1
Prose Comprehension Beyond the Page

Jennifer Wiley and Tricia A. Guerrero

A Brief Reflection of Emergence of Research on Discourse Comprehension

Over 35 years ago, *Prose Comprehension Beyond the Word* was published which explored the processes involved in constructing meaning across units of text that extended beyond single words and sentences (Graesser, 1981). A brief glimpse at some of Graesser's most impactful publications since that first monograph helps to illustrate how the field has moved toward the study of comprehension *beyond the page* by exploring how readers construct inferences in larger prose and discourse contexts (Graesser, Singer, & Trabasso, 1994), studying how learners interact with tutors and intelligent tutors (Franklin & Graesser, 1996; Graesser & Person, 1994), and promoting the development of tools (including Coh-Metrix) that can support automatic detection of readers' comprehension and emotional states (Graesser, McNamara, Louwerse, & Cai, 2004). As a young graduate student starting in the lab of James F. Voss, the first author read the 1981 book as she began her first research projects on the effects of prior knowledge and attitudes on text processing. Rereading it again recently along with the second author, it is clear that this work had a formative influence.

One salient theme that emerges from the opening chapter is the need for a new generation of research that considers how the meaning of prose (rather than word or sentence-level meaning) is constructed. The work presented in this early volume heralds the burgeoning of research exploring comprehension at a discourse level, with contemporary work from the expertise tradition on the effects of domain knowledge on prose comprehension (Chiesi, Spilich, & Voss, 1979; Spilich, Vesonder, Chiesi, & Voss, 1979), from the schema tradition with work using larger units of text (Anderson, Reynolds, Schallert, & Goetz, 1977), and from Garrod and Sanford who published a book in the same year (1981) with a very similar name, *Understanding Written Language: Explorations of Comprehension Beyond the Sentence*, which emphasized the need to take the context of language into account.

A second salient theme that immediately appeals to an inquisitive researcher's curiosity is the stark discrepancy painted throughout the monograph between the comprehension of narrative and expository prose. Readers prefer narratives, remember them better, and generate more inferences from them than from expository texts. These findings are at once intriguing and highly problematic for researchers who are interested in educational applications. In many educational contexts, students are routinely expected to gain an understanding of new concepts from expository texts, particularly in science. What is so different and difficult about comprehension of expository text? This is a central question that remains to this day. A later volume on *The Psychology of Science Text Comprehension* edited by Graesser and colleagues (Otero, Leon, & Graesser, 2002) provided some updates on this line of questioning, but there is still much more yet to discover about how and when learners can learn effectively from reading, and why they often fail to develop an understanding of new ideas from studying expository text.

There are surely many reasons why expository texts may cause comprehension difficulties. Texts vary in their structure and their complexity. And, readers vary in their familiarity with the subject matter, and whether they possess the prior knowledge needed to understand it or the working memory capacity that may be required to process the information. Narratives are generally about people and their actions and motivations over time. As readers, we generally have an abundance of prior knowledge about such things. In contrast, when we are given expository texts to read, it is often to teach us about phenomena, processes, or systems that we do not understand yet. This fundamental difference may be critical, or it may be just one of many reasons why comprehension from expository text is so challenging.

Answering the question of why expository text is so hard to comprehend, and under which contexts or conditions learning from text may be most successful, has been the motivation behind many studies done by Wiley and her colleagues. The general approach has been to identify when students experience obstacles to learning from expository texts; exploring which individual differences or conditions may relate to better or poorer performance; and testing hypothetical mechanisms by manipulating the amount of support that students receive through instructional prompts, learning activities, or instructional adjuncts such as animations and analogies. The ultimate goal of this work is to understand what features are critical for the design of effective instructional contexts so that students may acquire a deep understanding of subject matter as they read.

How Do We Distinguish “Deep” From “Shallow” Comprehension In Our Research, Both Theoretically and Empirically?

