Do People Remember Fiction and Nonfiction Texts Differently?

Maybe Not. Replicating Zwaan (1994) in 2020

Maria Triantafyllopoulos,[1,2] Binyan Li,[3] Margaret Schnabel,[2] Fritz Breithaupt[3,4, 5]

[1]Indiana University, Department of Psychological and Brain Sciences
[2]Indiana University, Hutton Honors College
[3]Indiana University, Cognitive Science Program
[4]Indiana University, Department of Germanic Studies

[5]Corresponding author: fbreitha@indiana.edu

Abstract

Do people remember texts differently based on whether they think the text is fiction or nonfiction? An influential study by Rolf Zwaan (1994) found that participants who read a text labeled as fiction demonstrated greater Surface Level comprehension (word memory), while those who read the same text as nonfiction demonstrated greater Situational Level (gist) comprehension of the text. More than twenty-five years later, we find no difference in participants' comprehension of fiction and nonfiction. This difference may be caused by the original study's small sample size, but it is also possible that cultural changes since 1994 (in particular the widespread digitalization of reading) may have affected how fiction and nonfiction are read. We replicated the original study in two alternative ways, and consider influences of age and reading habits.

Introduction

Beginning at a young age, we gain the ability to differentiate between fantasy and reality, fictional and factual accounts (Morison & Gardner, 1978; Samuels & Taylor, 1994; Wellman, 1992). Children at the age of three are already aware that characters in fictional stories are not real (Woolley & Cox, 2007). Because fiction and nonfiction are such distinct categories, it makes intuitive sense that our cognitive processing of the two genres could be different. Galak & Nelson (2011) and Jacobs (2011) argue that when people read nonfiction, they are more focused on gathering information; when they read fiction, they are more focused on their own enjoyment. And other studies suggest that when reading fiction, people are especially focused on cues of how the story will progress (Rapp & Gerrig, 2006), engaging in narrative empathy (Keen, 2006), or enjoying the beauty of aesthetic word choices (Kraxenberger & Menninghaus, 2017). For example, the phrase "Winter is coming" could, in a nonfiction context, direct us to reconsider certain choices in our wardrobe—or in the context of fiction, evoke emotions, invite an imaginative transportation into a fictional world, and lead to allegorical and metaphorical readings of the text (Mar et al., 2011; Green, Garst, & Brock, 2004; Green & Brock, 2000; Jacobs, 2015; Oatley, 2012; Oatley, Mar, & Djikic, 2012).

But to what degree do people process texts they believe to be nonfiction differently from texts they believe to be fiction? Notably, the novel study design by Zwaan (1994) showed significant differences in text comprehension based on whether participants were told that a short text was either an excerpt from a work of fiction or from a nonfictional news article: In the nonfiction condition, readers emphasized an understanding of the gist of the text, while in the fiction condition, readers of the same texts emphasized precise word memory in recall tests. This study laid the groundwork upon which dozens of other discourse-related studies are based.

Building on the insights of Zwaan (1994), Prentice and Gerrig (1999) suggest that nonfiction processing entails a higher degree of effort – involving the systematic integration of knowledge – than fiction processing does. Green et al. (2004) suggest that high reader engagement, commonly seen with reader transportation into fictional narratives, alter story processing, namely causing readers to be less critical of the story's factual claims. Furthermore, neuroimaging studies have alluded to a cognitive basis for the similarities and differences in reading fiction versus nonfiction texts (Altmann et al. 2012, Moss & Schunn, 2015). The reading of narrative texts, which are more commonly associated with fiction, involves the activation of many of the same brain regions as the reading of expository, i.e. information only, texts. The notable exception is dorsomedial prefrontal cortex (dmPFC) which is activated only for narratives, in addition to social cognition and Theory of Mind (Moss & Schunn, 2015).

However, other studies have not found a conclusive difference in how people process texts labeled as fiction or nonfiction. For example, in the context of persuasiveness, Green, Garst, Brock, and Chung (2006) did not consistently find higher persuasiveness for texts labeled as nonfiction compared to those labeled as fiction. Strange and Leung (1999) indicate that texts labeled as fiction or nonfiction yielded similar effects on how readers perceived social issues. Green and Brock (2000) and Green, Chatham, & Sestir (2012) find that the labeling of fiction or nonfiction does not impact the degree to which readers felt that they were "transported" into the text, the degree of emotional intensity experienced by the reader, or the general beliefs the reader had about the text. Hartung, Withers, Hagoort & Willems (2017) found no difference based on whether a story was labeled as based on true events or not. Hence, whether a study finds a difference in how readers process fiction and nonfiction seems to depend on which aspect of processing is studied (reader goals, expectations, persuasiveness, etc.) and how it is operationalized.

Yet it is also possible that such differences in processing are culturally dependent, and the cultural circumstances surrounding reading and reading habits have certainly changed with the rise of the internet. Today, fiction, nonfiction, and hybrids often share the same platform (Crossman, 1997). Hyperlinks within social media can point to texts that are not always clearly positioned as nonfiction or fiction. So far, results are mixed as to whether and to what degree text comprehension has been altered by new reading habits. Some studies have found no difference in recall ability or comprehension depending on whether the text is read in print or online/digital (d'Haenens, Jankowski, & Heuvelman, 2004; Singer & Alexander, 2017; Tran, Carrillo, & Subrahmanyam, 2013). Other studies have indicated that online reading leads to poorer text comprehension, perhaps linked to higher cognitive load, fatigue, and distraction (Coiro, 2014; DeStefano & LeFevre, 2007; Jeong, 2012; Macedo-Rouet, Rouet, Epstein, & Fayard, 2003; Mangen, A,Walgermo & Brønnick, 2013; Zumbach & Mohraz, 2008).

Having in mind the uncertainty regarding effects of genre expectation, as well as the small participant sample of the original study, we aimed to replicate Zwaan's 1994 experiment twenty-five years after the fact, bearing in mind that potentially different results might be partially attributable to cultural-technological changes. Zwaan selected several texts (roughly 200 words each), extracted from either a newspaper article or a piece of literature. He selected these texts based on their genre ambiguity—i.e., participants in a pretest could not reliably determine whether the text had come from a newspaper or a novel. In the main study, a different set of participants read one of these text pieces, which was labeled as either an excerpt from literature (fiction) or a newspaper article (nonfiction). After participants read the texts, text comprehension was analyzed by means of six questions that followed the Kintsch discourse comprehension model, in which comprehension is described by three interconnected categories or "Levels" (Fletcher & Chrysler, 1990; Kintsch, Welsch, Schmalhofer, & Zimny, 1990). These Levels are Surface Level, Textbase, and Situational Model.

Table 1. Terminology

Terminology	**Definition**
Genre	Fiction (novel) or nonfiction (news)
Level (of Representation)	Type of reading comprehension or memory: Surface, Textbase, Situational
Surface Level	Verbatim memory of text
Textbase	Basic comprehension of text and its semantic content
Situational Model	The overall situation implied by text; influenced both by prior knowledge of the reader, reader expectations, and strength of the Textbase

Zwaan (1994)'s found that reading comprehension was influenced by expected Genre: participants who read the text as fiction demonstrated better Surface Level comprehension, while participants who read the text as nonfiction demonstrated better Situational Model comprehension. Our replication reexamines Experiment 1 from Zwaan (1994). But to account for the potential influence of cultural changes, we also collected data on participants' reading habits to account for any possible interaction between genre expectation and a tendency to read online.

Study 1 Replication of Zwaan (1994)

In Study 1, we followed the procedures of Experiment 1 from Zwaan 1994, with very minor modifications as indicated. We also added some elements, such as improved statistical tools and considerations of age and reading habits by the participants.

Methods

Pretest 1: Stimuli Texts

We selected nine texts, 200-250 words in length, from news articles and literary works, that we predicted could reasonably pass as excerpts from either a novel or newspaper article. In the pretesting, we asked participants on Mechanical Turk whether they thought a particular text came from either a piece of fiction or nonfiction and then to assess the confidence of their opinion on a scale from 1 to 6 (N=59, mean age = 36, 41% female).

