Methods of Studying Text: Memory, Comprehension, and Learning

Kathryn S. McCarthy, Kristopher J. Kopp, Laura K. Allen, & Danielle S. McNamara

Institute for the Science of Teaching and Learning

Arizona State University

Methods of Studying Text: Memory, Comprehension, and Learning

Reading is an everyday activity that is important for our academic and workplace success, as well as engagement with our surrounding community. Despite this fact, the ability to deeply comprehend the texts that we read is a challenging skill that depends on complex interactions among our cognitive processes and prior knowledge (Gernsbacher, 1997; Kintsch, 1988, 1998; van den Broek, Young, Tzeng, & Linderholm, 1999). We must be able to decode the letters and words in a text, but also access these lexical elements from memory, process the syntax, generate inferences within the text, and integrate this information with prior knowledge (Magliano, Millis, Ozuru, & McNamara, 2007; McNamara & Magliano, 2009).

Importantly, research on the reading process is vast, spanning work across multiple domains. The field of reading research focuses on identifying and explaining lower-level processes involved in language understanding, such as word recognition, lexical decoding, or syntactic parsing (see Kamil, Pearson, Moje, & Afflerbach, 2011), whereas the field of discourse tends to focus on the higher-level cognitive processes associated with comprehending the underlying meaning of text and discourse (see Graesser, Gernsbacher, & Goldman, 2003; McNamara & Magliano, 2009; Schober, Britt, & Rapp, in press). An important component of this discourse comprehension research is the examination of the reader’s mental representation or mental model of the text, which includes information explicitly included in the text, similar information from the reader’s prior knowledge, and inferences generated during the reading process.
As these mental representations cannot be accessed directly, researchers must infer their structure and content based on behaviors and performance on comprehension tasks. To do so, researchers assess readers' memory for the information in the text, the inferences they generate (i.e., to establish connections between text information as well as between text information and information from prior knowledge), and their ability to learn from text and use this information in new situations (Goldman, 2004; Graesser, Gernsbacher, & Goldman, 1997; Kintsch, 1998; Magliano & Goldman, 1991; Magliano & Millis, 2003).

Though memory, comprehension, and learning are deeply intertwined, it is important to make distinctions between these constructs. Take for example, a student who is able to recognize or recall the definition of a particular term. This student has demonstrated memory for this information. However, even if the student could recite the definition of the term verbatim, they might have little to no understanding of what the definition means, thus, lacking comprehension. Likewise, the student may demonstrate an understanding of the concepts immediately after reading a text, but show poor memory after a delay or lack the ability to use information about this term to solve a new problem. The ability to do so would require learning from the text. In this chapter, we outline various methodologies that have been used to address research questions related to text memory, comprehension, and learning. Additionally, we discuss innovative methods that are currently being developed to address these questions in new ways.
THEORIES OF COMPREHENSION

The majority of contemporary models of text comprehension share a common goal of understanding the cognitive processes and knowledge that shape readers’ mental representations of texts (e.g., Gernsbacher, 1997; Kintsch, 1988, 1998; van den Broek et al., 1999; see McNamara & Magliano, 2009, for a review of contemporary models of comprehension). Though these models differ in their details and areas of emphasis, they share a number of common elements. Perhaps most important is the notion that readers construct representations of texts at multiple levels (Graesser, Millis, & Zwaan, 1997; Kintsch, 1998), and that successful comprehension relies on the reader’s ability to establish coherence within and among these levels (Gernsbacher, 1997; Zwaan & Radvansky, 1998).

Coherence of the mental representation is achieved by establishing meaningful connections among text information through the generation of inferences (Kintsch, 1993; Magliano, Trabasso, & Graesser, 1999). There are a number of different kinds of inferences that can be generated by readers. Whereas some inferences are made quickly and automatically (McKoon & Ratcliff, 1992; Myers & O’Brien, 1998), others are made strategically (Graesser, Singer, & Trabasso, 1994; van den Broek et al., 1999). These inferences can connect information across two proximal sentences (local inferences), or they can link information from disparate parts of the text, across multiple texts, or to prior knowledge (global inferences).

The most influential and pervasive model of text comprehension is the Construction-Integration model (CI model; Kintsch, 1988; 1998; van Dijk & Kintsch, 1983). This model argues that the mental representation consists of a surface code,
textbase, and situation model. These are not separate representations of the text, but rather different levels within a single mental representation. The surface code is the most basic level and reflects the verbatim wording and structure of the text. This level fades quickly, but in the time it is available, the reader uses it to identify semantic and syntactic relationships. This process helps in the construction of the textbase, which preserves the key pieces of information that are necessary to represent the “gist” of the text. Finally, the situation model builds upon the textbase with information from the reader’s long-term memory to construct a more elaborate representation of the text’s meaning. This level contains all inferences generated to establish connections among ideas in the text and connections to prior knowledge.

Beyond the levels outlined by the CI model, researchers have posited additional levels of representation to account for contextual information that can contribute to comprehension. The communicative (or pragmatic-communicative) level reflects an understanding of the author’s message or the text’s purpose (Graesser & McNamara, 2011; Graesser, Millis, & Zwann, 1997; Magliano, Baggett, & Graesser, 1996; McCarthy & Goldman, 2015). The task model level recognizes reading as a goal-oriented task and encompasses the reader’s goal or purpose (Doane, Sohn, McNamara, & Adams, 2000; Kendeou, Bohn-Gettler, & Fulton, 2011; McCrudden & Schraw, 2007). Finally, the documents model level was developed to capture information related to the processes involved in the comprehension of multiple documents, wherein readers must gather, integrate, and evaluate the validity of information across different sources (Perfetti, Rouet, & Britt,, 1999).
The components of these levels have important implications for the assessment of text comprehension. Broadly, the goal of these assessments should not be to measure an individual's memory for facts from the text, but rather to capture the state of their mental representation. Researchers must garner information related to the reader's prior knowledge of the text domain, as well as the inferences that are generated during reading (Johnson-Laird, 1980; McNamara & Magliano, 2009). In the remainder of the chapter, we describe various methods that have been used to assess text comprehension processes across multiple levels of representation.

