Automatic Evaluation of Cancer Treatment Texts for Gist Inferences and Comprehension

Christopher R. Wolfe, Mitchell Dandignac, Rachel Sullivan, Tatum Moleski, and Valerie F. Reyna

**Background.** It is difficult to write about cancer for laypeople such that everyone understands. One common approach to readability is the Flesch-Kincaid Grade Level (FKGL). However, FKGL has been shown to be less effective than emerging discourse technologies in predicting readability. **Objective.** Guided by fuzzy-trace theory, we used the discourse technology Coh-Metrix to create a Gist Inference Score (GIS) and applied it to texts from the National Cancer Institute website written for patients and health care providers. We tested the prediction that patient cancer texts with higher GIS scores are likely to be better understood than others. **Design.** In study 1, all 244 cancer texts were systematically subjected to an automated Coh-Metrix analysis. In study 2, 9 of those patient texts (3 each at high, medium, and low GIS) were systematically converted to fill-the-blanks (Cloze) tests in which readers had to supply the missing words. Participants (162) received 3 texts, 1 at each GIS level. **Measures.** GIS was measured as the mean of 7 Coh-Metrix variables, and comprehension was measured through a Cloze procedure. **Results.** Although texts for patients scored lower on FKGL than those for providers, they also scored lower on GIS, suggesting difficulties for readers. In study 2, participants scored higher on the Cloze task for high GIS texts than for low- or medium-GIS texts. High-GIS texts seemed to better lend themselves to correct responses using different words. **Limitations.** GIS is limited to text and cannot assess inferences made from images. The systematic Cloze procedure worked well in aggregate but does not make fine-grained distinctions. **Conclusions.** GIS appears to be a useful, theoretically motivated supplement to FKGL for use in research and clinical practice.

**Keywords**
patient education, psycholinguistics and medical decision making, readability, understanding cancer

Because anyone can have a serious disease such as cancer, written materials about complex diseases and their treatment must be presented in ways that everyday people with differing levels of education can understand and use in medical decision making. Recognizing this need, professionals developing written materials for patients often assess readability with tools such as the Flesch-Kincaid Grade Level (FKGL). FKGL has long been used in research and patient education and is a good predictor of which kinds of texts are most likely to be understood by children at lower grade levels. However, FKGL 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; specifically, the formula for FKGL is $0.39 \times (\text{words/sentences}) + 11.8 \times (\text{syllables/words}) – 15.59$. Using FKGL to estimate the appropriate grade level for a given text is undoubtedly useful and a ubiquitous technique in assessing the aptness of written communication for patients.

Department of Psychology, Miami University, Oxford, OH, USA (CRW, MD, RS, TM); Department of Human Development, Cornell University, Ithaca, NY, USA (VFR). The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors received no financial support for the research, authorship, and/or publication of this article.

**Corresponding Author**
Christopher R. Wolfe, Department of Psychology, Miami University, 115 Psychology Bldg, Oxford, OH 45045, USA (WolfeCR@MiamiOH.edu).
However, FKGL is neither data rich nor motivated by contemporary theory, and it does not take advantage of emerging discourse technologies. Recent research suggests that traditional readability formulas such as FKGL are less predictive than approaches using advanced natural language processing tools.6

In recent years, discourse technologies have grown tremendously in their sophistication and availability to both researchers and health professionals tasked with helping laypeople understand risk in complex contexts. Discourse technologies (i.e., digital natural language processing software with advanced algorithms interfacing with large databases) have been successfully harnessed for use in developing Intelligent Tutoring Systems (ITS) for medical decision making,7–9 eHealth packages,10 mHealth applications,11 and eLearning platforms.12 One powerful discourse technology is Coh-Metrix,5 which can be used to analyze any text and computes more than 100 linguistic and psycholinguistic variables ranging from single-word variables to measures that span entire texts. Coh-Metrix has been used for research in many contexts including assessing learning environments,13 measuring disorganized speech in schizophrenia,14 and analyzing dialogues about breast cancer and genetic risk.15 The research reported here used Coh-Metrix to analyze texts for patients about cancer, risk, and cancer treatments as a proximal index to the extent to which those texts are likely to help readers comprehend by making appropriate inferences.

