly confident that they adequately capture the gist-verbatim distinction. However, these brief 104 word descriptions made by research participants are less representative of published texts than those used in Studies 2 and 3. Thus, generalizations about specific variables made on the basis of Study 1 alone should be made with caution—but they are buttressed by results of Studies 2 and 3.

There is a kinship between GIS and Graesser's measure of Formality (Dowell et al., 2016; Graesser et al., 2014). GIS has three variables in common with Formality: Referential Cohesion, Deep Cohesion, and Word Concreteness. Thus, it is not surprising that with the text analyzed here, there was a positive correlation between Formality and GIS. However, GIS better discriminated among these texts than Formality (see Table 5). Moreover, it is not clear that in Study 1 verbal descriptions of a bank robber actually differed with respect to the construct "formality," or that researchers use more formal language in the Discussion section than the Method (Study 2), or that editorials are actually more formal than news articles (Study 3). In short, it appears that, despite the overlap, Formality and Gist Inference are two distinct constructs.

Study 4: Predicting knowledge and comprehension from gist inference scores

Having reliably made predictions about texts in three different studies, we turn to empirically testing predictions about the consequences of people reading texts in which GIS level was manipulated experimentally. Although there is clearly more to comprehending and learning from texts than GIS, we reasoned that people should better comprehend a text that has a

high GIS, and score better on a knowledge test after reading it, compared with a text that covers comparable content but has a low GIS.

We had previously developed and tested the BRCA Gist Intelligent Tutoring System to help women understand and make decisions about genetic testing for breast cancer risk (Wolfe et al., 2015; Wolfe et al., 2016). Because we had reliable instruments to assess knowledge and comprehension, we decided to use these didactic scripts as the basis of texts to manipulate. The basic approach was to use the *BRCA Gist* script as the base text and then create revised versions that scored one standard deviation higher on GIS while remaining less than one standard deviation different on any other Coh-Metrix variable. We then randomly assigned people to read different texts followed by tests of comprehension and knowledge. We predicted that people would score significantly higher on the knowledge and comprehension measures when randomly assigned to the higher GIS version.

Method

Materials We edited the script for the BRCA Gist Intelligent Tutoring System (Wolfe et al., 2015) first by removing all images and tutorial dialogues and all references to images, dialogues, the animated agents, or the process of interacting with BRCA Gist. This left one long didactic text that we divided into two didactic texts to build in replication with different materials. Particularly because the removed images and dialogues have a demonstrable effect on comprehension and knowledge (Wolfe et al., 2016), these base texts were not ideal gist communications from the standpoint of FTT. Then, we made a copy of each didactic text and edited them to increase each of the seven variables that make up the GIS. We focused on changing content words, verbs, connectives, and syntactic features, developing a method presented by Dandignac and Wolfe (2018). For example, we judiciously added words such as "however," "moreover," and "nonetheless" to increase construct deep cohesion. Terms that some readers may have difficulty interpreting, such as "they," were replaced by their referents, for example "malignant tumors" to increase referential cohesion. Conversely, we reduced the number of times in which the same *verb* was repeated, to decrease WordNet Verb Overlap. We also replaced more concrete nouns, as in "may be a threat to organs and tissues," with less concrete nouns, as in "may be a threat to life," to decrease word concreteness. To ensure that the texts differed by one standard deviation we made some changes to the original to decrease GIS, but the majority of changes were made to increase GIS. The Appendix presents both the high-GIS and low-GIS versions of one of the two Understanding Breast Cancer text pairs. For the first text pair, the high- and low-GIS versions differ by 1.051 GIS units, and for the second text pair, the high- and low-GIS versions differ by 1.006 GIS units. Table 6 presents the GIS scores and constituent Coh-Metrix variables for each of the four texts.