Based on these results, we selected six of the nine texts with the highest degree of ambiguity, such that the average rating for the six texts was 53% nonfiction and 47% fiction with a standard deviation of 10.9, which indicates that participants could not reliably distinguish between the genres. In addition, the average participant confidence rating for their choice of genre was 3.3 on a six-point scale with a standard deviation of 1.2. Three of the six texts selected as stimuli for the replication were excerpts from works of fiction and three from newspaper articles, such as *The New York Times;* for full texts see appendix S1. Zwaan (1994) did not report the percentage breakdown of fiction or news identification, only that subjects could not reliably determine the true genre of the texts in a pretest.

Pretest 2: Generating Test Items for Each Text

Zwaan (1994) and Schmalhofer & Glavanov (1986) measured the d' scores for each Level of Representation by applying signal detection theory to sentence test items specific to a piece of text. Four categories of test items were obtained for each text: verbatim, paraphrase, plausible inferences, and implausible inferences. We also followed this procedure by recruiting participants from Mechanical Turk (N = 48, mean age = 33, 56% females) and following the methods outlined by Zwaan (1994) to these test items. To generate verbatim and paraphrases items, sentences were randomly selected from the six texts obtained from Pretest 1 and were either kept word for word (making them "verbatim") or they were reworded and thus used as paraphase test items. To obtain the plausible and implausible inferences, the participants were randomly given one of the six texts – presented and described as either fiction or nonfiction – and after reading were asked open-ended questions about the passages, such as "Why did X event occur?" For example, we asked "Why did Mrs. Bohm shoot her husband?" for a text extracted from a news article that described a murder-suicide without providing a motivation for the event. These questions required participants to generate an inference about the text because the answer was not explicitly stated in the original text. Inferences that occurred frequently in both genre conditions were used as plausible inference test items for the main study. For the implausible inferences we used items that were only generated once by participants and that we rated as implausible.

Subjects

We recruited 173 participants on Amazon Mechanical Turk for this study who did not participate in the Pretests. 11 were excluded for failing a confirmation task, leaving us with 81 participants in the fiction condition and 81 participants in the nonfiction condition. The average age was 40.8 years, 63 were female, 97 male, and 2 declined to state gender.

Replication Experiment

We presented participants with a random sequence of four randomly selected texts from Pretest 1. Participants were randomly assigned a fiction or news condition for all these texts. In the original experiment, Zwaan's instructions for the fiction condition were: "The following texts are all excerpts from novels by famous Dutch and other European literary authors. Please read these texts just like you would normally read a novel." And for the nonfiction condition: "The following texts are all excerpts from news stories that appeared in either NRC-Handelsblad or de Volkskrant [two Dutch quality journals]. These stories describe important events that happened in the 1980s. Please read these stories just like you would normally read a news story." We followed the same format, replaced "European" and "Dutch" with "American" and did not include the phrase alluding to historic since it was not included in the fiction condition and seemed imbalanced. To limit participant bias, we chose not to include the names of real newspapers. In addition, after the instructions, we asked them on the next screen: "Which will you be reading?" with choices for "News from newspapers," "Fiction from novels," and "I do not know." We only counted particpants who made the correct choice. Zwaan did not issue such a genre confirmation. Moreover, in the original study, after each text was read the participant was issued a statement which was either consistent or inconsistent with the text, which the participant indicated by pressing a yes/no button. The goal of these statements was to ensure that the participants were focused on reading. However, Zwaan did not analyze the results from these questions and as we used a confirmation task earlier to only select participants who were accurately reading and understanding the task, we chose not to utilize this part of his procedure.

After reading each of the texts, participants were given a list of sentences or phrases in random order that were either: 1) taken *verbatim* from the text; 2) *paraphrases* of passages from the text; 3) *plausible inferences* about the text (as determined in Pretest 2); or 4) outlandish, *incorrect inferences* about the text. Participants were asked to determine whether each phrase was taken verbatim from the text they had just read by marking on a scale from 1 to 6 how confident they were that each sentence had appeared verbatim in the passage (6 being highly confident). Like

Zwaan (1994), we used 4 verbatim sentences for each passage, 2 paraphrases, 2 correct inferences, and 2 incorrect inferences for a total of related questions (one plausible inference and one implausible inference) for each text. Afterwards, we collected biographical data including age, sex, education, and reading habits of news and fiction, separately for reading in print and online.

The format of our replication introduced several small changes. In contrast to Zwaan (1994), we did not present subjects with a sentence by sentence display of each text since we were not interested in reading time (Experiment 2 in Zwaan (1994)). The test was entirely done on screen and hence did not involve a separate instruction booklet. Zwaan had a pool of 36 undergraduate students from a Dutch university, our pool was 162 participants recruited by Mechanical Turk. With this method, we collected an average 53.33 responses for each text each in the the fiction or news condition, while Zwaan only collected 16. Zwaan showed each of his participants six experimental texts, four of which came from originally nonfiction, news sources and two of which came from fictional, novel excerpts. However, Zwaan later excluded the results from one of the newssource text due to its responses being an outlier in comparison to the other texts. Thus, he calculated results from five texts for each participant. We used six texts, with equal numbers of nonfiction and fiction newssources, and we opted instead to show each participant four texts to limit fatigue effects.

Calculating d'

Zwaan (1994)'s principal analyses were performed on d' scores of Surface Level, Textbase, and Situational Model comprehension. We followed standard procedures for calculating d' score. The d' is a measure that makes use of signal detection theory and recognition memory models to produce a comprehensive measure of discourse comprehension (Kintsch, 1978; Kintsch, 1990; McNicol, 2005; Schmalhofer & Glavanov, 1986), for full explanations see S2 and S3. When a d' score is calculated based on so few inputs, there is a considerable chance that participants will answer all the questions correctly, resulting in an infinite d' score. To prevent this, we placed an artificial ceilings on d' scores (McNicol, 2005; see Appendix S3) although the cutoff point will necessarily arbitrary.

Zwaan's calculations used an unweighted scale in that a participant who gave a high confidence rating (4, 5, or 6 on the six-point scale) for a particular test item indicated that the participant viewed the response as an "old sentence" and that it was thus taken verbatim from the text passage (Zwaan 1994). This is considered a "yes" response. A score of 1, 2, or 3 out of 6 for a particular test item indicated that the participant did not believe that the phrase was taken verbatim from the text, making it a "no" response. The hit rate (HR) for Surface Level was a yes response to a true verbatim test item, while the false alarm rate was a yes response to a paraphrase test item. The HR for the Textbase was a yes response to a paraphrase test item, while the false alarm rate was a yes response to a plausible inference. The HR for the Situational Model level was a yes response to a plausible inference, while the false alarm rate was a yes response to an implausible inference.

To take the degrees of confidence into account, we first replicated Zwaan's calculation of unweighted responses. We then deviated from Zwaan by calculating weighted responses that account for the discrepancies between a 4/6, 5/6, and a 6/6 when calculating d' scores for each level. Weighted scores were calculated following McNicol (2005), by taking the average of the confidence scores for each category of sentence type and dividing them by 8 to transform them into a weighted value which would not yield a zero (0/6) or a perfect score (7/7). Finally, due to the low numbers of test items each participant answered for each text (10 sentences), we also took an aggregated d' score of the responses by calculating a d' score for each participant based on their cumulative responses. Rather than calculating a d' score for each text, all test items a participant answered (40 question total, 10 for each story), were combined to produce a single d' score for each Level for each participant. For a more detailed description, see Appendix 3 (S3).

Analysis Method

Zwaan (1994) used a mixed ANOVA to determine whether d' comprehension scores differed for the fiction and non-fiction conditions. d' was treated as the dependent variable; genre, Level (Surface, Textbase, Situational) were treated as the independent variables. Using an artificial floor and ceiling can render the data non-normal with values stacking up at the ends of the scale, which was confirmed by a Shapiro-Wilk's test of normality, $W = 0.95$, $p < .001$. Consequently, ANOVA is not an appropriate model for testing our hypotheses since it assumes normality in the data. (Still, in the spirit of a replication, we will also report the results of a mixed ANOVA using d').