A large body of research suggests that making appropriate inferences is a key aspect of comprehension.16,17 Reading requires people to make many different kinds of inferences. Psycholinguistic researchers differ in whether they argue that different kinds of inferences (e.g., those pertaining to causal consequences) are made “online” in real time during reading or shortly afterward when readers reflect on what they read.18 Nonetheless, researchers agree that inference making is essential to understanding texts.

When a patient or health care provider wants to learn more about different types of cancer and cancer treatments, an excellent source of information is the National Cancer Institute (NCI) website.19 Cancer Types on the NCI website has 244 texts about treatments for 48 different kinds of cancer. There is clear evidence that laypeople can learn from these NCI texts. For example, in 3 experiments on the efficacy of the BRCA Gist ITS, Wolfe and colleagues8 found that undergraduates, community members, and participants in a web-based experiment answered 76% of declarative knowledge questions correctly after being randomly assigned to read NCI texts about breast cancer compared with 55% to 57% for participants randomly assigned to a control group. Nonetheless, a critical reading of these texts from the standpoint of fuzzy-trace theory (FTT)20 suggests that there is room for improvement.

**Fuzzy-Trace Theory**

This work is guided by FTT,7,20 which is supported by many years of research and is a dual-process theory emphasizing comprehension, meaning making, and the process of developing mental representations. A key FTT concept is that people simultaneously encode information to create multiple representations that range from verbatim surface details to the bottom-line meaning or “gist.” The theory suggests that people independently process both gist and verbatim representations of experiences. Gist representations denote bottom-line meaning, whereas verbatim representations encode surface details. The theory indicates that gist and verbatim processing occur in parallel. Research suggests that people tend to reason with the most gist-like mental representation available; for example, the reasoning of experts is typically more gist-like than novices,21 and adults rely more on gist representations than children so.22 FTT suggests that authors who wish to create expository texts that facilitate understanding and sound medical decision making should write in ways that help readers form appropriate inferences about the bottom-line meaning of those texts. Thus, medical decision making is facilitated by helping people develop appropriate gist representations rather than emphasizing superficial verbatim details. In support of this goal, we developed an FTT-based method for automatically evaluating texts for gist inferences.

**Gist Inference Score**

Wolfe et al.23 developed a method using Coh-Metrix to create a Gist Inference Score (GIS), which was designed to predict the extent to which people will make meaningful inferences from a text. Coh-Metrix 3.0 is freely available to researchers via a web interface at http://cohmetrix.com/. Unlike traditional approaches that emphasize surface-level verbatim features of text, the idea behind GIS was to develop a proximal index of the likelihood with which readers will develop useful and appropriate gist inferences from a given text.

The GIS formula consists of Coh-Metrix’s 7 variables (see Table 1) converted to z scores and grouped into 3 areas: text cohesion, verb overlap, and word concreteness.
Referential and deep cohesion help facilitate comprehension and gist mental representations. Referential Cohesion assesses the extent to which words that refer to the same entities overlap across an entire text. It captures the degree to which referents in the text are related to each other, facilitating a unified situation model for the reader. A classic example is, “We checked the picnic supplies. The beer was warm,” in which the phrase picnic supplies connects through a bridging inference with the word beer.24 Deep Cohesion measures the degree to which texts contain causal, logical, and intentional connectives. Words such as but, however, because, resulting in, and additionally help readers make meaningful connections among different text passages. These connectives help readers create a situation model that captures the overall meaning of the text. Another important aspect of forming a coherent mental representation of a text is the extent to which actions, as represented by verbs, are related to one another throughout that text. FTT suggests that abstract, rather than concrete, verb overlap might help active readers construct gist mental representations. Coh-Metrix uses 2 variables to assess the extent to which verbs (actions) are interconnected across a text. Verb Overlap LSA uses latent semantic analysis25 to assess the connection between each pair of verbs.23 For example, verbs such as run and sprint score high on verb overlap LSA.23 Because this approach is more abstract, it is weighted positively in the GIS formula. Coh-Metrix also uses Verb Overlap WordNet to measure verb overlap. This approach counts identical verbs or those that are closely associated.23 One example is repeating the verb metastasize rather than using a mix of terms such as metastasize and spread.23 Because Verb Overlap WordNet is highly specific, FTT suggests that it is more likely to lead readers to form mental representations toward the verbatim end of the continuum and is thus weighted negatively.23 The theory suggests that concrete, imaginable words are likely to yield verbatim rather than gist representations.26 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 assesses the extent to which words that are concrete rather than abstract and evoke mental images. Chair is a word that scores high on concreteness, whereas justice would not. Imagability for Content Words is an index of how easy it is to construct a mental image of words.23 Hammer has been identified as high on imagability, and the word reason is low.23 Finally, Hypernymy for Nouns and Verbs represents the specificity of a word within a hierarchy. To illustrate, words such as people are superordinate to words such as patients. Each of the 7 variables in the GIS formula is expressed on a different numeric scale and for this reason were converted to z scores to put them on common footing (see Wolfe et al.23 for details). Once converted to z scores, the 7 variables were combined into the GIS as the unweighted mean (with some variables counted positively and others negatively; see Figure 1).