**MEASURES OF TEXT COMPREHENSION**

Text comprehension measures can be broadly grouped into two categories: those employed *during* reading (i.e., processing or “on-line” measures) and *after* reading (i.e., post-reading or “off-line” measures). Processing is commonly assessed using lexical decision tasks, think-aloud protocols, reading time measures, eye tracking, and neural measures such as fMRI and ERP. Common post-reading measures include sentence and inference verification tasks, constructed responses, free or cued recall, and essays. These measures are described in detail below. It should be noted that although we discuss these measures independently, researchers often employ a variety of processing and post-reading measures as a means of capturing multiple facets of the mental representation and to have converging evidence of the underlying comprehension processes.

**Post-Reading Measures**

Text comprehension is most commonly assessed using post-reading measures, which occur off-line and assess the mental representation after reading. Within
assessments that occur immediately following reading, differences in readers’ representations are assumed to reflect differences in processing. However, post-reading measures should not be considered merely a substitute for processing measures when processing measures are not available. They can be used to investigate how readers use the information in their representation or, when administered at a delay, to explore long-term retention as well as how the mental representation might develop or change over time. Given the rapid decay of information in the surface code level, immediate assessments are more likely to tap into a reader’s surface code whereas delayed assessments are more representative of a reader’s textbase or situation model understanding (Kintsch, Welsch, Schmalhoffer, & Zimny, 1990).

Post-reading assessments can require readers to determine the veracity of an answer, to identify the answer from a set of given responses, or to construct the answer on their own. Importantly, these different types of questions rely on different aspects of the comprehension process. Identification questions measure more passive processing (and, thus, rely more heavily on prior knowledge) and open-ended, constructed responses require more active comprehension processes (Ozuru, Kurby, Briner, & McNamara, 2013).

*Sentence Verification and Inference Verification Tasks*

Verification tasks ask readers to determine if a statement is true or false or if the statement appeared in the text. In the original version of the sentence verification task, readers were shown a picture of simple shapes (e.g., star and plus sign). The readers then determined if a sentence about the image (e.g., *The star is above the plus*) was
true or false (Clark & Chase, 1972). This task was designed as an assessment of reading comprehension ability, rather than a reader’s comprehension of a particular text; but the simplicity of the true or false choice has been useful for investigations into broader text and discourse comprehension.

In one such study, Rapp and colleagues investigated the degree to which readers rely on inaccurate information encountered in fictional texts (Rapp, Hinze, Kohlhepp, & Ryskin, 2014). Participants read a fictional story in which the characters stated inaccurate information such as *No one will admit that wearing a seatbelt can reduce your chances of living through an accident.* After a delay, participants took a general knowledge test in which they indicated if a statement was true or false. Some of these items were related to information provided in the story (e.g., *wearing a seatbelt can decrease your chances of living through an accident*). Those who had been exposed to the inaccurate information during reading were more likely to indicate that a false statement was true. The verification of these inaccurate statements suggests that information from texts (even when the source is fictional) is integrated into a reader’s real world knowledge. Importantly, when readers were asked to identify and revise inaccuracies during reading, they were less likely to verify these false statements during the general knowledge test. This indicates that potential harm from inaccurate information can be attenuated if it is addressed at the encoding stage.

Similar to sentence verification, the *inference verification technique* (IVT; Royer, Carlo, Dufresne, & Mestre, 1996) asks participants to make a true or false judgment on a series of statements about information from the text. Some of these statements are ideas that did not directly appear in the text, but are inferences that need to be made in
order to understand the passage. As previously noted, the construction of inferences is critical for the development of a coherent situation model (Kintsch, 1998). Thus, accuracy on this test is assumed to reflect deeper comprehension at the situation model level. Verification statements that require the reader to connect ideas from the same part of the text would assess the reader’s construction of local inferences, whereas statements that require the reader to connect information from distal parts of the text or to elaborate on ideas in the text with information from prior knowledge are assumed to assess the reader’s construction of global inferences (Graesser, Singer, & Trabasso, 1994).

Often the maintenance of a surface code representation comes at the expense of the situation model and, conversely, efforts to construct an elaborate, well-connected situation model can detract from the surface code. Consequently, sentence and inference verification tests are often used in tandem to assess how individual differences or experimental manipulations affect the mental representation as a whole. For example, Wiley and Voss (1999) used these two verification tasks to assess how different writing instructions (write an argument vs. write a narrative) affected readers’ representation of a text. There were no differences in sentence verification accuracy, indicating that the writing instruction manipulation had no effect on readers’ memory for the text. However, participants who were asked to write an argument demonstrated higher accuracy on the inference verification task, indicative of a more elaborate situation model.