In place of a frequentist mixed ANOVA, we rely on Bayesian parameter estimation. We used the hierarchical model introduced in chapter 9 section 2.4 of Kruschke (2015), which assesses percent correct for each condition collectively as well as for each individual subject. In signal detection, the advantage of d' over percent correct is that the former accounts for both "sensitivity" and "decision criteria" while the latter accounts for only "sensitivity." However, in our case, where we are comparing two groups who are responding to the same questions, there is no reason to expect the "decision criterion" to meaningfully differ between the two groups. Therefore, we can more simply and directly test Zwaan (1994)'s hypothesis by estimating percent correct with our

Bayesian hierarchical model. For the sake of replication, we will report the results of a frequentist mixed ANOVA using d'.

Three separate models were constructed, one for each Level. For this analysis all participants' responses were treated as either correct or incorrect, just as in Zwaan (1994). After estimating the collective correct response rate for each condition, we can take the difference between the two conditions' percent-correct rate (from the posterior distributions), and if the 95% most likely values (95% HDI) of the difference does not include zero, then we can reject the null hypothesis that there is no difference between the conditions. (The Bayesian 95% HDI can be interpreted as one interprets a frequentist 95% confidence interval.)

To determine whether participants' reading habits might influence the potential effects of expected genre, we used Bayesian two-way ANOVA to test the effects of 1) genre and 2) reading habits on d' (for more on Bayesian ANOVA see chapter 20 of Kruschke, 2015). We performed three tests, one with each Level of d' as the dependent variable. We recorded reading habits by asking the number of hours per week each participant read 1) online news, 2) online fiction, 3) physical news, and 4) physical fiction. We then calculated the proportion of print-physical reading each participant performed. Since the vast majority of our 154 participants read more online fiction or news than in print fiction or news, we decided to code reading habits as a two-class categorical variable: participants were binarily classified based on whether they were in the top 50th percentile of proportion of print reading in our dataset. If cultural changes toward online reading had a sizable influence on genre expectation and comprehension, then we would expect to detect an interaction effect between genre and reading habits in our data. For a more detailed explanation of hypothesis testing using Bayesian parameter estimation see Kruschke and Liddell (2018).

Results

Bayesian Hierarchical Model

Results of the Bayesian hierarchical model suggests that the difference in percent correct between the nonfiction and fiction conditions was not credibly nonzero for any of the three Levels. For the Surface and Situational Levels, the 95% HDI is essentially centered around zero; for Textbase, although the most likely estimate of the difference between nonfiction and fiction conditions is 4.11% in favor of nonfiction, the 95% HDI includes zero and 2.56% in favor of fiction. Hence, the difference in Textbase cannot be taken as a reliable and meaningful difference.

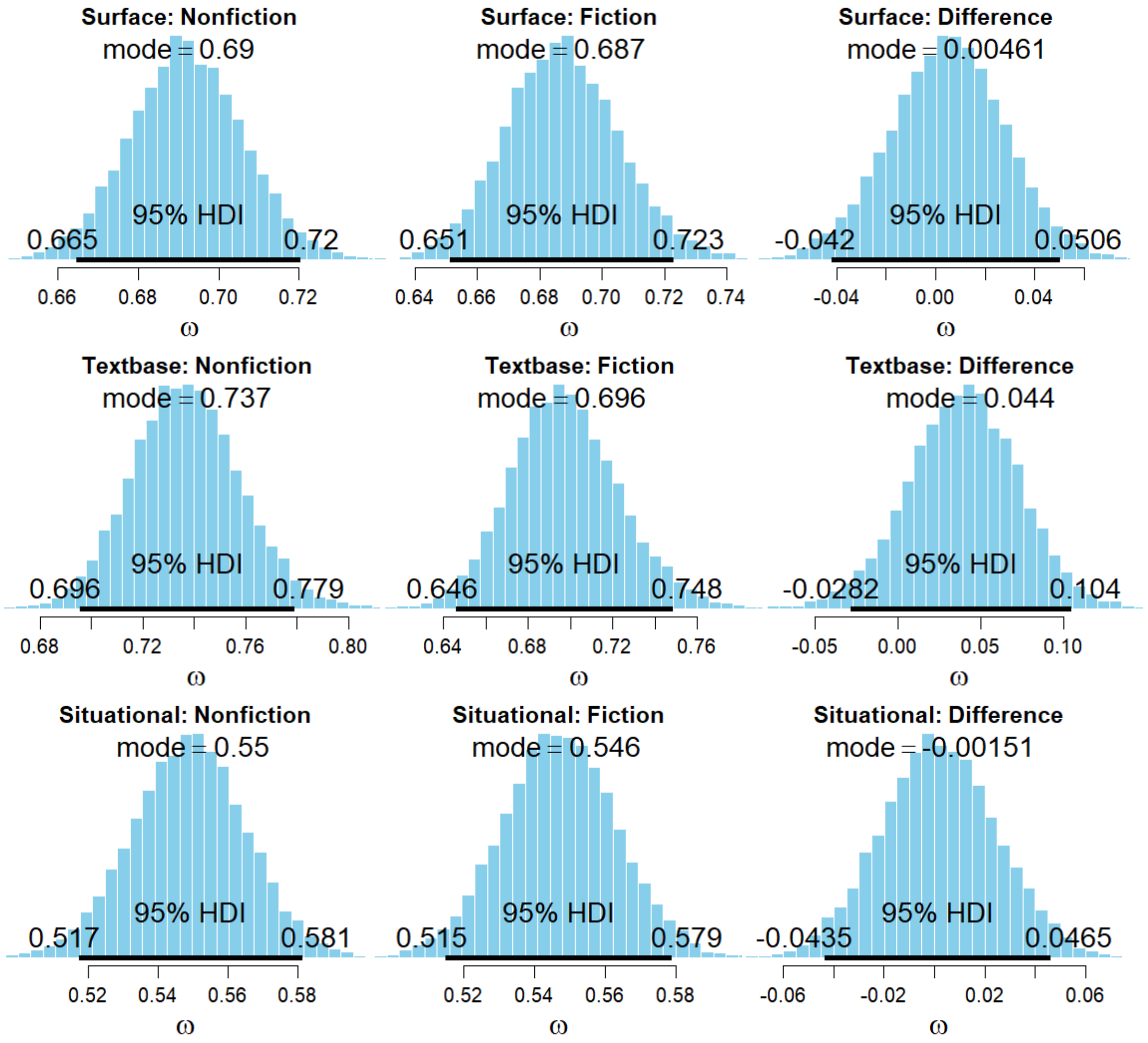

Figure 1. 95% HDIs of the posterior distributions. Left: modal estimate for percent correct in nonfiction condition. Middle: modal estimate for percent correct in fiction condition. Right: modal estimate for the difference in percent correct between the nonfiction and fiction conditions. We cannot conclude with 95% confidence that the difference is nonzero for any of the types of reading comprehension (since the 95% HDI crosses zero in all three cases).

Mixed ANOVA

Table 2. Mean Memory Scores as a Result of Genre Expectation from Zwaan (1994)

		Level of Representation for Zwaan 1994		
Genre		**Surface**	**Textbase**	**Sit. Model**
News				
	Mean	-0.052	1.20	0.66
	SD	0.54	0.74	0.80
Fiction				
	Mean	0.27	1.46	0.14*
	SD	0.42	0.90	0.71

* signifies not significantly different from zero

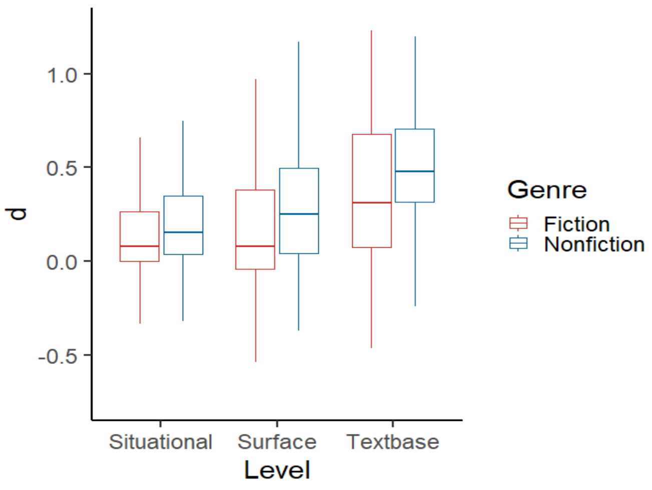

Figure 2. Memory Scores by Expected Genre in Study 1.