Wolfe et al.23 found evidence of reliability and validity analyzing verbal data from a memory study and texts from scientific journal articles, news reports, and editorials. In the memory study, gist-condition responses scored 1.96 standard deviations higher on GIS than verbatim-condition responses. GIS predictions were also confirmed in a text analysis study of 50 scientific journal articles. As predicted, texts selected from the Discussion sections of psychology journal articles scored significantly higher (0.83 standard deviations) on GIS than texts selected from the Methods sections of the same

### Table 1: Weight and Definitions of Gist Inference Score Variables

<table>
<thead>
<tr>
<th>GIS Variable</th>
<th>Weight</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Referential Cohesion</td>
<td>+</td>
<td>Overlap in words between sentences and paragraphs forming related threads</td>
</tr>
<tr>
<td>Deep Cohesion</td>
<td>+</td>
<td>The extent to which a text displays causal, logical, and intentional connectives across sentences and paragraphs, which supports representing a situation model of the text</td>
</tr>
<tr>
<td>Verb Overlap LSA</td>
<td>+</td>
<td>The extent to which verbs in a text are meaningfully related directly and indirectly as measured by the cosine of 2 LSA vectors for each pair of verbs</td>
</tr>
<tr>
<td>Verb Overlap WordNet</td>
<td>–</td>
<td>The extent to which verbs are related across a text as measured by verb pairs belonging to the same (closely associated) synonym set</td>
</tr>
<tr>
<td>Word Concreteness</td>
<td>–</td>
<td>The extent to which words in a text are concrete and evoke mental images</td>
</tr>
<tr>
<td>Imagability for Content Words</td>
<td>–</td>
<td>Mean imagability for content words rated in a database of 4825 words</td>
</tr>
<tr>
<td>Hypernymy for Nouns and Verbs</td>
<td>–</td>
<td>Word specificity based on depth in a WordNet hierarchy</td>
</tr>
</tbody>
</table>

*aSee also McNamara et al.5*
journal articles. Also as predicted, editorials from news outlets scored significantly higher (1.12 standard deviations) than news reports on the same topics from the same outlets. In a behavioral study, participants were randomly assigned to versions of a text about breast cancer and genetic risk that differed on GIS by more than 1 standard deviation but differed on other linguistic variables by less than half a standard deviation. Participants randomly assigned to the high-GIS version scored significantly higher on tests of knowledge and comprehension.23

These studies suggest that GIS is broadly applicable over different kinds of expository texts. The goal of the research presented here is to apply GIS to written communication about serious diseases, specifically cancer, risk, and cancer treatments. The current work tests the predictions that NCI texts for patients are helpful but suboptimal, and the GIS of NCI texts predicts reader comprehension.

Study 1: An Analysis of 244 Texts for Patients and Providers on the NCI Website

Methods
The purpose of this study was to automatically assess the extent to which texts for patients and providers help readers form appropriate gist inferences, comparing texts created for patients and health care providers on FKGL and GIS. Our critical reading of the texts through the lens of FTT suggested that those created for patients would score higher on FKGL but that they would actually score lower on GIS.