Alternatively, researchers can use a variation of the sentence verification task to assess the various levels of representation in a single question. In this task, readers are
asked which of the four versions of a sentence was the precise sentence from the text (Royer, Hastings, & Hook, 1979; Schmalhofer & Galvan, 1986; see Box 1 for an example). The correct answer is the \textit{verbatim} sentence that appeared. A second option is a close \textit{paraphrase} in which the syntax is the same, but the original text has been substituted with synonyms. Selecting this option indicates that the reader’s surface code representation may have decayed, but the gist is still available in memory. The third option preserves the meaning of the sentence or reflects a \textit{plausible inference} from the text. Selecting this option would indicate that this specific idea is no longer in memory, but that there is a coherent situation model. The final option is an incorrect or \textit{implausible inference} in which the information in the sentence is inconsistent with or irrelevant to the text suggesting that the reader either does not have the information in memory at all or that there is a misconception.

\begin{center}
\begin{tabular}{|l|}
\hline
\textbf{Box 1 Example of a sentence verification question from Zwaan (1994)} \\
\textit{Verbatim}: During the daytime he worked in a studio. \\
\textit{Paraphrase}: He worked in a studio during the daytime. \\
\textit{Plausible Inference}: But the police did not believe Sorin. \\
\textit{Implausible Inference}: The drawing was beautiful because Sorin was an artist. \\
\hline
\end{tabular}
\end{center}

Zwaan (1994) used this sentence verification method to test how readers’ expectations for a particular genre, rather than the features of the texts themselves, affected how texts were represented in memory. He provided the same text to all readers, but half were told that the text was a literary short story and the other half were told the text was a newspaper article. After reading, participants took the sentence verification test. There were few instances of readers selecting the \textit{implausible inference}
in either condition, suggesting that readers had little difficulty understanding the text. However, the participants who believed the text was a literary short story selected more of the *verbatim* sentences indicating a stronger surface code. On the other hand, those who believed the text was a newspaper article selected more of the *plausible inference* options indicative of a stronger situation model. These differences suggest that readers’ expectations and reading goals affect the kinds of information attended to and subsequently stored in the mental representation.

*Multiple-Choice Questions*

Possibly the most familiar comprehension assessment for students is the multiple-choice test. These tests are heavily favored in the classroom and on standardized tests such as the SAT or Gates MacGinitie Reading Test because they can be scored quickly and reliably. However, evidence suggests that test-takers can score above chance, even when they are unsure of the correct answer, through familiarity with one of the options, employing test-taking strategies, or relying on prior knowledge (Katz, Lautenshalger, Blackburn, & Harris, 1990; Valencia & Pearson, 1987). Thus, it is essential that researchers carefully construct these multiple-choice tests so that they reflect comprehension processes rather than test-taking skill. Questions can be written to specifically address information that is immediately in the text to assess the surface code or can be written to assess the reader’s ability to make both local and global inferences (McNamara & Dempsey, 2011). Thoughtful construction of the options can also provide additional evidence about what the reader has understood or misunderstood from the text. For example, Ozuru and colleagues (Ozuru, Best, Bell, Witherspoon, & McNamara, 2007) asked the question in Box 2.
The first option, “fission” is the correct answer. The second option is a near-miss distractor. Selecting this option indicates the reader had a shallow understanding of the text and was relying on familiarity. The third option is a common misconception that is related to the text topic, but indicates that the reader was drawing upon a misunderstanding or inaccuracy in prior knowledge. Finally, the fourth option is an unrelated or implausible option that is not related to the text, indicating that the reader did not read thoroughly or did not understand the main idea of the text. By constructing multiple-choice questions in this manner, researchers can assess comprehension not only in terms of whether the answer is correct or incorrect, but also infer the quality of the representation based upon which distractor is selected.

Another way researchers can be more confident that multiple-choice questions measure comprehension is to pilot the materials. Hinze, Wiley, and Pellegrino (2013) explored how constructive retrieval practice can support learning from text by using multiple-choice comprehension questions. They constructed both text-based questions that reflected answers that could be found directly in the text (e.g., [In the eye] What aids in seeing color and detail best?) and inference-based questions in which the
answer was not explicitly in the text (e.g., *If you have trouble identifying a color, it may help you to...*). To assess the validity of their questions, they conducted a pilot study in which three groups of participants were administered the multiple-choice comprehension test. One group of participants read the text and then immediately answered the questions from memory. A second group of participants had the text available when answering the questions. There was no difference in performance between the text available and text unavailable groups on text-based question performance. In contrast, there was a significant difference on inference-based questions, such that those who could refer back to the text scored significantly higher than those who did not have the text available (see also Ozuru et al., 2007; cf. Higgs, Magliano, Vidal-Abarca, Martinez, & McNamara, 2017). This supports the notion that the inference-based questions were indeed tapping different aspects of comprehension. Most importantly, a third group of participants completed the comprehension test without having read the text at all. Even without the text, participants' accuracy was above chance (around 45% for both types of questions), setting a baseline for performance in subsequent experiments. By testing the construct validity of their measures, the researchers could be more confident that this quick and familiar method of assessment was assessing comprehension in the ways that they intended.

*Constructed Response Questions*

Constructed response or *open-ended* questions are often used as an alternative to, or alongside, multiple-choice questions. Constructed response questions can be used to assess the degree to which information is *recollected* as opposed to being merely *familiar* (Yonelinas, 2002) because answering these questions requires the
reader to generate the answer, rather than select the answer from a given set of options. Take for instance this global inference comprehension question from Ozuru and colleagues (2007): *What kind of problem might a person experience as a result of damage to the cerebellum?*. In the multiple-choice version, the reader is given four options: *a) jerky and exaggerated motor movements, b) impairment in learning and memory, c) problems in execution of purposeful actions, or d) inability to regulate emotions*. It is possible that the reader could have generated the correct answer and searched for the answer amongst the options, but it is also possible to read the options to prompt recognition or use the process of elimination. In the constructed response version, however, readers have no recourse other than searching their mental representation in order to recall the information in the text regarding the cerebellum’s function. They must also then be able to manipulate this information in order to consider possible consequences from cerebellum damage.