Despite violation of the assumption of normality, for the sake of exact replication, we also performed a frequentist mixed ANOVA with d’ as the dependent variable, Level (Surface, Textbase, Situational) as a within-subject independent variable and genre as a between-subject independent variable. Mauchly's test indicated a violation of the assumption of sphericity for Level and the interaction between genre and Level (both W = 0.925, $p < .01$, ε = .930) so we used the Greenhouse–Geisser corrections of sphericity. We found a main effect of Level, F(1.86, 282.77) = 22.16, $p < .001$. But we found neither a main effect for genre nor Zwaan (1994)'s interaction between genre and Level, F(1.86, 282.77) = 0.24, p = .77.

Bayesian Two-Way ANOVA

To answer whether our divergent findings from Zwaan (1994) could be explained by cultural changes toward online reading we performed three two-way Bayesian ANOVAs with genre and reading habits (tendency toward online or print reading relative to others in the dataset) as independent variables and the the d' scores of the three Levels as the dependent variable in each analysis. In line with our Bayesian hierarchical model and frequentist mixed ANOVA, none of the three tests found a main effect of genre expectation on d' as the 95% HDI spanned zero in each case (Surface 95% HDI: -0.225 – 0.022; Textbase 95% HDI: -0.220 – 0.012; Situational 95% HDI: -0.132 – 0.027). For Surface and Textbase, but not Situational, we found a main effect of reading habits on d'. The group that read a smaller proportion of physical text scored higher on Surface and Textbase comprehension (Surface 95% HDI: 0.017 – 0.256; Textbase 95% HDI: 0.063 – 0.301; Situational: -0.042 – 0.115). We found no interaction effects between genre and reading habits for any Level of d' (Surface 95% HDI: -0.051 – 0.407; Textbase 95% HDI: -0.059 – 0.385; Situational 95% HDI: -0.145 – 0.126)

Study 2 Replication with One Text per Participant

We considered the possibility that genre-related effects could be mediated by the length of the testing session or through the repetition of reading multiple texts. Thus, to ensure that fatigue effects did not play a significant a role in our Study 1 and potentially in Zwaan (1994), we created a study in which participants only read a single text in either the fiction or nonfiction condition.

Methods

We used the same tasks, texts, instructions, materials, and methods of analysis from Study 1, but we reduced the number of texts for each participant to just one text. Participants were randomly assigned one text with the specific genre instructions. For the same reason, we also reduced the number of questions asked per participant. Instead of 10 questions, we used 6 test items with 2 verbatim, 2 paraphrases, and 2 inference-related questions (one plausible inference and one implausible inference) for each text. This method is more aligned with the Schmalhofer, F., & Glavanov, D. (1986) study, on which Zwaan (1994) is partially based. Schmalhofer and Galvanov had an equal ratio between verbatim, paraphrase, plausible inference, and implausible inference test items, and we decided to follow a similar approach to balance the Levels of Representation instead of skewing them heavily towards verbatim test items, as Zwaan (1994) did.

Subjects

We recruited 350 participants on Amazon Mechanical Turk for this study. There were 175 participants in the fiction condition and 175 participants in the nonfiction condition. The average age was 37.56 years, 49% were female, and 1 declined to state gender.

Results

Hierarchical Bayesian Model

Results of the Bayesian hierarchical model coïncide with Study 1's results: the difference in percent correct between the nonfiction and fiction conditions was not credibly nonzero any of the three Levels, see Fig. 3. For the Surface Level, the 95% HDI is essentially centered around zero; for Textbase the most likely estimate is 2.4% in favor of fiction, and for Situational the most likely estimate is 4.76% in favor of fiction. But in each case, the 95% HDI crosses zero and so we cannot conclude that any of these differences between the conditions are reliable or meaningful.

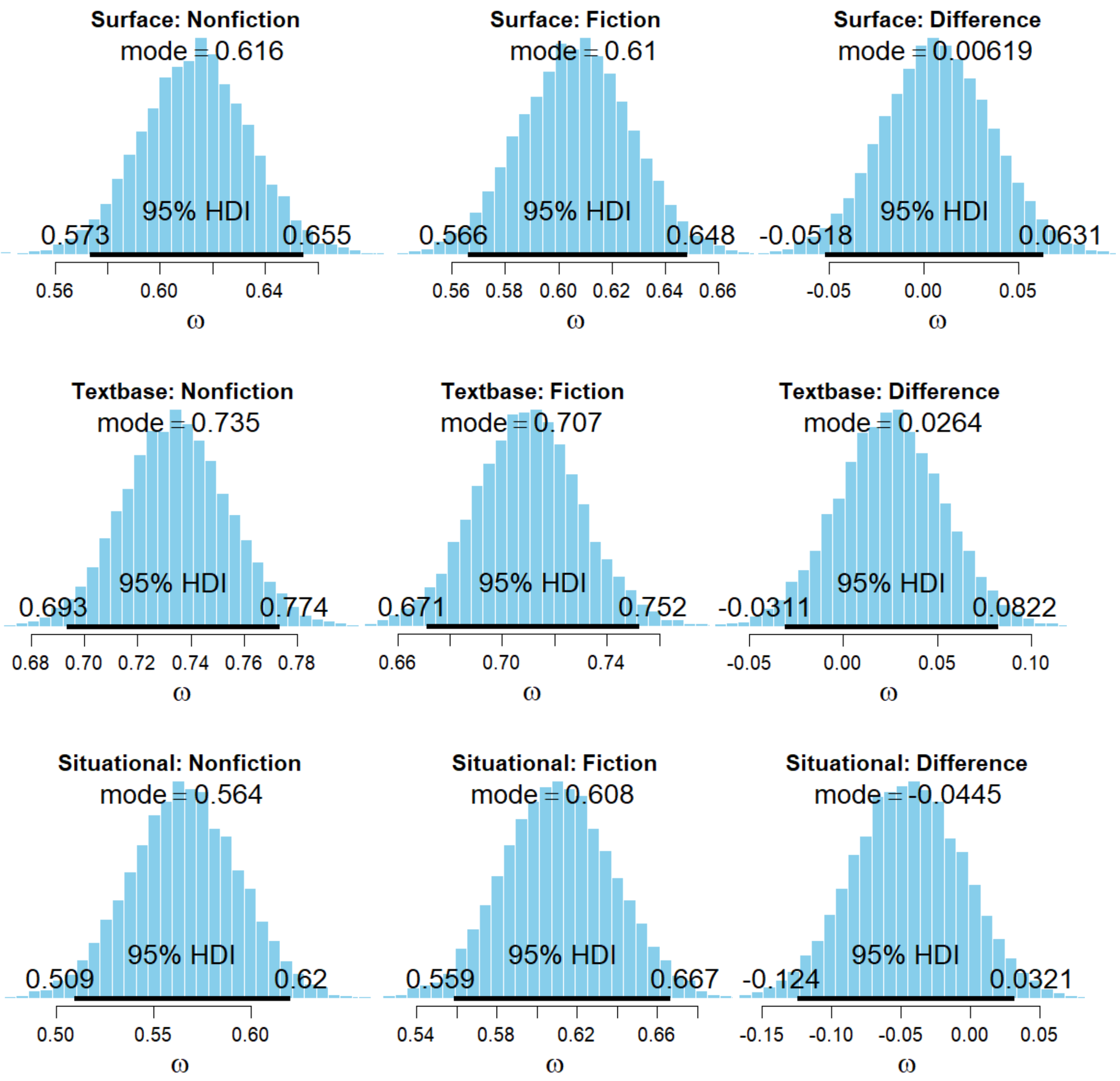

Figure 3. 95% HDIs of the posterior distributions. Left: modal estimate for percent correct in nonfiction condition. Middle: modal estimate for percent correct in fiction condition. Right: modal estimate for the difference in percent correct between the nonfiction and fiction conditions. We cannot conclude with 95% confidence that the difference is nonzero for any of the types of reading comprehension (since the 95% HDI crosses zero in all three cases).