We reasoned that versions of texts that facilitate inferences about their bottom-line meaning should be better understand by readers than texts that emphasize verbatim specifics. To test this hypothesis, we used two dependent measures that are represented in the literature. Declarative Knowledge of Breast Cancer and Genetic Risk is a 52-item, four-alternative multiple-choice assessment on breast cancer, genetic risk, and genetic testing (Wolfe et al., 2015; Wolfe et al., 2016), with items such as "Breast Cancer usually forms in which parts of the breast? (answer: ducts and lobules)." Cronbach's alpha for the instrument is .88 (Wolfe et al., 2016). Gist Comprehension of Genetic Breast Cancer Risk (Wolfe et al., 2015) is a 40-item, 1–7 Likert-scale instrument measuring gist comprehension of important information about breast cancer and genetic testing. 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 information conveyed in the texts presented to participants (see the Appendix)—the essential bottom-line meaning. People can strongly endorse statements such as these without remembering the precise verbatim details (Wolfe et al., 2016). The response format permits degrees of agreement: "Although these items are presented with a Likert scale, they are not opinion-based—all of the items have independently verifiable correct answers (the correct gist meaning). Thus, a participant who shows lower agreement with the statement 'The greatest danger of dying from breast cancer is when a tumor grows larger in the location where it started' can be said to possess a stronger gist understanding about breast cancer risk than one who shows higher agreement because this is not the correct gist meaning" (Widmer et al., 2015, p. 637). Cronbach's alpha for Gist Comprehension is .85 (Wolfe et al., 2016). Wolfe et al. (2015) found that participants randomly assigned to the BRCA Gist Intelligent Tutoring System scored significantly higher than control group participants. These findings were replicated by Widmer et al. in a study with web and community participants, suggesting validity. Before the beginning of the experiment, we determined which facts were covered in each pair of texts. This approach also permitted us to use items from materials not covered in each text pair as an index to preexisting knowledge.

Participants and procedure The participants were 169 native English-speaking undergraduates at Miami University, who participated for credit in a psychology course. Participants were randomly assigned to one of four groups: Text 1 High GIS, Text 1 Low GIS, Text 2 High GIS, and Text 2 Low GIS. Participants were run individually or in small groups at separate tables and work stations. Both the texts and dependent measures were presented on the Qualtrics platform using an ordinary web browser. Working at their own pace, participants

Table 6 GIS and raw constituent Coh-Metrix variables for each understanding breast cancer text

Text	GIS	PCCNCz	PCREFz	PCDCz	SMCAUSlsa	SMCAUSwn	WRDIMGc	WRDHYPnv
High 1	0.730	– 0.321	1.426	1.654	0.175	0.53	405.976	2.014
Low 1	– 0.321	0.756	0.600	0.675	0.121	0.629	427.630	2.333
High 2	0.702	– 0.646	1.509	1.325	0.166	0.554	404.825	1.974
Low 2	0.304	0.483	0.707	0.353	0.116	0.654	427.07	2.222

PCCNCz is Word Concreteness Z Score; PCREFz is Referential Cohesion Z Score; PCDCz is Deep Cohesion Z Score; SMCAUSlsa is LSA Verb Overlap; SMCAUSwn is WordNet Verb Overlap; WRDIMGc is Word Imageability; and WRDHYPnv is Hypernymy Nouns and Verbs

first read the randomly assigned text, then completed the Gist Comprehension task, and finally completed the declarative knowledge test. Upon completion, participants were thanked and debriefed.

Results

For Text Pair 1 (see the Appendix), participants who received the high-GIS text scored significantly higher on Declarative Knowledge than those who received the low-GIS version of the same content. For high-GIS participants, the mean proportion correct was .795 (SD = .020) and for low-GIS participants, the mean was .727 (SD = .019), $F(1, 80) = 6.36$, p = .014. Controlling for existing knowledge by entering items not covered in the text, pre-existing knowledge significantly predicted Declarative Knowledge (t = 5.20, $p < .0001$), but the high-GIS group was significantly higher on Declarative Knowledge (i.e., controlling for existing knowledge, p = .012). For Text Pair 2, participants who received the high-GIS text scored slightly but not significantly higher on Declarative Knowledge than those who received the low-GIS version. For high-GIS participants, the mean proportion correct was .799 (SD = .022), and for low-GIS participants the mean was .767 (SD = .020), $F(1, 88) = 1.21$, p = .27. Controlling for existing knowledge by entering items not covered in the text, again, preexisting knowledge significantly predicted Declarative Knowledge (t = 7.11, $p < .0001$), but the high-GIS group was not significantly higher on Declarative Knowledge when controlling for existing knowledge, p = .29. Thus, we found that high GIS led to higher declarative knowledge scores on one pair of texts.