One consideration when using open-ended responses is that readers can provide a variety of answers that may be more or less correct rendering it necessary to score responses for partial credit. Whereas moving beyond simple correct or incorrect responses is more time intensive, this scoring allows for greater specificity in measuring readers’ memory for and comprehension of the text.

*Free and Cued Recall*

Recall tasks in text comprehension are similar to those in memory research, but replacing list learning with longer passages. Anderson and Pichert (1978) used the following free recall instruction: *Please write down as much of the exact story as you can on these two sheets of paper. If you cannot remember the exact words of any*
sentence, but you do remember the meaning, write down a sentence or part of a sentence as close to the original as possible. It is extremely important that you write down every bit of the story which you can remember (p. 5). By asking for the text “exactly” or “accurately,” the reader must draw from a surface code level. Encouraging the reader to write down the ideas from the text even if the exact wording is forgotten allows the reader to draw on the textbase level. In a classic study by Bransford and Johnson (1972), free recall was used to show that readers who were provided with a title prior to reading recalled more from the text than those who were not provided with the title, suggesting that having the schema activated during reading allowed for a more connected and organized mental representation.

Cued recall is similar to free recall except that rather than being asked for any and all information that can be retrieved from memory, readers are cued to retrieve information related to a specific idea or concept. Such a method is appropriate when researchers want to check for the presence or absence of a particular concept or even as a manipulation to explore how different aspects of text are remembered. Both free and cued recall can be used immediately after reading or after a delay. Cued recall may be particularly appropriate for younger readers who may rely on the cues to fully complete the task (McNamara, Ozuru, & Floyd, 2011).

Though recall is generally assumed to reflect the reader’s textbase level of representation (McNamara & Kintsch, 1996), researchers can score the recall in a variety of ways as a means of assessing different levels of representation. A researcher interested in the surface code level would score the content of the recall for the amount of exact wording, whereas a researcher interested in the textbase level would assess
the number of ideas from the text. Researchers can also identify the quality of a situation model by scoring the recall for information that is consistent with, but not explicitly mentioned in the text. In addition, researchers can also identify inaccurate information that is contradictory to the text to assess readers’ misconceptions.

In addition to hand-scoring these responses, researchers can also use natural language processing (NLP) to assess the reader’s mental representation. NLP tools, such as Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014), can be used to compare the readability of texts used in discourse research, but they can also be used to score open-ended responses quickly with reasonable accuracy. The Constructed Response Analysis Tool (CRAT; Crossley, Kyle, Davenport, & McNamara, 2016) was developed specifically for assessing constructed responses. It uses semantic indices such as Latent Semantic Analysis (LSA; Landauer, Foltz, & Laham, 1998) as well as word-based indices to compare a response to the original text, producing overlap scores that can address similarities at both the word and meaning levels. For more information on natural language processing, Crossley, Allen, Kyle, and McNamara (2014) provide an excellent primer that demonstrates how a simple NLP tool (SiNLP) can be used in text comprehension research.

**Essays**

Essays are another form of constructed response, most often intended to investigate more complex aspects of comprehension and learning. Essay prompts generally ask readers to address a particular question or series of questions that require manipulating or restructuring information from the text in some way, such as making inferences, weighing evidence, or constructing arguments. These prompts can be
science questions such as *Write an argumentative essay on what causes ice ages* (Sanchez & Wiley, 2006), history questions, such as *To what extent did the U.S. Government influence the planning of the [Panamanian] revolution?* (Britt & Aglinskas, 2002), or prompts to think about deeper meanings in literary works, such as *Which do you think is the better interpretation? Use evidence from the text to support your claims* (McCarthy & Goldman, 2015). A written essay is representative of the reader’s ability to comprehend and learn from the text, rather than the ability to remember the text. As such, researchers allow, if not encourage, readers to use the text as a resource as they construct their response.

Like free recall, these essays can be scored using NLP tools (Hastings, Hughes, Magliano, Goldman, & Lawless, 2012; Wiley et al., 2017) or hand-scored by multiple (trained) raters who achieve some metric of reliability (McCarthy & Goldman, 2015). Responses can be parsed into sentences, clauses, or idea units and then categorized. Some of these categories include: Directly from the text (verbatim copying), paraphrased from the text, associations, bridging inferences, global inferences, or interpretive inferences that reflect a nonliteral meaning. Alternatively, or even in addition, the essays can be scored for the presence of certain concepts or relationships.

For example, Wiley and colleagues were interested in how readers integrate information when it comes from different web-based sources (Wiley et al., 2009). Participants were presented seven sources with information about volcanoes and were told that they would be writing a report on what caused the eruption of Mt. St. Helens. Though the texts were presented via computer to appear as a Google search, the text set was constructed by the researchers so that the set included all of the necessary...
information to create a complete causal model. Furthermore, the set included sources of varying degrees of credibility, and different texts offered overlapping, conflicting, and unique information. After reading, participants were either asked to construct a description (intended to bias summarization) or an argument (intended to bias elaboration) for why volcanoes erupt. The essays were compared to the causal model and were scored for both the number of core causes and erroneous causes that reflect a misconception. Participants who wrote an argument not only included more core causes in their essays, but they included fewer erroneous causes, indicating that the argument task orientation led readers to construct a more complete and accurate mental representation of the scientific phenomena.