Mixed ANOVA

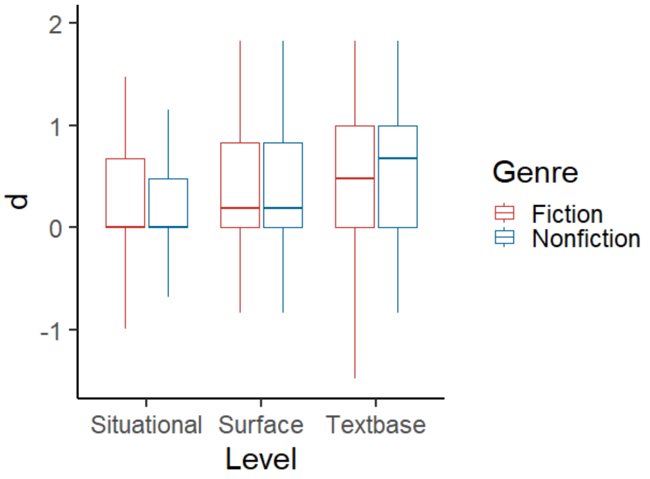

Figure 4. Memory Scores by Expected Genre in Study 2.

As in Study 1, we also performed a mixed ANOVA. Mauchly's test indicated a violation of the assumption of sphericity for Level and the interaction between genre and Level (both W = 0.969, p < .01, ε = .970) so we used the Greenhouse–Geisser corrections of sphericity. We found a main effect of Level, F(1.94, 675.12) = 11.71, p < .001. But we found neither a main effect for genre nor Zwaan (1994)'s interaction between genre and Level, F(1.94, 675.12) = 2.01, p = .136. Similarly to Study 1, there was no clear difference between the unweighted and weighted treatments of the results, see S5.

Bayesian Two-Way ANOVA

As with Study 1, we performed three Bayesian two-way ANOVAs, one for each Level of d'. In line with all other results Studies 1 and 2, none of the three tests found a main effect of genre expectation on d' as the 95% HDI spanned zero in each case (Surface 95% HDI: -0.189 – 0.123; Textbase 95% HDI: -0.328 – 0.019; Situational 95% HDI: -9.52e^{-5} – 9.68e^{-5}). For Textbase, but not Surface or Situational, we found a main effect of reading habits on d'. The group that read a smaller proportion of physical text scored higher on Textbase comprehension (Surface 95% HDI: 0.252 – 0.062; Textbase 95% HDI: 0.022 – 0.374; Situational: -9.73e^{-5} – 9.74e^{-5}). We found no interaction effects between genre and reading habits for any Level of d' (Surface 95% HDI: -0.399 – 0.166; Textbase 95% HDI: -0.221 – 0.409; Situational 95% HDI: -2.00e^{-4} – 1.85e-^{4}).

Overall Discussion

Our studies could not replicate the Zwaan (1994) finding of a difference in processing between fact and fiction. We tested whether paratextual information—i.e., whether text is labeled as fiction or news—makes a difference for comprehension and immediate recall. Zwaan (1994) had found that this genre information makes a difference with fiction leading to better Surface Level recall (verbatim memory), while nonfiction leads to a better Situational understanding of the overall meaning of the text. Our findings show that readers have near identical strength of comprehension for all Levels of Representation (Surface Level, Textbase, Situational Model), with a minor, potential difference only in Textbase. Neither of our experiments with differing number of texts for each participants produced a noticeable difference in comprehension under the two genre conditions.

A possible explanation of the differences between our results and those of Zwaan's is that Zwaan's findings are not replicable, due to an effect of small sample size or unreliable statistics. There were 36 student participants in the original 1994 experiment, while we include results from 512 participants. There were also small differences between our two studies and Zwaan's. While our studies aimed for a meaningful replication, we did not choose a laboratory setting due to its artificiality. Perhaps the state of being under surveillance in a laboratory enforced the genre priming for students to a higher degree. However, if this is the case, it might be more meaningful to know what happens without such high degree of artificial observation in a partly more naturalistic setting when participants are in a place of their own choosing. Part of Zwaan's text was conducted on paper and partly on screen, while our test relied entirely on the internet (onscreen). We cannot rule out that our test was influenced by the delivery format (see, for example, Noyes and Garland 2008).

There are some other possibilities for the difference in findings.

1. An explanation could be population differences between the Dutch undergraduate students, no average age given, from Zwaan (1994) and our US American participants on Mechanical Turk (2019-2020) with an average age of 38. While generational differences may account for the difference, our data do not clearly support such a hypothesis, since the different generations today score similarly on comprehension and we did not record interaction effects. Still, we cannot account for the Dutch generation 25 years ago.

2. A notable decline in general reading time occured from 1994 to 2020. The average reading of news and literature has been on the decline in the United States and in the Netherlands for the past few decades (Knulst & Kraaykamp, 1998; National Endowment for the Arts (NFAH), 2007; Perrin, 2018; Statistics Netherlands, 2008, March 18). The average Dutch person in 1994

read approximately 0.9 hours of books and 2 hours of newspapers daily, totaling about 4.2 hours when considering other reading sources such as personal communication and magazines (Knulst & Kraaykamp, 1998). This is twice as much as the average of our participants, who reported reading a combined 87.12 minutes of fiction and nonfiction. Perhaps, and this a speculative, longer average reading times make the development of different reading processes for each genre more likely. While our data do not offer support for this theory since reading times by our participants did not show interaction effects with comprehension scores in the genre groups, this does not rule out the possibility that long-term effects might be at work here (readers could have changed reading habits).

3. Another shift in reading habits concerns the move toward online reading for both news and fiction from 1994 to 2020. Instead of newspapers on printed paper and the aesthetically pleasing books, our subjects reported reading almost twice as much online text than physical text. The growing convergence of media into a single, online platform for all forms of communication and text types could precipitate a similarity of neural representations. Indeed, fMRI and MEG studies have demonstrated that similar visual stimuli provoke similar neural activation patterns (Connolly et al., 2012; Mur et al., 2013; Wardle, 2016). The typography and physical format of texts, such as line spacing, coloring and the number of columns, can also influence the reader's comprehension, reading speed and exhaustion (Dyson, 2005; Ganayim & Ibrahim, 2013). It could thus be a possible explanation of our findings that new online reading habits have reduced previous differences in processing of fact and fiction. In our studies, online reading did improve memory scores, but we found no evidence of reading habits influencing genre expectations. Still, we cannot rule out the possibility that online reading nullified the genre difference that might have existed prior to online reading.

The general evidence concerning the impact of online reading is large. For example, online reading is connected with skimming information and has affected suspectablility to fake news (Bronstein, M. V., 2019; Carr, 2008; Carr, 2011; Lazer et al., 2018; Wolf, 2018). Media theorist Fredrich Kittler famously predicted changes brought forth by technological media, remarking, "The general digitization of channels and information erases the differences among individual media. Sound and image, voice and text are reduced to surface effects, known to consumers as interface" (Kittler, 1999, pg. 1).

4. In addition to changes in the delivery format, the genres themselves and genre expectations may have undergone significant changes that bring them closer together. It is possible that news today appears more like fiction and, likewise, that fiction deemphasizes typical fiction elements. Furthermore, people today may be habituated to see them as related; for example, literature courses at school are including more nonfiction than in previous decades and do not emphasize precise aesthetic word choices, but rather social dynamics that drive actions. While historically people may have read fiction partly for entertainment in the form of aesthetic precision of word choices and news for the sake of better understanding larger contexts, today similar expectations may be increasingly govern both fiction and nonfiction. That is, differentiated goals for reading may be changing. It seems that readers today reach an increasingly similar understanding and mental reconstruction of text pieces regardless of the whether it appears to be fiction or news (Sage, Piazzini, Downey, & Masilela, 2020).