For Text Pair 1, the high-GIS text yielded a slightly but not significantly higher Gist Comprehension, with high-GIS texts producing a mean Gist Comprehension score of 5.24 (SD = 0.075), and low-GIS texts yielding a mean of 5.085 (SD = 0.073), $F(1, 78) = 2.25$, $p = .14$. When controlling for existing knowledge by entering items not covered in the text, preexisting knowledge significantly predicted Gist Comprehension ($t = 4.09$, $p < .0001$), and the high-GIS group was not significantly higher on Gist Comprehension, controlling for existing knowledge, $p = .16$. For Text Pair 2 there was a significant effect, with the high-GIS text yielding higher Gist Comprehension: High-GIS texts produced a mean Gist Comprehension score of 5.37(SD = 0.074), and low-GIS texts produced a mean of 5.153 (SD = .073), $F(1, 87) = 4.44$, p = .038. When cntrolling for existing knowledge by entering items not covered in the text, preexisting knowledge significantly predicted Gist Comprehension (t = 6.67, $p < .0001$), and the high-GIS group was also significantly higher on Gist Comprehension when controlling for existing knowledge, p = .023. As in the case of Declarative Knowledge, we found that high GIS led to higher Gist Comprehension scores on one pair of texts.

Discussion

Although the distinction between text pairs was arbitrary, the high-GIS version of each pair produced a statistically significant difference on one outcome measure. We characterize this as converging evidence that GIS is measuring characteristics of text associated with gist inferences—the reader's ability to make inferences about the bottom-line meaning of texts.

General discussion

Coh-Metrix (McNamara, et al., 2014) is a powerful data-rich discourse technology that is easy to use and available to the research community free of charge. We developed the formula for Gist Inference Scores using Coh-Metrix following fuzzy-trace theory (Reyna, 2008; Reyna et al., 2016) to predict the extent to which people will make meaningful inferences from a text. In three text analysis studies with three very different kinds of text, our predictions were confirmed with large effect sizes. A behavioral study with two pairs of texts that were otherwise closely matched suggested that people randomly assigned to high-GIS versions scored higher on declarative knowledge and gist comprehension. Moreover, GIS was consistently better at discriminating among texts than alternative measures (see Table 5). Taken collectively, this provides evidence of reliability and validity.

On one level, it appears that our GIS approach makes predictions that contradict the construction-integration model (Kintsch, 1988; Kintch & van Dijk, 1978; van Dijk &

Kintsch, 1983) in that GIS scores will be higher for texts with more abstract terms whereas in the construction–integration model, concrete words might be predicted to produce greater activation of concepts in a network representation of a text, thus facilitating integration. However, prior work has separated associative activation from meaningful gist (Brainerd et al., 2008). Furthermore, as previously noted, there are many different kinds of inferences that readers must make, either online, or offline after reading (Graesser et al., 1994). To illustrate, inferences of the sort described by van Dijk and Kintsch (1983, p. 191) such as, "the sequence ('X goes to the airport', 'X checks in', and 'X waits for boarding') entails the macroproposition 'X is taking a plane', given the appropriate world knowledge," and this might be argued to be an inference facilitated by concrete words associated with airplanes. However, evidence shows that these pragmatic (world-knowledge-based) inferences follow the same representational principles as other kinds of inferences, and associative explanations have been ruled out (e.g., Reyna & Kiernan, 1994). For example, "X goes to airport" could be take the A subway train, take a Yellow Cab, order an Uber, get a ride from a friend in a red Corvette, and so forth. Thus, pragmatic inferences about the bottom-line meaning used in decision making should be captured by GIS, although chains of concrete associations may not be captured. A deeper investigation of the relationship between GIS and other extant theories of discourse comprehension should be a subject for future investigations.

There are a number of shortcomings and limitations to our approach. First, GIS does not account for reader characteristics. Attributes such as reading ability, domain knowledge, and readers' goals and motivation all have demonstrable effects of comprehension and inference making. Yet, by focusing solely on observable text characteristics, GIS cannot include these individual-level reader traits. However, Gist Inference Scores may be used to develop research materials (stimuli) for empirical studies of psychological processes including memory, comprehension, and subsequent decision-making. (Dandignac & Wolfe, 2018, developed an approach to writing and revising texts to manipulate GIS while keeping other text variables relatively constant.)

A second shortcoming is that GIS does not capture memorable phrases and other aspects of good writing that help readers form gist representations. For example, to discourage base rate neglect in diagnosis, medical students are sometimes admonished "When you hear hoof beats, think horses not zebras" (Wolfe, 1995). Concrete examples and clever phrases sometimes help people understand the bottom line meaning (gist) of concepts, such as the base rate fallacy (though novel metaphors tend to support more memorable verbatim representations; Reyna & Kiernan, 1995). These literary devices might not be captured with our GISs. Finally, it is premature to use GIS for absolute magnitude judgments. If we believed that all of our estimated Z scores were perfectly representative and that each variable should truly be equally weighted, then a GIS = 0 would correspond to the gist inferences yielded by an average text. We are not ready to make this strong claim at this time, as further research is necessary. We suggest that GIS is better suited for relative, comparative judgments, such as different texts in the same general domain or different versions of the same text.