When using essays as a measure of text comprehension, it is important to remember that the prompt may bias what information from the text is task-relevant and, as a result, the reader may choose to omit information that does not directly address the prompt. With this in mind, researchers cannot assume that the essay reflects the full extent of a reader’s mental representation. In addition, essays require additional skills beyond reading comprehension, such as proficiency in writing (Ackerman & Smith, 1988; Graham & Perin, 2007; Weston-Sementelli, Allen, & McNamara, 2016) and argumentation (Goldman & Lee, 2014). Indeed, research indicates that there is some overlap in the underlying cognitive processes involved in reading and writing, but many of the higher-order cognitive skills that strongly predict reading comprehension ability, such as inferencing skill and knowledge integration, do not predict writing quality (Allen, Snow, Crossley, Jackson, & McNamara, 2014). This suggests that researchers must consider that essays may not provide a perfect window into a reader’s mental model.
Measuring Behaviors during Post-Reading Assessments

In addition to these post-reading assessments, researchers have also examined the online behaviors that occur while completing post-reading assessments. These fine-grained measures can reveal further information about the mental representation a reader has constructed. For example, Read&Answer (Vidal-Abarca et al., 2011) is a tool that can be used to measure on-line reading processes, but was also designed to capture readers’ question-answering behaviors. Read&Answer uses a moving window paradigm to mask coarse-grained (e.g., sentence or paragraph level, rather than individual words) regions of interest. After reading, comprehension questions appear in a window on the side of the screen, allowing the reader to refer back to the text to answer. In order to see the different parts of the text, readers click on the region of interest and the system records their moves. Researchers can then make inferences about the reader’s representation and their question-answering strategy based on whether the reader looks back at the text at all and what parts of the text the reader reviews. Read&Answer is often used in studies that investigate how different reading or comprehension goals affect the way readers process and comprehend the text (e.g., Higgs et al., 2014; Vidal-Abarca & Cerdán, 2013).

A more recently developed measure is the logging of keystrokes during composition. These logs have been used to develop indices, such as number of keystrokes, latencies between keystrokes, and the degree of uniformity across multiple time windows, to capture information about individuals’ composition behaviors. Research indicates that these keystroke indices are significantly correlated with both the quality and characteristics of students’ essays (Allen et al., 2016a). Further, these
indices have been used to model students’ affective states during writing (Allen et al., 2016b; Bixler & D’Mello, 2013). As these measures can be used to assess readers’ understanding of texts, keystroke logging is a potentially powerful tool in assessing underlying comprehension processes.

PROCESSING MEASURES

Process, or on-line, measures have been developed to move beyond readers’ text understanding and tap into the cognitive processes (e.g., inference generation) that occur during reading (Singer, 1990; Zwaan & Singer, 2003). Some of these methods measure behavior continuously during reading, whereas others interrupt reading to measure progress. The overall assumption of these methods is that temporal measures of reading, such as pause times and eye movements, reflect critical differences in the cognitive processes employed during comprehension.

Lexical Decision Task

In a lexical decision task (Myer & Schvaneveldt, 1971), readers are asked to judge if sets of words are related or unrelated or if a string of letters is a word or a non-word. Non-words that do not follow linguistic rules (e.g., pwxl) are rejected more quickly than those that follow linguistic rules, but have no semantic meaning (e.g., marb). Similarly, pairs of words that are semantically related are identified more quickly than those that are unrelated. Although readers typically achieve high accuracy rates on these tasks, the speed at which they respond provides information about the availability of this lexical information in memory. The underlying assumption of these tasks is that concepts must be active in memory in order to be integrated into the mental
representation. Thus, words that are related to the text will have stronger activations and be more quickly retrieved than information that is unrelated to the text.

Lexical decision tasks have been used by text researchers to test whether inferences are generated on-line and if these inferences are generated automatically or strategically. These studies again rely on the assumption that the construction of an inference depends on having that information active in memory. Readers are asked to judge if the presented word appeared in the text they just read, based on the assumption that words related to a particular inference should be recognized more quickly than unrelated words.

McKoon and Ratcliff (1992) had participants read a narrative text in which a character pursued a main goal and subordinate goals. Participants read one of three versions of the story: 1) main goal complete, 2) main goal incomplete, sub-goal complete, 3) main goal incomplete, sub-goal incomplete. After each text, participants were presented with a target word and asked to judge if it appeared in the passage. There were no differences in response time across the three conditions for target words related to the main goal. In contrast, response times for target words related to the sub-goals were affected by text condition. Participants in the sub-goal complete condition responded more slowly than those in the sub-goal incomplete condition, suggesting that readers keep active only the most basic information necessary to maintain coherence from sentence-to-sentence. This is in contrast to keeping active information related to maintaining global coherence across the entire text. McKoon and Ratcliff used these and other findings to argue that inferences are made automatically and only when absolutely necessary.
In contrast, Millis and Graesser (1994) used a lexical decision task to show that some strategic inferences are generated during reading. They had participants read short scientific texts that described a mechanism, such as how steam powers a turbine to produce electricity. Response times indicated that words related to causal inferences (e.g., steam rises) were identified more quickly than words related to elaborative inferences (e.g., electricity will be produced) or non-words. These findings suggest that readers do generate some inferences strategically to maintain coherence at a global level.

Lexical decision tasks are useful in that they are relatively easy to implement on a computer. However, the task is relatively invasive and can disrupt natural reading processes.