A key feature of the research that has been done by Wiley and her colleagues that resonates with Graesser’s work is the critical distinction between learning that is based in memory versus understanding of text, or alternatively between “shallow” and "deep" comprehension processes. Otero et al. (2002, p. 6) provide this helpful characterization:

> Shallow knowledge consists of explicitly mentioned ideas in a text that refers to: lists of concepts, a handful of simple facts or properties of each concept, simple definitions of key terms, and major steps in a procedure… . Deep knowledge consists of coherent explanations of the material that fortify the learner for generating inferences, solving problems, making decisions, integrating ideas, synthesizing new ideas, decomposing ideas into subparts, forecasting future occurrences in a system, and applying knowledge to practical situations.

This distinction between shallow and deep comprehension is key when studying learning from expository science texts. The goal for reading is not for students to merely have a superficial representation of the exact words or sentences that were read. Rather, the goal is for them to develop a coherent situation model (Kintsch, 1994), causal model (Graesser & Bertus, 1998; Millis & Graesser, 1994; Singer, Harkness, & Stewart, 1997; Wiley & Myers, 2003), or mental model (Mayer, 1989) of a system or phenomenon. This generally requires generating connections not explicitly mentioned in the text, by integrating information across different parts of text or with prior knowledge. It can also involve reasoning through ideas stated in the text to arrive at implications or consequences.

Different Conceptions of Comprehension

The theoretical differences between these two types of learning outcomes (memory and understanding, or shallow vs. deep comprehension) lead to differences in measures of comprehension. Memory for a text requires shallow, surface-level processing. Typical memory-based questions will prompt the reader to recall ideas mentioned directly in the text such as facts or definitions, or to interpret the meaning of particular words or phrases. On the other hand, understanding of text requires a deeper level of processing. Questions that test for understanding often require the reader to think about "how" and "why," and when given after reading, such questions can assess "deep learning" or "deep comprehension" of a system, process, phenomenon, or concept.

There are many ways of assessing comprehension of text, and test questions vary in the "depth" of the comprehension that is required. A consideration of two standardized comprehension tests (Nelson-Denny, ND, and Gates-MacGinitie, GM, reading tests) helps to provide examples on a continuum of possible question types. The college-level forms of the ND (Forms G and H) contain both narrative and expository passages that are about 200 words in length. The scoring manual categorizes questions into two categories: literal and interpretive. The literal items are "text-based" questions that largely assess surface-level features of the text. A large proportion of these questions can be found verbatim in the text, and answered by using lexical overlap or simple search. Interpretive questions usually require a reader to assess the author's point of view or the appropriate audience for the text. Because of the explicit relations between the questions and the texts, along with the minimal inferencing needed for items of either type, the ND can be categorized as requiring "shallow" comprehension abilities to arrive at correct responses.

The highest-level version (10/12) of the GM also contains both narrative and expository passages. As shown in Table 1.1, the GM expository passages tend to be shorter and more variable in length than the ND, yet comparable in difficulty. Although not defined by the GM, questions can also be categorized similarly to the ND. Literal questions require shallow processing, although they have more syntactic and lexical variability. They also require slightly more interpretation than the ND because the distractor choices require that the reader develops a basic understanding of the passage in order to select the correct answer, as opposed to conducting a verbatim search for words associated with the question. The non-literal questions are variable in the level of processing required. Some are lower-level inference questions that require minor connective or bridging inferences across sentences to arrive at the answer. There are also higher-level inference questions that require the reader to reason with information from the text, integrate it with prior knowledge, or apply it in new contexts.

How do we measure "deep" comprehension? As shown in Table 1.1, whereas the standardized tests provide strong coverage in "shallow" comprehension processes, they provide less coverage of "deep" comprehension processes. Few questions require the reader to use the information that they read as part of a reasoning process or as applied to a new situation. However, there are some standardized tests that do have more items that require readers to think about inferences or implications that follow from the text including both the ACT Reading Comprehension subtest and the MCAT Critical Reasoning section. For example, the ACT Reading Comprehension subtest includes questions that ask readers to use reasoning skills to understand sequences of events; make comparisons; comprehend cause-effect relationships; and draw generalizations. The passages are representative of the kinds of text commonly encountered in first-year college curricula. And, there is now even a section that requires reasoning across multiple texts.