We conclude that GIS as a proximal index represents a valuable supplement to (or, for some purposes, an alternative to) traditional data-lean readability measures. For practical purposes, such as communicating complex information to medical patients (Reyna, 2008) and in empirical research on discourse, comprehension, and other cognitive processes, the preliminary evidence suggests that GIS is a viable approach to predicting the likelihood that readers will make meaningful inferences from expository text.

Acknowledgments We thank the following Miami University undergraduates for their capable assistance with text analysis: Sam Almond, Navdeep Bais, Luejack Baker, Aiyana Green, Pamela Ianiro, Vasanthi Kalindi, Tatum Moleski, Jordan Oldham, Michael Suponic, Chase Tirey, Eleni Vidalis, Katie Weidner, and Vince Werthmann. None of the data or materials for the experiments reported here are available, and none of the experiments was preregistered.

Appendix: High-GIS and low-GIS texts about breast cancer used in Study 4

High-GIS version

Genetic testing for breast cancer is important because it is the second most common form of cancer among women. Additionally, it is also the kind that causes the most deaths. It is a serious disease, but with treatment breast cancer is not fatal for most people because they are successfully treated. The risk of getting it increases as you get older, and most are over sixty years old when they are first diagnosed.

Some mutated genes including BRCA 1 and BRCA 2 increase your risk of cancer. Gene mutations have been found in many families with a history of breast cancer and some in these families have also had ovarian cancer. Overall, inherited genetic mutations make up only about 5% to 10% of all breast cancer.

Typically, the BRCA 1 and BRCA 2 genes suppress excess cell growth. However, when these genes have a harmful mutations they cannot function properly and cells can grow uncontrollably, leading to tumors. Having mutations does not automatically guarantee that a person will get breast cancer. However, the risk of someone with a mutation of developing it becomes five times higher. Therefore, between 55% and 80% of women with a BRCA mutation will develop the disease.

Tumors in the breast can become benign or malignant. Benign ones only grow in one place and are not as harmful

as malignant ones. Unlike benign tumors, malignant ones are cancerous and they may become a threat to life. They can often be removed, but sometimes grow back. As a result, they spread to other parts of the body and damage organs and tissues.

A negative test result means that your risk for developing cancer is no greater than the general population, but it still is no guarantee that malignancy will not develop. Because of this, it is sometimes impossible to determine if the test is positive or negative and you are left with no more information than before. This result is called ambiguous and it happens about 10% of the time.

Genetic testing affects your family. Think about who in your family might want to know your results, such as your children, because one of the decisions you'll need to make is who you would tell, and what it might mean for their lives.

The price of testing varies and is often not covered by insurance. Before testing, ask your health care provider for more information on genetic testing, privacy issues, and insurance coverage. Some advantages of having a test are that it may help you to make lifestyle choices, clarify your cancer risk, allow you to consider surgery, and give your family useful information. Overall, a positive result may help explain why you or other family members have developed cancer.

There are three ways that cancer spreads. First, it can invade the surrounding normal tissue. Second, it can invade the lymphatic system and travels through the lymph vessels to lymph nodes throughout the body. Lastly, it can invade the veins and capillaries and travels through the blood and lymphatic system.

The metastatic tumor is the same type of cancer as the original tumor. For example, if breast cancer spreads to the bones, the cancer cells in the bones are actually breast cancer cells. In this case, the disease is metastatic, or "distant," breast cancer, not bone cancer. Metastasis becomes deadly when it spreads to vital organs, and when other tissues of the body form tumors.

There are three common ways doctors think about risk. Those are absolute risk, relative risk, and 5-year risk. It becomes important to know the difference between these risks. The first kind is absolute risk which means the overall chances of a person getting breast cancer. When someone talks about absolute risk they mean the risk that a person will develop cancer some time during her lifetime.

The absolute risk for a baby girl is about 12% sometime during her lifetime. As she gets older her absolute risk will decrease but her risk of getting cancer in the next 5 years will increase. Different from absolute risk is relative risk. The relative risk of breast cancer and genetic mutations means the change in risk due to risk factors. To further explain relative risk, imagine two scenarios. Scenario A has 2 out of 200 people with that cancer, or a 1% chance of getting it and Scenario B has 3 out of 200 people with that cancer, or a 1.5% chance of getting it. As a result, there is an increase from 1% to 1.5% in terms of absolute risk from Scenario A to Scenario B but a 50% increase in relative risk. As a result, most with a risk factor for breast cancer mutations do not have mutations themselves. For example, having a relative with breast cancer or being of Ashkenazi Jewish descent are risk factors for mutations, but most people with afflicted relatives or in a higher risk ethnic group do not possess mutations themselves.