*Think Aloud/Talk Aloud*

Think-aloud protocols, or simply *think-alouds*, ask readers to report their thoughts as they read through a text (Cote & Goldman, 1999; Magliano, Trabasso, & Graesser, 1999). Think aloud or “talk aloud” procedures ask readers to “turn up the volume on their inner thoughts” as they read through the text (Peskin, 1998). Analysis of the reader’s utterances can then be used to assess what information is active as well as what conscious, strategic processing is taking place (Ericcson & Simon, 1980). Though think-alouds should not be considered a one-to-one mapping of the ongoing cognitive processes that occur during reading, they are useful for capturing online processes that might not be apparent in post-reading assessments (Burkett & Goldman, 2016).

Think-alouds, like open-ended post-reading measures, can be coded for evidence of particular behaviors (e.g., explaining) or idea units that are reflective of a
particular category of behaviors (e.g., comprehension monitoring). Trabasso and Magliano (1996) used a think-aloud method to investigate the extent to which global, causal inferences are made during the reading of narrative texts. Participants read eight short stories (13-16 sentences) and were prompted to think-aloud after each sentence. These utterances were coded as explanation, association, or prediction inferences as well as the associated memory operation: activation of general world knowledge, retrieval of text, paraphrase or maintenance of information. Readers’ statements, such as “Ivan does not want to be seen” in response to the text Ivan waited until dark, reflected the presence of explanation inferences to resolve why a character engaged in a particular behavior. The think-alouds also indicated that readers relied on activation of general world knowledge rather than retrieval or maintenance of text information, emphasizing the importance of prior knowledge in the construction of coherent and elaborate mental representations.

Wolfe and Goldman (2005) scored readers’ think-alouds to analyze both the cognitive and metacognitive behaviors that the readers engaged as they read multiple documents about a historical event (i.e., the fall of Rome). Specifically, these researchers proposed the nomenclature presented in Table 1.

Table 1. Scoring categories from Wolfe and Goldman (2005)

<table>
<thead>
<tr>
<th>Category</th>
<th>Subcategory</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paraphrase</td>
<td></td>
<td>adds no novel information because statement is semantically similar to the original text</td>
</tr>
<tr>
<td>Evaluation</td>
<td></td>
<td>conveys either a positive or negative view of the text or the author</td>
</tr>
</tbody>
</table>
Comprehension
Problem indicates a lack of understanding at the word, phrase, or sentence level

Comprehension
Success indicates understanding of the text

Elaborations embellishes the current sentence

<table>
<thead>
<tr>
<th>Elaborations</th>
<th>Adds information to connect to the current text at the situation model level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-explanations</td>
<td>Connects the current sentences to another sentence based on surface features</td>
</tr>
<tr>
<td>Surface Text</td>
<td>Connects the current sentences to another sentence based on surface features</td>
</tr>
<tr>
<td>Connections</td>
<td>Adds no new relevant information even though information is new</td>
</tr>
<tr>
<td>Irrelevant</td>
<td>Adds no new relevant information even though information is new</td>
</tr>
<tr>
<td>Associations</td>
<td>Indicates what the reader expects to happen next</td>
</tr>
</tbody>
</table>

Using this rubric, Wolfe and Goldman (2005) were able to investigate not only how the readers comprehended the individual texts, but also how they reasoned from the multiple documents to develop a coherent representation of events. They found that the most prevalent behavior was to generate elaborative inferences in which they self-explained the information to connect ideas across the texts. In addition to the think-aloud, readers answered post-reading questions that required them to explain the fall of Rome in their own words. The amount of self-explanation in students’ think-alouds was positively correlated with the quality of the post-reading explanation.

Notably, some researchers have raised concerns about the conclusions that can be drawn from think-alouds because thinking aloud while reading can either disrupt the natural reading process or encourage readers to engage in strategies or processes that may not have occurred without verbalization (Branch, 2000; Nisbett & Wilson, 1977). Nonetheless, research has demonstrated that think-alouds largely reflect natural comprehension processes (Magliano & Millis, 2003; Zwaan & Brown, 1996).
Reading Time

Another methodology that has been used to investigate comprehension processes is to assess reading time during critical moments of the text. Text is presented on a computer screen and the reader progresses to the next screen with a key press. The computer records the length of time between key presses as an indicator of the amount of time spent on each word or sentence. Researchers compare these times across experimental conditions or compare target sentence reading times to a baseline reading time. Longer reading times are assumed to be evidence of additional processing, whereas shorter reading times are assumed to be evidence of facilitation.

In such studies, the text presentation progresses either word-by-word or sentence-by-sentence. In non-cumulative self-paced reading, the old text disappears as new text appears. In some versions of the task, each sentence is presented in the center of the screen. In a moving window version of the task, the entire passage is presented with only one sentence visible at a time, as shown in the example below:

```
xxxx xx xxx xxxxx xxxxxxxxx xxxx xxx xxxx. xxxx xx xxx xxxxxxxxx xxxxxxxxxx. This is the sentence currently being read.
Xxxx xxx xxxxx xxxx xxx xxx xx xxxxxxx xxxxxxxx xxx xx xxxxx xx xxxxxxxxx xxxxxxxxx xxx xxxxxx xxxxxxxxx xxxxxxxxx xxxxxxxxx xxxxxxxxx xxxxxxxxx.
```

In cumulative self-paced reading, the new text is added on while the old text remains visible as seen below:
Though non-cumulative self-paced reading better isolates the target information, it is somewhat restrictive because in normal reading, readers often look back to previous text as part of normal comprehension. Thus, having to rely only on the currently visible sentence increases cognitive demands (Rayner, 1998).