Table 1.1 Descriptive Statistics for Features of Different Comprehension Tests

	Nelson-Denny	*Gates-MacGinitie*	*Griffin, Wiley & Thiede (in press)*
FKGL (Texts)	11.7 (3.2)	12.1 (2.0)	11.16 (1.24)
FRES (Texts)	44.6 (16.0)	48.7 (9.4)	48.02 (7.54)
Word count (Texts)	203.4 (22.8)	123.9 (43.0)	799 (198)
Text-based test items	63.16%	22.92%	50%
Low-level inference items	36.84%	66.67%	25%
High-level inference	0%	10.42%	25%

items

Note: FKGL = Flesch-Kincaid Grade Level, FRES = Flesch Reading Ease Score

The materials from Griffin, Wiley, and Thiede (in press) (and Thiede, Wiley, & Griffin, 2011) serve as examples of attempts to assess comprehension at both surface and deep levels. These materials include a set of six expository science texts that describe phenomena, and support construction of a causal understanding of the process or system (see Wiley, Griffin, & Thiede, 2005 and 2008 for more discussion). In contrast to memory-based test items, the inference questions used by Wiley and her colleagues primarily test for understanding of a causal or mental model of phenomena. These questions ask readers to apply knowledge they gained from the text to new situations, analyze causal relationships and relationships between ideas and implications, and predict possible consequences when factors are changed. In order to support the construction of a mental model from the text, and to allow for multiple different test questions to be created, the texts are typically much longer than ND and GM passages, while being comparable in difficulty as shown in Table 1.1. A further contrast with traditional assessments regards whether texts are available during testing. While reading comprehension assessments traditionally allow readers to answer with the texts available, when the goal of research is to measure what a student has *learned* from a passage, then it becomes important to test what understanding *remains* after the text is no longer present.

A final point is that the nature of questions that can be asked on a comprehension test is inextricably linked to the quality of the text that the reader is asked to read. If the text does not include the information needed to construct a causal model of a phenomenon, then it makes no sense to include comprehension items based on causal inferences. The expository texts that are provided to learners need to be carefully written. Test items also need to be carefully created so that they require that students have constructed a coherent situation-model representation of the text in order to answer questions correctly. To ensure test items are measuring understanding, they cannot be answerable simply by using a surface-level representation of the text. To ensure that readers need to engage in active comprehension, the texts also cannot be fully explicit. That is, some key connections needed to be inferred by the reader. If the texts are fully explicit with respect to a causal model, then students could potentially answer inference questions by using verbatim memory for the text.

To explore differences between memory for text and understanding (or shallow and deep comprehension), two distinct sets of test items (memory and inference) were developed for these six texts. The distinction between these two types of test items can be empirically demonstrated in a number of ways. For example, in one study we found that undergraduates are able to reliably differentiate between memory and inference questions by recognizing when answers could be "found directly in the text" (memory questions) as opposed to answers that could not, and need to be inferred from the text (inference questions). A different study asked participants to answer both types of test questions with the texts present. Performance on the memory questions in this condition was at ceiling, whereas performance on the inference questions did not improve. However, a third study found that performance on the inference questions did improve when readers were prompted to generate an explanation of how or why a phenomenon is occurring before taking the tests (Guerrero & Wiley, 2018).

Further results discussed later provide additional support for the distinction between the two types of test questions, as several manipulations have produced dissociable effects. One line of work explores conditions that underlie successful understanding of (as opposed to memory for) explanatory, expository science texts. Two other lines of work extend in two different directions: across multiple documents (when readers attempt to acquire new understanding from multiple sources), and into the reader's mind (when readers need to assess their levels of comprehension). All three areas of work show the critical importance of the reader adopting an appropriate task or activity model that supports reading for understanding, as well the importance of engaging in constructive processing at the level of the situation model to achieve deep comprehension.