When we take the statistical approach the base rate becomes the starting point. While we need to consider other information to make good decisions, the base rate keeps things in perspective.

Another mistake that people make is believing the risk of having two factors is higher than having either one. For example, imagine a person has risk factors for a breast cancer mutation. Some people may think their risk for developing breast cancer and having a mutation is higher than just having a mutation. However, this is wrong. If you have two or more close relatives who are diagnosed with ovarian cancer, regardless of what age they were diagnosed at, your risk is increased. This is because ovarian cancer has been shown to be highly linked to breast cancer, unlike other cancers.

Similarly, your risk is also increased if you have a close relative who was diagnosed at any age with both breast and ovarian cancer. Lastly, your risk is also increased if you have a male relative who has been diagnosed with breast cancer. It is so rare that having a male relative with it is considered high risk of having mutations.

There are no guidelines for recommending when someone should be tested for BRCA mutations. However, there is agreement on what may increase the likelihood of having one. Identifying families with a history of breast or ovarian cancer is a first step to gathering information about a person's risk. Because of this, working with a genetic counselor helps in detecting and explaining risks and will also provide information about genetic tests.

However, there are other costs of having genetic testing. First, if you are at low risk for a BRCA mutation, there may be cheaper options. Second, your test results will become a part of your medical records. There are laws prohibiting employers and health insurance companies from discriminating against you. However, they do not cover life insurance, disability insurance, long-term health insurance, or members of the military. As a result, many people are also concerned about the possibility of discrimination.

People tested for BRCA mutations usually possess a family history of breast cancer and ovarian cancer. In these cases, the best approach is to have the person with cancer tested for mutations first. Imagine that several people in the family have breast or ovarian cancer but they all test negative for BRCA mutations. If this family has a history of either cancer but no mutations have been found, a negative test is not informative.

Consequently, they can't tell whether the mutation was a false negative or a true negative. The family may also have a rare or unknown mutation.

As a result, a mutation in a gene other than BRCA 1 or BRCA 2 could increase cancer risk but may not be detectable.

One option you have after you find a mutation is active surveillance. Here, that means that you regularly screen yourself for cancer to detect it earlier. The goal is to find it early, when it is most treatable. However, this does not change the risk of developing cancer; though it decreases the risk of dying from it.

A second option after finding a mutation is surgery that removes the at-risk breast tissue. By removing healthy tissue, a person's risk of developing breast or ovarian cancer decreases. However, this does not guarantee that it will not develop.

Lastly, you can choose chemo prevention. Here, a person would take a drug such as Tamoxifen, which has been shown to reduce the risk of developing breast cancer by about 50% for at-risk women. However, it is unknown how effective Tamoxifen really is for prevention with women not at-risk.

Since genetic information is considered health information, it gets covered by HIPAA. Within HIPAA, the Privacy Rule requires that health care providers and others protect that information. As a result, it sets boundaries on the use and release of health records, and it empowers individuals to control disclosures of them. Genetic discrimination occurs when people are treated differently by insurance companies or employers because they have a mutation that increases their risk of a disease. When this happens, GINA protects U.S. citizens against discrimination due to their genetic information in relation to health insurance and employment. However, there are several items that this does not cover. For example, life, disability, and long-term care insurance are excluded.

Low-GIS version

In this study you will be reading and learning about genetic testing for breast cancer risk. Among women, breast cancer is the second most common form of cancer after skin cancer. Breast cancer is also the kind of cancer that causes the most deaths after lung cancer. Breast cancer is a serious disease, but it is important to remember that with treatment breast cancer is not fatal for most women. Most women with breast cancer are successfully treated. The risk of getting breast cancer increases as you get older. Most women are over sixty years old when they are first diagnosed with breast cancer.

Some altered genes including BRCA 1 gene and BRCA 2 gene can increase your risk of cancer. Gene alterations have been found in many families with a history of breast cancer. Some women in these families have also had ovarian cancer. People inheriting genetic mutations from family make up about 5% to 10% of all breast cancer cases.

Under normal circumstances the BRCA 1 gene and BRCA 2 gene stop excess breast cell growth. When the BRCA 1 gene and BRCA 2 gene have a harmful mutation they cannot function properly and breast cells can grow uncontrollably, leading to tumors. Having a BRCA gene mutation does not automatically guarantee that a person will get breast cancer. The risk of women with a BRCA gene mutation of developing breast cancer is about five times that of the rest of the people, and between fifty-five and eighty percent of women with a BRCA gene mutation will develop breast cancer.