In short, we could not replicate Zwaan (1994). It is possible that larger participant numbers nullified the original results. These results do not mean that people cannot distinguish between fact and fiction (Altmann et al., 2012; Green et al., 2004). They also do not mean that there are no processing differences between fiction and nonfiction, but it suggests that strong differences in comprehension and recall do not exist.

References

Altmann, U., Bohrn, I. C., Lubrich, O., Menninghaus, W., & Jacobs, A. M. (2012). Fact vs fiction--how paratextual information shapes our reading processes. *Social cognitive and affective neuroscience*, *9*(1), 22–29. doi:10.1093/scan/nss098

Baron, N. S. (2015). *Words onscreen: The fate of reading in a digital world*. Oxford University Press, USA.

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, *8*(1), 108-117.

Carr, N. (2008). Is Google making us stupid?. *Yearbook of the National Society for the Study of Education*, *107*(2), 89-94.

Carr, N. (2011). *The shallows: What the Internet is doing to our brains*. WW Norton & Company.

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y. C., ... & Haxby, J. V. (2012). The representation of biological classes in the human brain. *Journal of Neuroscience*, *32*(8), 2608-2618.

Coiro, J. (2014). Online reading comprehension: challenges and opportunities.

Crossman, D. M. (1997). The evolution of the World Wide Web as an emerging instructional technology tool. *Web-based instruction*, 19-23.

DeStefano, D., & LeFevre, J. A. (2007). Cognitive load in hypertext reading: A review. *Computers in human behavior*, *23*(3), 1616-1641.

Dyson, M. C. (2005). How do we read text on screen. *Creation, use, and deployment of digital information*, 279-306.

D'Haenens, L., Jankowski, N., & Heuvelman, A. (2004). News in online and print newspapers: Differences in reader consumption and recall. *New Media & Society*, *6*(3), 363-382.

Einstein, G. O., McDaniel, M. A., Bowers, C. A., & Stevens, D. T. (1984). Memory for prose: The influence of relational and proposition-specific processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 133-143.

Fletcher, C. R., & Chrysler, S. T. (1990). Surface Forms, Textbases, and Situation Models: Recognition Memory for Three Types of Textual Information. *Discourse Processes*, *13*(2), 175-190.

Galak, J., & Nelson, L. D. (2011). The virtues of opaque prose: How lay beliefs about fluency influence perceptions of quality. *Journal of Experimental Social Psychology*, *47*(1), 250-253.

Ganayim, D., & Ibrahim, R. (2013). How do typographical factors affect reading text and comprehension performance in Arabic?. *Human factors*, *55*(2), 323-332.

Genette G. Fictional narrative, factual narrative. Poetics Today. 1990;11(4):755–74

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*, *79*(5), 701.

Green, M. C., Garst, J., Brock, T. C., & Chung, S. (2006). Fact versus fiction labeling: Persuasion parity despite heightened scrutiny of fact. *Media psychology*, *8*(3), 267-285.

Green, M. C., Chatham, C., & Sestir, M. A. (2012). Emotion and transportation into fact and fiction. *Scientific Study of Literature*, *2*(1), 37-59.

Green, M. C., Garst, J., & Brock, T. C. (2004). The power of fiction: Determinants and boundaries. In L. J. Shrum (Ed.), *The psychology of entertainment media: Blurring the lines between entertainment and persuasion* (pp. 161-176). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Hamburger, K. The Logic of Literature. 2nd edn. Bloomington: Indiana University Press, 1973.

Hartung, F., Withers, P., Hagoort, P., & Willems, R. M. (2017). When Fiction Is Just as Real as Fact: No Differences in Reading Behavior between Stories Believed to be Based on True or Fictional Events. *Frontiers in psychology*, *8*, 1618. doi:10.3389/fpsyg.2017.01618

Jacobs, A. M. (2011). "Neurokognitive poetik: elemente eines modells des literarischen lesens (Neurocognitive poetics: elements of a model of literary reading)," in Gehirn und Gedicht: Wie Wir Unsere Wirklichkeiten Konstruieren (Brain and Poetry: How We Construct Our Realities), eds R. Schrott and A. M. Jacobs (München: Carl Hanser Verlag), 492–520.

Jacobs A. M. (2015). Neurocognitive poetics: methods and models for investigating the neuronal and cognitive-affective bases of literature reception. Front. Hum. Neurosci. 9:186. 10.3389/fnhum.2015.00186

Jeong, H. (2012). A comparison of the influence of electronic books and paper books on reading comprehension, eye fatigue, and perception. *The Electronic Library*, *30*(3), 390-408.

Keen, S. (2006). A theory of narrative empathy. *Narrative*, *14*(3), 207-236.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological review*, *85*(5), 363.

Kintsch, W., Welsch, D., Schmalhofer, F., & Zimny, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and language*, *29*(2), 133-159.

Kittler, F. A. (1999). *Gramophone, film, typewriter*. Stanford University Press.

Knulst, W., & Kraaykamp, G. (1998). Trends in leisure reading: Forty years of research on reading in the Netherlands. *Poetics*, *26*(1), 21-41.

Kraxenberger, M., & Menninghaus, W. (2017). Affinity for poetry and aesthetic appreciation of joyful and sad poems. *Frontiers in psychology*, *7*, 2051.

Kruschke, J. K. (2015). *Doing Bayesian data analysis, second edition: A tutorial with R, JAGS, and Stan*. Academic Press/Elsevier, Burlington, MA, 2nd edition.

Liddell, T. M., and Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, 79, 328-348.

Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D. and Schudson, M., (2018). The science of fake news. *Science*, *359* (6380), 1094-1096.

Lindsay, J. (2010). Children's access to print material and education-related outcomes: Findings from a meta-analytic review.

Macedo-Rouet, M., Rouet, J. F., Epstein, I., & Fayard, P. (2003). Effects of online reading on popular science comprehension. *Science Communication*, *25*(2), 99-128.

Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International journal of educational research*, *58*, 61-68.

Mar R, Oatley K, Djikic M, et al. (2011) Emotion and narrative fiction: Interactive influences before, during, and after reading. Cognition and Emotion 25(5): 818–833.

McNicol, D. (2005). *A primer of signal detection theory*. Psychology Press.

Morison, P., & Gardner, H. (1978). Dragons and Dinosaurs: The Child's Capacity to Differentiate Fantasy from Reality. *Child Development, 49*(3), 642-648. doi:10.2307/1128231

Moss, J., & Schunn, C. D. (2015). Comprehension through explanation as the interaction of the brain's coherence and cognitive control networks. *Frontiers in human neuroscience*, *9*, 562.

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in psychology*, *4*, 128.

National Endowment for the Arts (NFAH), Washington, DC. (2007). *To Read Or Not to Read: A Question of National Consequence. Research Report# 47*. ERIC Clearinghouse.

Noyes, J. M., & Garland, K. J. (2008). Computer-vs. paper-based tasks: Are they equivalent?. *Ergonomics*, *51*(9), 1352-1375.

Oatley K. (2012). The cognitive science of fiction. WIREs Cognitive Science, 3(4), 425–430. https://doi.org/10.1002/wcs.1185.

Oatley, K., Mar, R. A., & Djikic, M. (2012). The psychology of fiction: Present and future. *Cognitive literary studies: Current themes and new directions*, 235-249.

Perrin, A. (2018). Nearly one in five Americans now listen to audio-books. Pew Research Center.

Perrin, A. (2018). Who doesn't read books in America. *Pew Research Center*, *23*.

Prentice, D. A., & Gerrig, R. J. (1999). Exploring the boundary between fiction and reality.

Rapp, D. N., & Gerrig, R. J. (2006). Predilections for narrative outcomes: The impact of story contexts and reader preferences. *Journal of Memory and Language*, *54*(1), 54-67.

Sage, K., Piazzini, M., Downey IV, J. C., & Masilela, L. (2020). Reading from print, laptop computer, and e-reader: Differences and similarities for college students' learning. *Journal of Research on Technology in Education*, 1-20.

Samuels, A. and Taylor, M. (1994), Children's ability to distinguish fantasy events from real-life events. British Journal of Developmental Psychology, 12: 417-427. doi:10.1111/j.2044-835X.1994.tb00644.x

Schmalhofer, F., & Glavanov, D. (1986). Three components of understanding a programmer's manual: Verbatim, propositional, and situational representations. *Journal of Memory and Language, 25*(3), 279-294.