What Does Your Research Indicate About the Factors That Enhance Comprehension or Its Assessment?

Obstacles to Deeper Comprehension of Expository Science Texts

Instructional conditions that improve memory for text are not necessarily the same as those that improve understanding of text (Kintsch, 1994). So, to the extent that one cares about deep comprehension, it is important to include measures of learning that represent understanding, and not just surface or verbatim memory for what was read. Further, it is measures of deep comprehension that are the better predictors of whether readers will be able to apply their learning from the texts in novel contexts (Mayer, 1989, Wiley & Voss, 1999). As summarized by Otero, et al. (2002, p. 6), it is deep comprehension from expository science texts that is "essential for handling challenges and obstacles because there is a need to understand how mechanisms work and to generate and implement novel plans."

One main line of inquiry has explored how to support readers in the generation of coherent situation models as they learn from complex expository science texts. This work builds on the theoretical importance of causal or mental models as the representation that best captures the deeper meaning of explanatory science texts. Several studies have involved providing students with animations or analogies that can support the construction of mental models. Sanchez and Wiley (2014) showed that students given animations alongside text wrote better explanation-based essays and performed better on inference tests about volcanic eruptions. This was especially important for students who had low-spatial skills. Similarly, Jaeger, Taylor and Wiley (2016) found that including an analogical example in a text about El Niño led to better explanation essays and inference test performance. Again, this manipulation was particularly important for low-spatial students. In another line of work, Hinze, Wiley, and Pellegrino (2013) explored the role of different types of study and testing activities on learning. When readers were given practice recall tests, it improved their performance on memory-for-details questions. On the other hand, when practice tests involved the generation of an explanation, it improved performance on inference tests. These studies suggest that many readers struggle with generating appropriate mental models of scientific phenomena. They do not generally engage with expository science texts at a level of deeper comprehension, and need support in engaging in the kinds of processes that can support the construction of coherent situation models.

Effective Multiple Source Comprehension

Engaging in an inquiry task with multiple sources is a complex activity because it requires all the processes necessary for comprehending individual expository texts, plus an additional set of processes that become particularly important when readers are confronted with information from more than one source. Comprehending phenomena from multiple sources means that there needs to be a level of representation that reflects the understanding derived from integrating information across sources (an *integrated model*). In a multiple source context, it is this level that best represents a reader's understanding of the content, or the mental model that has been constructed about the causes of the phenomenon that is the focus of inquiry (Wiley, Jaeger, & Griffin, 2018).

As with single-text studies, materials used for multiple-source-comprehension studies have been designed so that important connections are not explicitly made for the reader. In addition, the reader must have to re-purpose the texts in order to answer the inquiry question (i.e., there should be no verbatim sentences that can be copied as a response). The main results emerging from this research suggest that comprehension is critically influenced by the reader's interpretation of the task and the processing they should engage in to achieve it. As defined by the MD-Trace framework (Rouet & Britt, 2014), the *task model* includes the reader's goals and subgoals. In addition, the *task model* subsumes an *activity model*that informs the reader about which specific actions, procedures, or behaviors one might engage in to reach those goals during the task (Wiley, Jaeger, & Griffin, 2018). The extent to which readers understand that they need to actively engage in constructive processing, attempt to build connections across ideas, and try to form a coherent, integrated model of the phenomena, as well as the extent to which they engage in those activities, are critical determinants of learning from multiple sources as measured by both comprehension tests and analyses of the quality of their inquiry task essays.

Multiple measures of essay quality have been shown to correlate with inference test performance (Wiley & Voss, 1999). Students who demonstrate better understanding of the material on inference tests also tend to write essays that have more connected ideas and more causal connections. Further, the conditions that led to the best memory for the text were not the same conditions that led to the best understanding of the text as measured by either inference test performance or essay quality.