NCI text corpus on cancer types and treatments. The NCI Cancer Types text corpus19 is composed of 244 texts about treatments for 48 different kinds of cancer. The site has different texts for patients and health care providers on the same topics, although there can be little doubt that some patients read the texts intended for health care providers. There are 145 texts for patients (including overviews) and 99 texts for physicians and other providers. The texts average 425 words in length for a total of 103,786 words of text analyzed.

For both patient versions and provider versions, texts were collected from each of the 48 types of cancer. Under each section of cancer, there were separate sections for each type of treatment. The texts were collected by selecting the first 7 paragraphs of each cancer treatment text. We set several guidelines for “cleaning” the text since Coh-Metrix does not compute irregular text features (i.e., bullet points). First, text that was listed as bullet points was reformatted into paragraph form. If a bullet-pointed sentence ended with a comma, then a comma was placed at the end of that sentence, connecting it to the next sentence. If a bullet-pointed sentence ended with a period, it was treated as a standalone sentence. Headlines were treated as the first sentence in the paragraph that followed. Citation numbers following American Medical Association format (i.e., [1, 2]), were removed from the text because Coh-Metrix cannot provide a meaningful interpretation. Many texts contained a brief paragraph that provided links to other treatments that did not offer additional information on the relevant treatment. These were removed during the text-cleaning process. These data-collecting and -cleaning procedures are unlikely to bias our analyses because the formatting changes were needed for both patient and provider versions of each text. In addition, the sample gathered from the text met the minimum length (>200 words) suggested to compute a meaningful FKGL score.

After collecting and preparing the text, each cancer treatment text was individually submitted to Coh-Metrix for analysis. The data output was saved and coded so that the patient and provider versions of each treatment were paired together (i.e., PatientT1, ProviderT1, PatientT2, ProviderT2, etc.).

Results
Table 2 presents the FKGL and GIS for NCI texts for patients and providers. As expected, texts for patients scored significantly lower on grade level than those for
Table 2  Mean (s) Flesch-Kincaid Grade Level and Gist Inference Score for National Cancer Institute Texts for Patients and Health Care Providers

<table>
<thead>
<tr>
<th></th>
<th>Patients</th>
<th>Providers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flesch-Kincaid Grade Level</td>
<td>9.3 (1.9)</td>
<td>14.5 (2.1)(^a)</td>
</tr>
<tr>
<td>Gist Inference Score</td>
<td>–0.311 (0.40)</td>
<td>–0.011 (0.45)(^a)</td>
</tr>
</tbody>
</table>

\(^aP < 0.0001.\)

providers, \(F(1, 242) = 409.6, P < 0.0001.\) Texts for patients were written at a ninth-grade level according to FKGL compared with a college sophomore level for texts for providers. However, those same texts for patients actually scored lower on GIS than those for providers, \(F(1, 242) = 29.9, P < 0.0001.\) This suggests that it would be difficult for people to form gist inferences from reading these texts.

To illustrate, consider this second paragraph from the NCI text for patients on skin cancer\(^27\) totaling 475 words and earning a GIS of \(-0.93.\)

**Melanoma is a disease in which malignant (cancer) cells form in melanocytes (cells that color the skin).**

The skin is the body’s largest organ. It protects against heat, sunlight, injury, and infection. Skin also helps control body temperature and stores water, fat, and vitamin D. The skin has several layers, but the two main layers are the epidermis (upper or outer layer) and the dermis (lower or inner layer). Skin cancer begins in the epidermis, which is made up of three kinds of cells:

- Squamous cells: Thin, flat cells that form the top layer of the epidermis.
- Basal cells: Round cells under the squamous cells.
- Melanocytes: Cells that make melanin and are found in the lower part of the epidermis. Melanin is the pigment that gives skin its natural color. When skin is exposed to the sun or artificial light, melanocytes make more pigment and cause the skin to darken.