Searle, J. R. (1975). The logical status of fictional discourse. *New literary history*, *6*(2), 319-332.

Singer, L. M., & Alexander, P. A. (2017). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The journal of experimental education*, *85*(1), 155-172.

Sperduti M. (2016). The paradox of fiction: emotional response toward fiction and the modulatory role of self-relevance. Acta Psychol. 165, 53–59. 10.1016/j.actpsy.2016.02.003

Statistics Netherlands. (2008, March 18). Dutch read fewer books, but bookshops record higher turnover rates. Retrieved from https://www.cbs.nl/en-gb/news/2008/12/dutch-read-fewer-books-but-bookshops-record-higher-turnover-rates

Strange, J. J., & Leung, C. C. (1999). How anecdotal accounts in news and in fiction can influence judgments of a social problem's urgency, causes, and cures. *Personality and Social Psychology Bulletin*, *25*(4), 436-449.

Tran, P., Carrillo, R., & Subrahmanyam, K. (2013). Effects of online multitasking on reading comprehension of expository text. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, *7*(3).

Wardle, S. G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S. M., & Carlson, T. A. (2016). Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. *Neuroimage*, *132*, 59-70.

Wellman, H. M. (1992). *The MIT Press series in learning, development, and conceptual change. The child's theory of mind.*Cambridge, MA, US: The MIT Press.

Wolf, M. (2018). *Reader, come home: The reading brain in a digital world*. Harper.

Woolley, Jacqueline & Cox, Victoria. (2007). Development of beliefs about storybook reality: PAPER. Developmental science. 10. 681-93. 10.1111/j.1467-7687.2007.00612.x.

Zumbach, J., & Mohraz, M. (2008). Cognitive load in hypermedia reading comprehension: Influence of text type and linearity. *Computers in Human Behavior*, *24*(3), 875-887.

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 920.

Appendix

S1. All texts we used.

Text 1 (Nonfiction – The New York Times, 1980)

When the official driver knocked at the door of the East German Finance Minister's country house one morning in May, there was no answer. Inside, the man found the two occupants, Finance Minister Siegfried Bohm and his wife, Ruth, lying dead.

According to a police investigation that has never been published but that was disclosed privately to diplomats and other officials, Mrs. Bohm killed her husband with his revolver and then turned the gun on herself.

Faced with a scandal, the Communist leadership tried to hush up the family tragedy. As in any authoritarian system where bad news is considered no news, the party ordered all details of the case suppressed. News organizations did not report that a shooting had occurred and that two persons had been killed.

Instead, readers of Neues Deutschland and other party newspapers were told the day after the deaths that Mr. Bohm, a 51-year-old member of the central Committee and Finance Minister since 1966, had died "as a result of an accident." The papers did not mention Mrs. Bohm's death nor did they explain when and how the "accident" had taken place.

Text 2 (Fiction, ~2000)

In Ignatius' first major policy statement to Parliament he appealed to his fellow countrymen to tighten their belts in the face of economic problems and promised to end graft and corruption in public life and warned the electorate that no one holding an official position could feel safe unless he led a blameless life. He ended his maiden speech with the words, "I am disposed to clear out Nigeria's Augean stables." Such was the impact of the minister's speech that it failed to get a mention in the Lagos Times. Perhaps the editor considered that, since the paper had covered the Speeches of the previous sixteen ministers in extension, his readers might feel they had heard it all before. Ignatius, however, was not to be disheartened by the lack of confidence shown in him and set about his new task with vigor and determination. Within days of his appointment he had caused a minor official at the Ministry of Food to be jailed for falsifying documents relating to the import of grain. The next to feel the bristles of Ignatius' new broom was a leading Lebanese financier, who was deported without trial for breach of the exchange control regulations.

Text 3 (Nonfiction, 2010)

Growing up, Rod Scurry never doubted he would play in the majors, if not as a pitcher then as a hitter. In high school he once hit a five-hundred-foot home run. But despite his batting prowess, he had always been a pitcher at heart. In the 1960s, when he was just a child, he stacked mattresses against the wooden fence in the backyard of his Nevada home and hurled fastballs at them. He had always had power. But then there was the hook. He could sweep his curveball in at such an angle the ball would bend between a batter's legs. Frequently compared to the preeminent lefty of all time, Sandy Koufax, Scurry drove himself to live up to the complement. This desire propelled him out of bed at 5:30 a.m. to jog to school through high mountain air and sometimes freezing temperatures just so he could get extra pitching practice in at the Hug High gym before the opening bell rang. On game days, when his teachers believed him to be studiously tending to his work in the classroom he would in fact be poring over index cards he had made that listed the tendencies of the opposing team's big hitters.

Text 4 (Nonfiction – The New Yorker 1988)

When Lieutenant Colonel John Paul Vann, in starched cotton khakis and a peaked green cap, strode through the swinging doors of Colonel Daniel Boone Porter's office in Saigon, shortly before noon on March 23, 1962, he struck Porter as a man no one could keep down. Porter soon had the feeling that if the commanding general were to tell this junior lieutenant colonel he was surrendering direction of the war to him John Vann would say, "Fine, General," and take charge.

The commanding general in Vietnam then was Paul Harkins, who had made his reputation as the principal staff aide to George Patton in the Second World War. In December, 1961, President Kennedy had committed the arms of the United States to the task of suppressing a Communist-led rebellion and preserving South Vietnam as a separate state governed by an American-sponsored regime in Saigon. The following month, he had decided to create the United States Military Assistance Command Vietnam (macv), in Saigon, and appoint Harkins to head it. During 1962, the President was to nearly quadruple the number of American military men in South Vietnam, from thirty-two hundred to eleven thousand three hundred.

Text 5 (Fiction – 2014)

Gorbachev in his deepest soul believed in the truth. Since Lenin died, every Soviet leader had been a liar. They had all glossed over what was wrong and declined to acknowledge reality. The most striking characteristic of Soviet leadership for the last sixty-five years was the refusal to face facts. Gorbachev was different. As he struggled to navigate through the storm that was battering the Soviet Union, he held on to that one guiding principle, that the truth must be told. Dimka was full of admiration.

Both Dimka and Gorbachev were pleased when Erich Honecker was deposed as leader of East Germany. Honecker had lost control of the country and the party. But they were disappointed by his successor. To Dimka's annoyance, Honecker's loyal deputy, Egon Krenz, took over.

All the same...The Soviet Union could not permit the collapse of East Germany. Perhaps the USSR could live with democratic elections in Poland and market forces in Hungary, but Germany was different. It was divided, like Europe, into East and West, communist and capitalist; and if West Germany were to triumph that would signal the ascendancy of capitalism, and the end of the dream of Marx and Lenin.

Text 6 (Fiction, 2015)

Sahira had been five when Saleh was captured, twenty-three when he was released. Sahira and her mother waited as the first, second, and third round of prisoners of war exchanged between Iraq and Iran, long after the end of the eight-year war. They asked returning prisoners if they had met Saleh, if they had known him. No one had. Sahira's mother died in 1996. Saleh made it out in 1998.

That winter, Sahira slept three nights in her car at Al Nusoor Square in Baghdad, where it was promised the last of the POWs would be brought home. Sahira brought an old ID photo of her father, which she'd enlarged and put in a bright gold frame. Sahira hoped that, even if Saleh didn't recognize her, he'd at least recognize his old self. Sahira slouched down in the seat of her car, pulled her sleeves to cover her cold fingertips, and dozed off. She woke up when her car began rocking as people squeezed past it in the mayhem. The crowds made her car move with them.

At first, Saleh didn't believe it when the prison doors flung open and the guards yelled at them in Farsi to get out. Saleh though, as his cellmates did, that they would be executed.