Another measure considers the origin of information included in the essay, and the extent to which students plagiarize or copy information from the original texts. Students who demonstrate better understanding of the material on inference tests tend to write essays that contain a lower proportion of borrowed or copied sentences. Similarly, using automated plagiarism detection techniques, negative correlations can be found between plagiarism scores and explanation quality (Wiley, Hastings et al., 2017). These results suggest that both comprehension and the quality of the inquiry product suffer as students fail to transform information and fail to engage in constructive activity as they attempt to learn from multiple documents. The extent to which the student essay is responsive to the inquiry prompt and re-purposes text from the documents, rather than using segments of text as written for their original purpose, also predicts both the quality of essays as well as inference test scores (Wiley, Hastings et al., 2017).

Students often tend to engage in passive, superficial representation of the information they read, rather than actively selecting and transforming information to construct a coherent, integrated mental model of the phenomenon (Kintsch, 1994). This tendency is the primary issue that instructional manipulations have been designed to address. Studies that have manipulated the features of the inquiry task that is given have supported the theoretical importance of the *task model*, and have highlighted that a learner's *task model* can be affected by the wording of the inquiry prompt (Wiley et al., 2009; Wiley & Voss, 1999). Studies that have manipulated features of the task environment, such as changing the type of documents included in a set, have also highlighted how sensitive the *task model* is to subtle changes (Blaum et al., 2017; Wiley, Ash, Sanchez, & Jaeger, 2011). This suggests that learners may often lack a well-developed *activity model* that would guide them to consciously implement basic subgoals during inquiry tasks. Instead, learners may depend on contextual cues that make different subgoals salient.

Multiple-document inquiry tasks provide an opportunity to help students to gain a richer understanding of content and to develop skills that are important for seeking, selecting, evaluating, re-purposing, and integrating information from various sources (Graesser et al., 2007). Such skills are vital to lifelong learning outside of laboratories and classrooms. This research suggests that students need help developing these skills and need to learn to engage in constructive and "deep" comprehension behaviors.

Effective Metacomprehension of Expository Science Texts

A third line of research on comprehension has explored how accurate (or inaccurate) students are at predicting their own understanding after studying a set of texts (i.e., their metacomprehension accuracy). Most work in this tradition has utilized a judgment paradigm that involves participants reading a series of expository texts, providing predictive judgments of comprehension for each text, and answering test questions for each text. In general, most work finds that students' ability to accurately predict their own level of comprehension for each text is quite poor (Griffin, Mielicki, & Wiley, in press).

In contrast to poor accuracy in baseline conditions, a substantial body of evidence emerging from studies inspired by the *situation-model approach to metacomprehension* (Wiley, Thiede, & Griffin, 2016) indicates that there are a number of conditions that can lead to significant improvements. Several studies have demonstrated that conditions that prompt readers to rely on situation-model-based cues for judging their comprehension (such as delayed keyword, delayed summarization, and self-explanation tasks) can improve their accuracy when criterial tests also assess the situation model.

The results of other studies using manipulations that specify goals for reading and prompt readers to expect inference-level tests suggest that it is important for students to have a clear understanding of what comprehension means. Thiede et al. (2011) found improved metacomprehension accuracy when graduate students were provided with an explicit statement about their goals for reading as being reading for understanding, an explicit statement about the kind of test items they should expect, and example test items for a practice text (on a different topic than the target texts). Griffin et al. (in press) found similar improvements among a sample of undergraduates. Participants in the inference-test-expectancy condition were better able to predict performance on inference tests (as opposed to memory tests) over no-expectancy and memory-expectancy conditions. Conversely, no-expectancy and memory-expectancy conditions showed more accurate prediction of memory test performance. Further, when a self-explanation activity prompt was combined with inference test expectancy, it led to the highest levels of metacomprehension accuracy (i.e., the most accurate predictions of relative performance across topics on inference tests). It is encouraging that similar benefits have also been found for this combined manipulation in a real classroom context (Wiley, Griffin et al., 2016). Further, Thiede et al. (2012) found that middle-school students whose early literacy education focused on deep understanding and experience with inference tests demonstrated better metacomprehension accuracy than students with more typical schooling experience.