An example of the use of reading times was illustrated by O’Brien and colleagues (Albrecht & O’Brien, 1993; Meyers, O’Brien, Albrecht, & Mason, 1994). These researchers assessed how readers update their mental representations by presenting target sentences that were either consistent or inconsistent with previous information in the text. In a now classic example, participants read a short passage about a woman named Mary. Mary was described as either a healthy vegetarian (inconsistent), a junk food addict (consistent), or without additional information about her diet (neutral). Reading times for the target sentence, *Mary ordered a cheeseburger*, indicated that readers in the inconsistent condition took longer to read this sentence than those in the neutral or consistent condition. This finding suggests that the information related to Mary’s diet was still active in memory, and that readers required more time to integrate this conflicting information into their mental representation of the text.

Reading time assessments can also be used to test whether or not an inference has been generated during reading. In these studies, the texts contain a target sentence that require an inference, such as *Silicone tiles were used for the space shuttle because*
the total weight had to be kept low (Noordman, Vonk, and Kempff, 1992). In an explicit condition, participants are given the inferred information (e.g., Silicone is much lighter than the material used for traditional heat shields). In an implicit condition, this sentence is omitted so that participants must generate the inference on their own. Longer reading times in the implicit condition are indicative of the inference being generated on-line.

Research using this paradigm indicates that readers generate inferences on-line when reading narrative texts, but not when reading science texts (Noordman, Vonk, and Kempff, 1992; Singer, Harkness, & Stewert, 1997).

Thus, these different reading time paradigms can be used to isolate and assess a variety of rapid comprehension processes (e.g., inference generation and integration of current information) that occur during reading.

**Eyetracking**

Recording eye movement during reading provides both a more fine-grained method of analysis and more natural reading experience than word-by-word or sentence-by-sentence reading time paradigms. A calibrated camera records gaze location, duration, and moment-by-moment movement by tracking the pupil as a reader reads from the computer screen. The assumption made in this research is that eye movements reflect, in real-time, the underlying cognitive processes that are occurring during reading (Rayner, Pollatsek, Ashby, & Clifton, 2012). Importantly, the eyes do not move in a smooth, sweeping left to right motion (or right to left or up or down depending on the language). Instead reader’s eyes tend to “jump” across words, landing on some (mostly content words) and skipping others (mostly function words). Researchers are particularly interested in the location and duration of fixations (i.e., when the eyes land
on a particular word), *saccades* (i.e., when the eyes move forward), and regressive saccades or *regressions* (i.e., when the eyes move backward toward previously read parts of the text; Rayner et al., 2012). Longer fixations and more regressions indicate attempts to resolve issues in comprehension. For example, readers engage in more rereading and regressions when a sentence is ironic because they attempt to resolve the literal inconsistency (Kaakinen, Olkonimi, Kinnari, & Hyönä, 2014).

Wiley and Rayner (2000) used eyetracking to further understand how activating information prior to reading influenced online processes. Capitalizing on the title manipulation used by Bransford and Johnson (1972), participants read short passages with or without a title. In addition to replicating the beneficial effect of title on recall, they showed that a title increased reading speed at the end of a sentence, indicative of faster integration of new information. The eye movement patterns also indicated that readers who were provided a title engaged in fewer regressions and shorter fixations on content nouns. Taken together these eye movement patterns suggest that activation of prior knowledge supports comprehension at both the surface and situation model levels.

Notably, eyetracking has become increasingly easy to use. In early eyetracking research, the technology was both expensive and cumbersome, as it required a head stabilizer, such as bite bar or chin rest. However, more modern eyetrackers are significantly more affordable and can be head-mounted or even affixed to the top or bottom of the monitor, allowing for a more natural reading experience. However, one limitation is that it is still difficult to track eye movement across multiple screens, so research is limited to relatively short passages. There have also been developments in wearable eyetrackers embedded within eyeglasses, which may someday allow us to
capture eye movements of reading in more natural reading environments. However, the technology is not yet advanced enough to track these fine-grained reading processes (Ye, et al., 2012). An extensive review of the kind of research that uses eyetracking can be found in a thorough tutorial on how to use an eyetracker (including a short video demonstration) by Raney, Campbell, and Bovee (2014).

**Neural Measures**

Neural measures provide insight into the biological foundations of text comprehension processes revealed by behavioral studies (e.g., Ferstl & Von Cramon, 2001). Text comprehension researchers measure event-related potentials (ERPs) in electroencephalograms to assess the time-course of these neural responses and functional magnetic resonance imaging (fMRI) to identify which regions of the brain are being recruited in a given reading task (see Dehaene, 2009 and Mason & Just, 2006, for reviews of neuroimaging research in discourse processing).

The most common form of neural imaging is the assessment of event-related potentials (ERPs). ERPs measure electrical activity over time through electrodes placed in specific locations on the scalp. Researchers then use an averaging method to isolate the ERP signal from the background “noise.” Benefits of this method are that it allows readers to read text in a relatively normal environment (as opposed to lying still for an fMRI) and that the cost is significantly more affordable.

The N400 component of the ERP signal is of particular interest to text researchers. The N400 is a negative deflection in the signal that occurs approximately 400 ms after the detection of an anomaly and is indicative of semantic processing (Kutas & Hillyard, 1980). Nieuwland and Van Berkum (2006) presented participants with
target sentences with semantic violations (e.g., *The peanut fell in love*). When this sentence was preceded by a realistic context, the typical N400 effect emerged. In contrast, when readers were given a cartoon-like context in which peanuts appeared animate, this target sentence showed no N400, suggesting the readers did not detect a semantic violation. Interestingly, when a semantically-correct, but globally-incoherent statement like *The peanut was salted* was presented in this cartoon-like context, it produced the N400 effect. These findings indicate that readers are sensitive to violations not only at the sentence level, but also at the more global situation model level.