S2: Background Information on Signal Detection Theory

Signal detection theory is a statistical method that judges the presence of a signal based on the reaction/response pattern of hits and false alarms (Swets, 1965). Hits and false alarms were computed in a manner to create a d' score for each text level comprehension by using recognition test-items when examining memory which has been done in previous studies as well (Swets, 1965). A d' score is a sensitivity measure that reveals the extent to which a signal is present, with high scores near and above 1.0 indicating that the signal is present, and low scores near 0.0 indicating that the signal is absent. In this case, the d' score is a measure of comprehension for each unique level of text-understanding in Kintsch's Construction-Integration Model (Kintsch 1994). Furthermore, this application of signal detection when used in conjunction with recall has been applied in previous studies to ascertain the strength of each text comprehension level in the Construction-Integration Model, such as in one study where participants were told to either read a text for knowledge acquisition or for summarization (Schmalhofer, F., & Glavanov, D., (1986)).

To calculate the d' score, the traditional formula is as follows: d' = Z(HR)-Z(FA). The FA is the false alarm rate, which is the rate at which people marked a signal present when it was in fact, not present. The HR is the hit rate, which is the rate at which people correctly determined that a signal was present when it was present. The Z scores are taken for each rate and the resulting difference is the d' value, which indicates the strength of comprehension for a particular discourse comprehension level (surface, textbase, situational). We will discuss the problems with this approach below.

As the levels are somewhat intertwined with each other to provide a complete mental representation of the text, determination of a d' value for each distinct level requires a careful method of analysis. Thus, similarly to traditional signal detection theory, responses to test items are examined, however, in this particular case, a response can count as either a "Hit" or a "False Alarm", depending on which level of text comprehension is being examined (Kintsch, 1990; Zwaan, 1994). To calculate a d' score for the surface level text comprehension, Zwaan's FA was the rate at which people stated that a paraphrase response was a verbatim response. (Note: scores of 4,5,6 were scored as statements affirming that the sentence was taken verbatim from the original text while scores of 1,2,3 were interpreted as meaning that the phrase in question was not taken verbatim from the text). The HR was the rate at which people determined that a verbatim phrase in the questionnaire was indeed verbatim. These responses were found by counting how many people responded "Yes, this phrase is verbatim" by answering either 4,5,6. That means, Zwaan 1994 did not make a difference between the levels of confidence between 4, 5 or 6 and 1, 2, and 3; the differences were collapsed as either yes or no statements. (We will add a weighted analysis below).

According to both Kintsch and Zwaan, the main idea is that affirmative responses (as in yes, this is a verbatim statement) to certain phrase categories (verbatim, paraphrase, correct inference, incorrect inference), will register as either a hit or a false alarm depending on which level of text comprehension is being investigated. To analyze a participant's surface level of comprehension, the crux of this matter is the participant's ability to remember and note whether general gist/meaning of a sentence appeared in the original passage or whether the phrase, word-for-word, appeared in the original text. To analyze a participant's textbase level of comprehension, this involves the difference between the participant remembering the meaning of the text and the explicit statements/actions versus the participant drawing conclusions about the text which are not explicitly stated. Finally, to analyze a participant's situational level, the participant must identify the difference between an accepted, plausible inference and a wild, far-fetched inference. For example – in one text, a woman shoots and kills her husband, who is a high-ranking government official. One can generally infer that the woman did not like her husband due to marital issues. It is much less plausible that the woman is a part of a secret organization whose main mission is to bring down the tyrannical government.

S3. McNicol's Approach to Signal Detection Theory with Example

The combination of scores is derived from the method presented by McNicol (2005), in which categories, which designate degree of confidence on a confidence scale, of either signal (which form the Hit Rate) or noise responses (which form the False Alarm rate) are summed to prevent a Z-score value from reaching infinity. Thus, there were four resulting scores, one for each type of sentence (verbatim, paraphrase, plausible inference, implausible inference). These scores were then divided by 8 as opposed to 6 to eliminate the possibility of a person having hit rate or a false alarm rate of 0% or 100%. This action was also inspired by McNicol's approach to the utilization of Signal Detection Theory because a Z-score response of 0.0 or 1.0 is once again prevented. In addition, McNicol also incorporates the variability in confidence responses by weighing each probability rate (Hit Rate or False Alarm) through systematically summing all responses recorded to a signal/noise in a particular confidence category or stricter (McNicol, 2005, Chapter 5). This results in probability rates which are weighted according to the number of participants who select a particular confidence value on the scale. Moreover, by dividing all the participant-given value by 8, the rates more accurately reflected the degree of confidence of a participant rather than the "all or nothing" approach of Zwaan's application of Signal Detection Theory. An example value is shown below.

Assume a participant scored the recognition sentence test items as follows:

	Phrase Type					
Participants	Verbatim 1	Verbatim 2	Paraphrase 1	Paraphrase 2	Plausible	Implausible
1	6	6	5	4	2	1
2	5	6	4	2	3	2

The averaged scores would therefore be:

	Phrase Type			
Participants	Avg. Verbatim	Avg. Para	Plausible	Implausible
1	6	4.5	5	4
2	5.5	3	4	2

Next, the scores were weighted by dividing by 8 to acquire rates which could be used as either false alarm rates or hit rates, depending on the specific comprehension level, when acquiring d' values. This differs from traditional signal detection theory which would designate values 4 through 6 as being considered 100% confidence and thus a "hit" while designating values 1 through 3 as being "misses", without taking into account the discrepancies in confidence between a score of 4 and a score of 3. In this modified, weighted method, all values are taken into consideration to provide a more comprehensive portrait of the cognitive representations of text.

	Phrase Type			
Participants	Verbatim	Avg. Para	Plausible	Implausible
1	0.75	0.563	0.625	0.5
2	0.688	0.375	0.5	0.25

Then, the formulas referenced earlier were applied.

Surface Level Comprehension

d’ = Z(HR)-Z(FA) = Z(Verbatim as Verbatim) – Z(Paraphrase as Verbatim)

Textbase Level Comprehension

d’ = Z(HR)-Z(FA) = Z(Paraphrase as Verbatim) – Z(Correct Inference as Verbatim)

Situational Level Comprehension

d’ = Z(HR)-Z(FA) = Z(Correct Inference as Verbatim) – Z(Incorrect Inference as Verbatim)

Thus, the participants 1 and 2 would have the following d’ scores which reflect the strength of each text comprehension level:

Surface Level Comprehension

Participant 1 : d’= Z(HR)-Z(FA) = Z(0.75) – Z(0.563) = 0.516
Participant 2 : d’= Z(HR)-Z(FA) = Z(0.688)-Z(0.375) = 0.809

Textbase Level Comprehension

Participant 1 : d’ = Z(HR)-Z(FA) = Z(0.563) – Z(0.625) = 0.16
Participant 2 : d’ = Z(HR)-Z(FA) = Z(0.375)-Z(0.5)= 0.319

Situational Level Comprehension

Participant 1 : d’ = Z(HR)-Z(FA) = Z(0.625) – Z(0.5) = 0.319
Participant 2 : d’ = Z(HR)-Z(FA) = Z(0.5)-Z(0.25) = 0.674

Appendix S4. Study 1 Calculated d’ Scores

Levels of Representation Study 1

Surface	Textbase	Situational Model

	Unweighted	*Weighted*	*Unweighted*	*Weighted*	*Unweighted*	*Weighted*
News						
Mean	0.917	0.328	1.532	0.528	0.556	0.160
SD	0.093	0.033	0.117	0.034	0.109	0.027
Fiction						
Mean	0.805	0.270	1.243	0.450	0.480	0.156
SD	0.093	0.033	0.117	0.034	0.108	0.027

Table 1. Study 1 Unweighted and Weighted d’ Scores for Levels of Text Representation

Appendix S5. Study 2 Calculated d’ Scores

Table 2. Study 2 Unweighted and Weighted d’ Scores for Levels of Text Representation

	Levels of Representation Study 2					
	Surface		**Textbase**		**Situational Model**	
	Unweighted	*Weighted*	*Unweighted*	*Weighted*	*Unweighted*	*Weighted*
News						
Mean	0.934	0.288	2.099	0.690	0.503	0.157
SD	0.134	0.035	0.172	0.038	0.175	0.036
Fiction						
Mean	0.799	0.268	1.656	0.575	0.859	0.212
SD	0.141	0.035	0.181	0.039	0.184	0.036