Tumors in the breast can be benign or malignant. Benign tumors can only grow in one place and are not as harmful as malignant tumors. Malignant tumors are cancerous tumors and they may be a threat to life. Tumors often can be surgically removed, but sometimes grow back. Malignant tumors can disperse to other parts of the human body and invade and damage nearby organs and tissues.

A negative test result means that your risk for developing breast cancer is no greater than the average person, but a negative test result still is no guarantee that breast cancer will not develop. Sometimes it is impossible to determine if the test result is positive or negative, and you are left with no more information about your risk than before getting the test result. This test result is called ambiguous and ambiguous tests happens about 10% of the time.

Genetic testing can affect relationships with family members. You may want to think about who in your family might want to know your test results, such as your children. One of the decisions you need to make if you get tested is which people you would like to share the results with, and what it might mean for their lives.

The price of genetic testing differs and is often not covered by health insurance. Ask your doctor or other health professionals for more information on testing, privacy issues, and insurance coverage.

Some advantages of having a genetic test are that the test may help you to make medical and lifestyle choices, clarify your cancer risk, decide whether or not to have risk-reducing surgery, and give other family members useful information. A positive test result may help explain why you or other family members have developed cancer.

There are three ways that cancer travels through the body. Cancer can travel through tissue; cancer invades the surrounding normal tissue. Cancer can also travel through the lymph system, cancer invades the lymphatic system and travels through the lymph vessels to lymph nodes in other places in the body. And cancer can also travel through the blood, cancer invades the veins and capillaries and travels through the blood circulation and lymphatic system to other places in the body.

The metastatic tumor is the same type of cancer as the original tumor. For example, if breast cancer travels to the bones, the cancer cells in the bones are actually breast cancer cells. The disease is metastatic, or "distant," breast cancer, not

bone cancer, and should be treated as breast cancer. Metastasis is deadly when the breast cancer cells travels to vital organs, and when other critical tissues of the body form tumors that stop them from working properly.

There are several ways of talking about the risk for developing breast cancer. The three most common ways doctors think about risk are 1 absolute risk; 2 relative risk; and 3, 5 year risk. These three types of risk are quite different. It is important to know these differences between the three types of risk. The first kind of risk is absolute risk. The absolute risk of breast cancer is the overall chances of a person getting breast cancer. When people talk about absolute risk they mean the chance that a woman will develop breast cancer some time during her life. The lifetime risk of breast cancer for a baby girl is about 12% sometime during her life. As she grows older her lifetime risk will decrease but her chances of getting breast cancer in the next 5 years will increase.

Another type of risk is relative risk. The relative risk of breast cancer, genetic mutations, or diseases is the change in risk based on specific risk factors. Imagine two women. One has 2 out of 200 women with breast cancer or a 1% chance of getting breast cancer and the other has 3 out of 200 women with breast cancer, or a 1 and 1/2% chance of getting breast cancer. There is an increase from 1% to 1 and a half percent in terms of absolute risk for breast cancer from the woman in the left to the woman in the right, but a 50% increase in relative risk for breast cancer. That means that the majority of people with a risk factor for breast cancer mutations do not have breast cancer mutations. For example, having a sister with breast cancer or being of Ashkenazi Jewish descent are risk factors for breast cancer mutations, but most women of Ashkenazi Jewish descent and with sisters with breast cancer do not have the breast cancer mutation themselves. The base rate is the starting point when we approach thinking about breast cancer risk from a statistical perspective. Patients need to learn a lot of other information to make good decisions, but thinking about the base rate helps keep things in perspective. There is one final mistake that people commonly make. This mistake happens when people wrongly believe that the risk of having two factors is higher than having either factor. For example, imagine a woman has many risk factors for a breast cancer mutation. Some people may think her risk for developing breast cancer and having a breast cancer mutation is higher than simply having a breast cancer mutation. It is easy to see this is wrong. If you have two or more first- or second-degree relatives who are diagnosed with ovarian cancer, regardless of what age they were diagnosed at, your breast cancer risk is increased. Ovarian cancer has been shown to be highly linked to breast cancer, unlike other cancers.

Your risk for breast cancer is also increased if you have a first- or second-degree relative who was diagnosed at any age with both breast and ovarian cancer. Lastly, your risk for breast cancer is also increased if you have a male relative who has been diagnosed with breast cancer. Breast cancer in men is so rare that having a male relative with breast cancer is considered high risk of having harmful BRCA gene mutations.