The passage is rich in potentially useful information, but it is difficult for readers to ascertain what is most important or make inferences beyond the stated information. Figure 2 provides a visual representation of how the text on melanoma (partially excerpted above) scores on each of the 7 components of GIS from left to right: Referential Cohesion, Deep Cohesion, Verb Overlap LSA, Verb Overlap WordNet, Word Concreteness, Imagability for Content Words, and Hypernymy for Nouns and Verbs. Here, the bars with horizontal lines represent variables that are weighted positively in determining the GIS, and those with dots are weighted negatively. This text scores appropriately higher on Verb Overlap LSA than Verb Overlap WordNet and also positively on Referential Cohesion. However, it is quite low on Deep Cohesion, and all 3 measures of concreteness (Word Concreteness, Imagability for Content Words, and Hypernymy for Nouns and Verbs) are quite high.

An example of an NCI text with a high GIS score is the patient text on pituitary tumors\(^28\) totaling 579 words with a GIS of \(+0.41.\) The last paragraph of that text reads, “Signs of a pituitary tumor include problems with vision and certain physical changes. Signs and symptoms can be caused by the growth of the tumor and/or by hormones the tumor makes or by other conditions. Some tumors may not cause signs or symptoms.” Although not perfect, with this text, it is easier for readers to understand that there is not a “necessary and sufficient” link between vision problems and pituitary tumors.

**Discussion**

The low GIS scores for the patient texts correspond with our own reading, that it is difficult for readers to make inferences about the bottom-line meaning of these texts. As exemplified in Figure 2, the biggest reason for this difficulty with the melanoma text is the high level of concreteness that can be readily seen in the passage above. A lack of deep cohesion is another problem with this text. It seems to move from one topic to the next with few cues to help the reader understand how the propositions fit together. Having found evidence that the NCI texts for patients score relatively low on GIS, our next step was to see if the automatically assessed observable characteristics of text measured by GIS predict the extent to which people actually comprehend those texts.

**Study 2: Comprehension of Cancer Treatment Texts by GIS**

To assess readers’ text comprehension of patient texts from the NCI website, we employed a fill-the-blanks or Cloze procedure.\(^29,30\) With the Cloze procedure, words in a text are systematically replaced with blanks of a standard length and presented to readers. Their task is to fill the blanks with the missing word or another word that retains the meaning of the sentence, with a higher percentage correct corresponding to greater comprehension. The Cloze procedure has been used by researchers and classroom teachers for more than 65 years\(^29\) to assess and improve student comprehension\(^31\) and, more recently, to measure the ability of machines in reading
comprehension.\textsuperscript{32} Whereas in most studies the Cloze procedure is used to measure the comprehension of the reader (human or machine), we adapted the procedure here to assess the comprehensibility of texts for patients about cancer types and treatments. As Friedman and Hoffman-Goetz note in their review of readability and comprehension instruments used for print and web-based cancer information (p. 362), “Cloze is a valid and reliable measure of patient comprehension that can be used to pretest written health information . . . though it has a heavy time burden for administrators and participants.”\textsuperscript{33} Our prediction is that, in aggregate, texts scoring high on GIS should yield a greater percentage of correct missing words, including others that convey a comparable meaning of the sentence.

**Methods**

*Sample and recruitment.* Participants were 162 Miami University undergraduate psychology students who participated for course credit. The participant pool had a mean age of 19 y ($s = 1.4$) and were 80.4\% white and 70.4\% female. The majority, 60.5\%, were psychology majors, with 54\% being first-year students, 27\% sophomores, and 11\% juniors. An additional 10 participants started the study but did not complete it because of technical difficulties or failure to follow the procedure. Those 10 participants were replaced by others in the same experimental conditions. They were recruited online and participated at computers in the first author’s laboratory.

*Experimental design.* Each participant received 3 Cloze-modified NCI patient cancer texts, with 1 each of high-, medium-, and low-GIS score selected from 9 total texts with 3 at each GIS level. The design was completely balanced, controlling for all permutations of high-, medium-, and low-GIS texts (27) times all permutations of the order in which they received the texts of each GIS level first, second, and third ($6 \times 27 = 162$ unique combinations).

*Materials.* Each text was selected from the NCI website for patients\textsuperscript{19} as being low, medium, or high on GIS. The 3 low-GIS texts, on the topics of leukemia, myeloproliferative neoplasms, and skin cancer, yielded GISs of $-0.8$, $-0.5$, and $-0.93$, respectively. The 3 medium-GIS texts on gallbladder cancer, plasma cell neoplasms, and soft-tissue sarcoma produced GIS of $-0.34$, $-0.33$, and $-0.32$, respectively. Finally, the 3 high-GISs texts on
pheochromocytoma and paraganglioma, pituitary
tumors, and thyroid cancer had GISs of 0.33, 0.38, and
0.7, respectively.