These results speak to the importance of helping students to appreciate that their goal for reading expository science texts is to understand how or why a phenomenon or process occurs, and that they will need to make connections and causal inferences across sentences in order to correctly answer test questions. Students also need to appreciate the difference between what it means to remember a text versus what it means to understand a text. They need to keep this distinction in mind while reading, and use it to help them make better judgments of whether they have actually understood a text. Improving the ability to judge deep comprehension is important because it will impact how effective learners can be when they make decisions about what to re-study.

Conclusions

The results from these lines of work echo several themes introduced in Graesser's (1981) *Prose Comprehension Beyond the Word.* Deep comprehension from expository text is not automatic, not effortless. Complex processes are required to build deep comprehension. The deeper meaning is not obvious. It requires engaging in active construction of meaning and reasoning processes. Readers benefit from support in constructing causal or mental models from expository science texts. Readers benefit from an appropriate task model, goals that support reading for understanding, inference-based test expectancies, and activities that support constructive processing.

At the time of the 1981 monograph, text-processing research was largely using reading times on sentences or response times to probes as the main dependent measures. However, reading time and reaction time data (alone) are problematic to interpret on their own. Much has been learned by adding other methods to the researchers' arsenal including having readers engage in think-aloud protocols, providing responses to how-and-why questions, or generating explanations or other products that can be analyzed to better understand the reader's mental representation of what they read. To determine the contexts that can support successful comprehension from text, we need measures that assess understanding and not just memory for text. We also are in the best position to make recommendations for educational purposes when we use converging measures that provide information about the processes that the reader engages in, as well as the depth of learning that results from those processes.

Finally, the Graesser (1981, p. 63) monograph also highlighted the need for work that considers the role of text processing in authentic, real-world contexts: "Psychologists have generally lost sight of self-induced acquisition. We know next to nothing about how adults acquire information in daily life." As we have moved text-processing research into exploring how readers use internet sources and how they engage in self-regulated learning, we have started on some initial steps toward the spirit of this vision.

Acknowledgements

The authors were supported by grants from the National Science Foundation (DUE 1535299) and the Institute of Education Sciences (R305A160008) during the preparation of this chapter. All opinions expressed herein are those of the authors and do not necessarily reflect those of the funding agencies.

References

Anderson, R. C., Reynolds, R. E., Schallert, D. L., & Goetz, E. T. (1977). Frameworks for comprehending discourse. *American Educational Research Journal*, *14*, 367–381.

Blaum, D., Griffin, T. D., Wiley, J., & Britt, M. A. (2017). Thinking about global warming: Effect of policy-related documents and prompts on learning about causes of climate change. *Discourse Processes*, *54*, 303–316.

Chiesi, H. L., Spilich, G. J., & Voss, J. F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, *18*, 257–273.

Franklin, S., & Graesser, A. (1996). Is it an agent, or just a program? A taxonomy for autonomous agents. In *International workshop on agent theories, architectures, and languages* (pp. 21–35). Berlin, Heidelberg: Springer.

Graesser, A. C. (1981). *Prose comprehension beyond the word.* New York: Springer-Verlag.

Graesser, A. C., & Bertus, E. L. (1998). The construction of causal inferences while reading expository texts on science and technology. *Scientific Studies of Reading*, *2*, 247–269.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, *36*, 193–202.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, *31*, 104–137.

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

Graesser, A. C., Wiley, J., Goldman, S. R., O'Reilly, T., Jeon, M., & McDaniel, B. (2007). SEEK Web tutor: Fostering a critical stance while exploring the causes of volcanic eruption. *Metacognition and Learning*, *2*, 89–105.

Griffin, T. D., Mielicki, M. K., & Wiley, J. (in press). Improving students' metacomprehension accuracy. To appear in J. Dunlosky & K. Rawson (Eds.), *Cambridge handbook of cognition and education.* New York: Cambridge University Press.

Griffin, T. D., Wiley, J., & Thiede, K. W. (in press). The effects of comprehension-test expectancies on metacomprehension accuracy. *Journal of Experimental*