There are no standard criteria or rules for recommending when people should be tested for BRCA gene mutations. There is wide agreement on factors that may increase a woman's likelihood of having a breast cancer mutation. Identifying families with a history of breast cancer or ovarian cancer is a first step to gathering information about a person's risk. A genetic counselor can help in detecting and explaining potential risks and will also provide additional information about BRCA genetic tests.

There are other hidden costs of having genetic testing. First, if you are at low risk for a BRCA 1 or BRCA 2 gene mutation, there may be better ways to get the most from your health dollars. Second your genetic test results will be a part of your medical records. There are laws prohibiting employers and health insurance companies from discriminating against you due to genetic testing. The law does not cover life insurance, disability insurance, long-term health insurance, or members of the military. Many women are also concerned about the possibility of illegal discrimination.

People who are tested for BRCA gene mutations usually have a family history of breast cancer and ovarian cancer. Often the best approach is to have the person with cancer tested for genetic mutations first. Now suppose that several women in the family have breast or ovarian cancer but they all test negative for BRCA gene mutations. If a family has a history of breast cancer and or ovarian cancer but no mutations in BRCA genes have been found, a negative test result is not informative. There is no way to tell whether the harmful BRCA gene mutation was not detected by genetic testing, a false negative, or whether the result is a true negative. The family may also have a rare genetic mutation, or a mutation unknown to medical science. A mutation in a gene other than BRCA 1 gene or BRCA 2 gene could increase breast cancer risk but may not be detectable with the genetic tests used today.

One option you have after you have screened positive for a genetic cancer risk is active surveillance. Surveillance means that you screen yourself for cancer in order to detect the disease earlier. The goal of surveillance is to find cancer early, when the cancer is most treatable. Surveillance does not change the risk of developing cancer; though surveillance can decrease the risk of dying from cancer.

Another option for women to prevent breast cancer is surgery that involves removing as much of the at-risk tissue as possible. By removing healthy breast, called a bilateral prophylactic mastectomy or removing healthy fallopian tubes and ovaries, a woman's risk of developing breast or ovarian cancer can be reduced. Surgery does not guarantee that cancer will not develop.

One more method for prevention that women should be aware of is chemo prevention. For chemo prevention a woman would take a drug such as Tamoxifen, which has been shown to reduce the risk of developing breast cancer by about 50 percent for women who are at an increased risk. However, the jury is still out as to how effective Tamoxifen really is in preventing breast cancer.

Since a person's genetic information is health information, it is covered by HIPAA. The Privacy Rule requires that health care provider and others protect the privacy of health information. It sets boundaries on the use and release of health records, and it empowers individuals to control when their health-related information is disclosed. Genetic discrimination occurs when people are treated differently by insurance companies or employers because they have a gene mutation that increases their risk of a disease, such as breast cancer. GINA protects U.S. citizens against discrimination based on their genetic information in relation to health insurance and employment. It is important for U.S. citizens to realize that there are several items that GINA does not cover. Some of these exclusions in GINA are life insurance, disability insurance, and long-term care insurance.

References

Abadie, M. , & Camos, V. (2018). False memory at short and long term. *Journal of Experimental Psychology: General*. Advance online publication. https://doi.org/10.1037/xge0000526

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Cohesive features of deep text comprehension processes. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 2681–2686). Austin, TX: Cognitive Science Society.

Babor, T. F., Higgins-Biddle, J. C., Saunders, J. B., & Monteiro, M. G. (2001). *AUDIT: The Alcohol Use Disorders Identification Test. Guidelines for use in primary care* (2nd ed.). Geneva, Switzerland: World Health Organization, Department of Mental Health and Substance Dependence.

Brainerd, C. J., Reyna, V. F., & Holliday, R. E. (2018). Developmental reversals in false memory: Development is complementary, not compensatory. *Developmental Psychology, 54*, 1773–1784.

Brainerd, C. J., Yang, Y., Reyna, V. F., Howe, M. L., & Mills, B. A. (2008). Semantic processing in "associative" false memory. *Psychonomic Bulletin and Review, 15*, 10351053. https://doi.org/10.3758/PBR.15.6.1035

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology, 2,* 311–350.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language: An introduction to psycholinguistics*. New York, NY: Harcourt Brace Jovanovich.

Cole, P. (September 25, 2008). News writing. *The Guardian*. Retrieved December 5, 2018, from https://www.theguardian.com/books/2008/sep/25/writing.journalism.news.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology, 33A,* 497–505.https://doi.org/10.1080/14640748108400805

Dandignac, M., & Wolfe, C. R. (2018, November). *Writing for Coh-Metrix: A systematic approach to revising texts to foster gist inferences*. Article presented at the 48th meeting of the *Society for Computers in Psychology*, New Orleans, LA.