The procedure for creating the fill-the-blanks task was
identical for each of the 9 NCI texts for patients. First,
we used the versions previously used to assess GIS, with
all figures, headings, and references to figures removed
and presented in paragraph form. The first sentence was
presented intact to assist readers as a topic sentence.30
Starting with the second sentence, every 10th word was
replaced by a standard-length blank skipping articles and
other small function words (e.g., the, and, of) and exclud-
ing acronyms (e.g., VHL for “von Hippel-Lindau” syn-
drome). Filled blanks were then scored as correct if the
same exact word was used (excluding spelling mistakes
and typographical errors) or if they retained the same
meaning (e.g., “the risk of reoccurrence is 50%” com-
pared with “the chance of reoccurrence is 50%”). Two
research assistants made judgments about 1404 Cloze
items, and they made the same judgments in 93% of the
cases (the interrater reliability was 0.93). The characteris-
tics of each text are presented in Table 3. The (untrans-
formed) constituent Coh-Metrix variables for these 9
texts are presented in Supplementary Appendix A. Note
that participants never saw the complete original text,
and thus, they filled in words based on the meaning of
surrounding text (rather than verbatim memory for the
omitted words).

Procedure. Upon arrival, participants signed an
informed consent sheet. They were seated at a computer
with a mouse and keyboard and given verbal instructions
on how to complete the study. Participants opened a
digital folder containing a high-, moderate-, and low-
GIS text, numbered in the order they were assigned to
read and complete them. Each text contained empty text
boxes in which participants entered the best-fitting word
that fit the sentence. Participants were given explicit
instructions to not return to a previous text after com-
pleting it. The entire procedure took approximately 30
min. Participants were given course credit following the
end of the study.

Results
The mean percentage of Cloze items correct (with exact
words or different words retaining the same meaning)
were subjected to a repeated-measures analysis of var-
iance assessing the GIS level across 3 different sets of 3
texts. As predicted, there was a significant effect for GIS
level, $F(2, 477) = 3.94, P = 0.020$. High-GIS texts scored
a mean of 0.631 (63% correct), which was significantly
greater than 0.580 for medium-GIS texts and 0.598 for
low-GIS texts, using Fisher’s exact test to contrast high
GIS to medium and low GIS, $F(1, 192) = 6.77, P =
0.0096$. There was also a significant effect for text sets
with $F(2, 477) = 12.20, P < 0.0001$, meaning that there
were differences in percentage of Cloze items correct
among texts at the same GIS level.

Interestingly, a post hoc analysis separating exact
word matches from other words retaining the same
meaning showed the opposite pattern, with a mean cor-
rect of 0.424 for high-GIS texts, 0.462 for medium-GIS
texts, and 0.496 for low-GIS texts, $F(2, 477) = 10.08,
P = 0.0001$. The reason for this discrepancy is that for
high-GIS texts, 0.21 (21%) of the blanks were filled with
different words retaining the same meaning of the sen-
tence, compared with 0.12 for medium-GIS texts and
0.10 for low-GIS texts, $F(2, 477) = 51.25, P < 0.0001$. Over
the 9 texts, the correlation between GIS and per-
centage of blanks filled with different words retaining

<table>
<thead>
<tr>
<th>GIS Level</th>
<th>Title (T = Text Number within Cancer Type)</th>
<th>Word Count</th>
<th>No. Cloze</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Leukemia T7</td>
<td>520</td>
<td>44</td>
</tr>
<tr>
<td>Low</td>
<td>Myeloproliferative Neoplasms T2</td>
<td>316</td>
<td>28</td>
</tr>
<tr>
<td>Low</td>
<td>Skin Cancer T2</td>
<td>474</td>
<td>43</td>
</tr>
<tr>
<td>Medium</td>
<td>Gallbladder Cancer T1</td>
<td>434</td>
<td>40</td>
</tr>
<tr>
<td>Medium</td>
<td>Plasma Cell Neoplasms T1</td>
<td>434</td>
<td>39</td>
</tr>
<tr>
<td>Medium</td>
<td>Soft Tissue Sarcoma T3</td>
<td>493</td>
<td>45</td>
</tr>
<tr>
<td>High</td>
<td>Pheochromocytoma and Paraganglioma T1</td>
<td>408</td>
<td>37</td>
</tr>
<tr>
<td>High</td>
<td>Pituitary Tumors T1</td>
<td>579</td>
<td>50</td>
</tr>
<tr>
<td>High</td>
<td>Thyroid Cancer T1</td>
<td>409</td>
<td>36</td>
</tr>
</tbody>
</table>