Dowell, N. 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.

Eurispes & Telefono Azzurro (2012). Indagine conoscitiva sulla condizione dell'infanzia e dell'adolescenza in Italia [Explorative investigation about Italian condition of infancy and adolescence]. Retrieved from http://www.azzurro.it/sites/default/files/SintesiIndagineconoscitivaInfanziaAdolescenza2012_1.pdf.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32,* 221–233.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3,* 371–398.https://doi.org/10.1111/j.1756-8765.2010.01081.x

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. *Elementary School Journal, 115,* 210–229.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40,* 223–234.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101,* 371–395.https://doi.org/10.1037/0033-295X.101.3.371

Haviland, S. E., & Clark, H. G. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior, 13,* 512–521.

Kintsch, W. (1974). *The representation of meaning in memory.* Hillsdale, NJ: Erlbaum.

Kintsch, W. (1988). The role of knowledge in discourse comprehension—A construction integration model. *Psychological Review, 95,* 163–182.https://doi.org/10.1037/0033-295X.95.2.163

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

Magliano, J. P., & Graesser, A. C. (1991). A three-pronged method for studying inference generation in literary text. *Poetics, 20,* 193–232.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27,* 57–86.

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

Merica, D. (2018). Officials from four countries discussed exploiting Jared Kushner. *CNN*, February 28. Retrieved December 5, 2018, from https://www.cnn.com/2018/02/27/politics/jared-kushner-manipulation-mexico-israel-china-uae/index.html

Morelli, M., Bianchi, D., Baiocco, R., Pezzuti, L., & Chirumbolo, A. (2017). Sexting behaviors and cyber pornography addiction among adolescents: The moderating role of alcohol consumption. *Sexuality Research and Social Policy, 14,* 113–121.

Psaki, J. (2018). Jared Kushner's problems are only just beginning. *CNN*, February 28. Retrieved December 5, 2018, from https://www.cnn.com/2018/02/28/opinions/jared-kushners-problems-are-only just-beginning-psaki/index.html

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

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision Making, 7,* 332–359.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences, 7,* 1–75. https://doi.org/10.1016/1041-6080(95)90031-4

Reyna, V. F., Corbin, J. C., Weldon, R. B., & Brainerd, C. J. (2016). How fuzzy-trace theory predicts true and false memories for words,

sentences, and narratives. *Journal of Applied Research in Memory and Cognition, 5,* 1–9. https://doi.org/10.1016/jjarmac.2015.12.003

Reyna, V. F., & Kiernan, B. (1994). The 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., & Kiernan, B. (1995). Children's memory and metaphorical interpretation. *Metaphor and Symbol*, *10*, 309–331. https://doi.org/10.1207/s15327868ms1004_5

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.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*, 36–71. https://doi.org/10.1016/0010-0285(90)90003-M

Singer, M., & Spear, J. (2015). Phantom recollection of bridging and elaborative inferences. *Discourse Processes*, *52*, 356–375. https://doi.org/10.1080/0163853X.2015.1029858

Smith, R. J. (2017). A fuzzy-trace theory approach to exploring verbal overshadowing (*Unpublished Master's thesis*). Miami University, Oxford, Ohio.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York, NY: Academic Press.

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. https://doi.org/10.3758/s13428-015-0592-1

Wilhelms, E. A., Fraenkel, L., & Reyna, V. F. (2018). Effects of probabilities, adverse outcomes, and status quo on perceived riskiness of medications: Testing explanatory hypotheses concerning gist, worry, and numeracy. *Applied Cognitive Psychology*, *32*, 714–726. https://doi.org/10.1002/acp.3448

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., Reyna, V. F., Widmer, C. L., Cedillos, E. M., Fisher, C. R., Brust-Renck, P. G., & Weil, A. M. (2015). Efficacy of a web-based intelligent tutoring system for communicating genetic risk of breast cancer: A fuzzy-trace theory approach. *Medical Decision Making, 35,* 46–59.

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

Wolfe, C. R., Widmer, C. L., Torrese, C. V., & Dandignac, M. (2018). A method for automatically analyzing intelligent tutoring system dialogues with Coh-Metrix. *Journal of Learning Analytics, 5,* 222–234. https://doi.org/10.18608/jla.2018.53.14

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123,* 162–185. https://doi.org/10.1037/0033-2909.123.2.162