FKGL, Flesch-Kincaid Grade Level; GIS, Gist Inference Score.
the same meaning was \( R = 0.447 \), meaning that GIS accounted for about 20% of the variance.

We subjected a subset of texts to additional qualitative analysis. First, we chose 2 texts, on myeloproliferative neoplasms with a GIS of –0.80 and pituitary tumors with a GIS of 0.38 as most representative of low- and high-GIS texts, respectively. To control for differences in participant characteristics, we analyzed each of the Cloze items for those 2 texts using data from the 18 participants who were randomly assigned to both of these texts. Thus, we controlled for individual differences in reading ability, vocabulary, and other characteristics that generally predict text comprehension.

Table 4 presents the words generated for 6 Cloze items by the same 18 participants for 3 of the highest sentences on comparable meaning words for the pituitary tumors text (GIS = 0.38) and 3 for the myeloproliferative neoplasms text (GIS = –0.80). For 3 of the 6 Cloze sentences, none of the participants used the word that was replaced by a blank in the original NCI text. In the high-GIS text, there were not only more participants who generated an alternative word compared with the low-GIS text (38 v. 25), but they also generated more different terms (14 v. 7). For example, words that adequately filled the blank for “control the menstrual cycle in women and the ___ of sperm in men.” include production, creation, development, level, amount, and number. By way of contrast, for the low-GIS text with the highest number of participants generating alternative words that retain a comparable meaning, “Having certain genetic conditions increases the ___ of developing a pituitary tumor.” yielded just 2 reasonable alternatives to “tissues”: parts and area (see Table 4). It appears that some Cloze items in some texts lend themselves more readily to being filled with words that retain the same meaning than others. This appears to be a potentially important property of the texts themselves, rather than a product of participant characteristics. In other words,
some texts lend themselves more easily to retaining their basic meaning using different words, and this characteristic of texts appears to be captured by GISs.

**Discussion**

Applying the Cloze procedure to “authentic” texts for patients found on the NCI website, we found that the GIS, derived solely from the observable characteristics of texts analyzed with a publicly available discourse technology, predicted subsequent comprehension among readers who were randomly assigned to read those texts and fill-the-blank tasks. These results are predicted by FTT,20 which emphasizes the dual-process nature of encoding and the primacy of forming appropriate gist representations. Apparently, texts that score relatively high on GIS are able to cue a larger range of semantically viable responses than those with low scores. Although the research paradigm is dramatically different, these findings are conceptually compatible with models derived from FTT predicting both true and false memories for words, sentences, and entire narratives.17

**General Discussion**

The evidence from these 2 studies suggests that GISs are a useful tool for predicting the likelihood that readers will make useful inferences from reading texts about cancer and cancer treatments, which is a key dimension of comprehension. Although the patient texts on the NCI website have been demonstrated to be effective in contributing to reader knowledge,7 these data suggest that there are reasons to be concerned that many readers will find it difficult to infer the bottom-line meaning of these texts.

We argue that part of what makes high-GIS texts more likely to yield inferences is that they have the characteristic of making it easier for readers to represent the meaning of those texts in their own words rather than in the exact printed words in the text. This explains the finding that superior Cloze performance on high-GIS texts primarily stemmed from different words retaining a similar meaning rather than exact matches. This is consistent with the research on false memories22,26 and suggests that in future research, it may be theoretically interesting to examine readers’ recall and false memories after reading cancer texts.

GIS is a proximal variable, and a higher GIS suggests that readers are more likely to make gist inferences. Lower GIS scores may suggest that readers are more likely to form a verbatim representation of the text, recalling the exact wording, although perhaps at the expense of deep understanding. However, texts that are incoherent may also draw low GIS scores. Thus, in working with particular texts, it is often useful to examine the complete constellation of GIS variables, as represented in Figure 2. Although FTT suggests that helping readers form appropriate gist representations is central to good medical decision making,8 for some purposes, authors may wish to ensure that key information is encoded verbatim. GIS graphs such as Figure 2 suggest that repeating concrete words, as indicated by the 3 bars on the right of Figure 2, may help encourage verbatim memories. However, it is difficult to think of cases for which low cohesion scores on the left side of Figure 2 would be helpful for any purpose. With respect to authoring texts to promote gist inferences, Dandignac and Wolfe34 have developed guidelines for revising texts to achieve higher GIS. Providing direction for authors is another respect in which GIS adds value above and beyond traditional FKGL, which suboptimally suggests using shorter words in short, choppy sentences.

**Limitations**

Although promising, these studies suffer from a number of shortcomings and limitations. First, GIS (and indeed all Coh-Metrix variables) are limited to texts and cannot assess images or the integration of text and image. The texts on the NCI website typically have illustrations, and the experience of readers visiting those pages and looking at those images while reading accompanying texts may be quite different than that predicted by GIS or experienced by participants in study 2.

Second, although working with authentic texts from the NCI website increases the utility of the research, it does come at the expense of some experimental control. Thus, although we conducted research on multiple texts at each level, in study 2 we tested only 9 of the 244 texts that were analyzed in study 1. We do not know whether our results might have differed had we sampled a wider range of texts. It is also the case that the college student participants used in study 2 are different from cancer patients in many respects. Our focus here was on the texts themselves rather than reader characteristics, but assessing GIS with cancer patients differing in reading ability, active vocabulary, and other characteristics will have to wait for future work. A follow-up study investigating knowledge effects with patients and with health care providers might also be worthwhile to examine whether GIS can differentiate expert comprehension from that of novices.

Another shortcoming stems from weakness in comparing different texts with the Cloze procedure compared
with the more typical case in which researchers are interested in comparing different kinds of readers on the same texts. In selecting every 10th word, we took a systematic approach. However, differences among texts within GIS level may be due to the chance of selecting Cloze words (for example, those repeated throughout the text) or due to fundamental differences between texts not captured by GIS. The strategy of selecting every 10th word seems to work well in aggregate over many texts at the same GIS level, but in future research, it would be helpful to develop another systematic method that targets more theoretically relevant words. Another potentially promising approach would be to create high-GIS and low-GIS versions of the same text using the same Cloze terms.

Finally, we do not know how high GIS scores need to be to make a difference in comprehension and inference making. In study 2, we found no improvements from GIS scores of about $-0.8$ to about $-0.33$. Positive GIS scores were associated with superior Cloze performance, and although we can make relative comparative judgments, additional research will be needed before we can interpret specific scores (e.g., GIS = 0) in absolute terms.

Conclusions

Although they are written at about a ninth-grade level as measured by FKGL, the NCI texts for patients are suboptimal in helping readers form appropriate inferences. GIS predicts comprehension of these authentic texts, suggesting the utility of GIS in research and the clinical potential of editing texts to achieve higher GIS scores and better inference making and comprehension among readers. Thus, GIS appears to be a useful and theoretically motivated supplement to traditional readability measures such as FKGL for use in research and clinical practice.

Acknowledgments

We would like to thank the following Miami University students who assisted in analyzing texts and collecting data from research participants: Sam Almond, Navdeep Bais, Luejack Baker, Alyssa Conyers, Natalie Goldfein-Killilea, Ayiana Green, Pamela Janiro, Vasantha Kalindi, Jordan Oldham, Sheysel Sanchez, Michael Suponic, Chase Tiry, Eleni Vidalis, Ellie Wallace, Margo Weber, Katie Weidner, and Vince Werthmann.

ORCID ID

Christopher R. Wolfe https://orcid.org/0000-0001-6905-9457

Supplementary Material

Supplementary material for this article is available on the Medical Decision Making Web site at http://journals.sagepub.com/home/mdm. –RJF, 8/16/19

References