In early brain imaging studies, researchers relied on positron emission tomography (PET). However, this method has fallen largely out of favor given that it relies on the injection of radioactive tracer and has a relatively coarse-grained unit of analysis. Instead, modern brain imaging is largely conducted via functional magnetic resonance imaging (fMRI). fMRI is a technique that measures changes in oxygenation to various part of the brain during moment-by-moment processing. Areas that receive more oxygen are assumed to be more active (Cohen & Bookheimer, 1994). For example, Prat, Mason, and Just (2011) used fMRI to investigate how individual differences in neural efficiency relate to the generation of causal inferences. In the fMRI scanner, participants read high or low coherence passages. The less coherent passages showed greater activation in several regions of the left hemisphere (angular gyrus, superior parietal region, and inferior and middle temporal gyri). However, those with lower reading skill (based on their performance on the Nelson Denny Vocabulary Test) showed greater right-hemisphere activation when reading the less-coherent texts.
compared to those with higher reading skill. The authors interpreted these findings in terms of a *spillover* effect, such that readers recruit the right hemisphere when the language network in the left hemisphere cannot keep up with current processing demands. Indeed, these results are consistent with other studies indicating the recruitment of the right hemisphere during comprehension, particularly during inference generation (e.g., Beeman, Bowden, & Gernsbacher, 2000).

Often the findings of fMRI research are limited due to small sample size because the costs of owning and maintaining the technology are quite high. In addition, participants must be able to remain completely still during the task. Another consideration is that imaging research is generally limited to right-handed participants, because left-handed individuals tend to show less lateralization than their right-handed counterparts (Desmond et al., 1995). This makes identifying regions of interest more difficult to sample.

Despite some limitations, fMRI data provides a wealth of neural information on how readers understand text. These techniques allow researchers to investigate structures of the brain involved in different aspects of text memory, comprehension, and learning as well as provide information regarding how these processes unfold biologically in real time.

**THINKING DYNAMICALLY ABOUT COMPREHENSION**

The wide assortment of assessment methodologies described in this chapter have afforded researchers the ability to develop a robust understanding of text comprehension. Post-reading measures are able to tap into readers’ mental text representations at multiple levels, whereas on-line measures identify the specific
processes that are involved in the construction of this representation. One limitation of these current approaches, however, is that they provide a relatively static and summative picture of comprehension. Although on-line assessments do seek to examine the learning processes underlying comprehension, these measures are often aggregated across entire reading sessions. For instance, researchers might examine the frequency with which readers employ specific strategies or engage in metacognitive monitoring; however, the dynamic nature of these processes is not preserved.

This gap is problematic because text comprehension is far from a linear, ordered process. Rather, the processes involved in learning from texts are complex and multi-dimensional, and they draw on multiple resources, sources of knowledge, strategies, and abilities in a non-linear fashion. Although traditional measures that rely on inferential statistics have successfully captured some of this complexity, they cannot provide a complete picture of the dynamics. Consequently, these assessments can lead researchers to characterize text comprehension processes in overly simplistic terms.

One perspective that may help researchers to effectively model this complexity comes from Dynamical Systems Theory (DST; Gallagher & Appenzeller, 1999). DST is an emerging theoretical and methodological approach to psychology that has largely been popularized due to its emphasis on systems that are complex, non-linear, and highly interactive (Dale, Fursaroli, Duran, & Richardson, 2013; Kello, 2013; Richardson et al., 2015; Vallacher, Read, & Nowak, 2002). The principal aim of this approach is to consider human behavior as a complex system that emerges non-linearly through numerous interactions.
Particularly relevant to the assessment of text comprehension is DST’s emphasis on complex systems and self-organized behaviors. The term complex system refers to a system of multiple interacting components, such as people or words, that cannot additively explain the overall behavior of the system (Gallagher & Appenzeller, 1999). In other words, the critical importance of describing systems as complex is in the acknowledgment that these systems are interaction-dominant. This means that the individual system components act as a function of the other system components, such that the system cannot be reduced to the sum of its individual parts.

In considering these properties of complex systems, it naturally follows that the system cannot be driven by a central controller. Thus, the cognitive mechanisms commonly attributed to control processes, such as working memory or attention control, are no longer needed to explain how humans are able to integrate and regulate the multitude of interacting components involved in natural psychological tasks (see Turvey, 1990). In contrast to standard cognitive science approaches, DST emphasizes the fact that these processes can be coordinated through self-organization. Self-organization is simply the process by which a structured system of behaviors can emerge through interactions among components at the local level of the system, rather than through controlled planning by a central executive. This concept has been used to describe processes in a number of domains within (Dixon, Stephen, Boncoddo, & Anastas, 2010) and outside of psychology (see Kauffman, 1996).

Given the complexity of the text comprehension process, DST may prove to be an extremely useful perspective that can contribute to the extension of both theoretical and methodological approaches. From a theoretical perspective, learning from text can
be considered to be a complex system that achieves coherence through self-organization. In this vein, the multiple layers and interactions involved in text comprehension, such as decoding, sentence processing, and inference generation, can be considered part of the same system. From a methodological standpoint, DST can provide sophisticated new approaches to understanding these processes. DST measures aim to characterize systems, rather than individual components, and place a strong emphasis on understanding how these systems emerge and change over time. Thus, time is treated as a critical variable in this approach, and these temporal analyses can provide important information about human behaviors.

One dynamic method that may be useful for studying comprehension processes is Recurrence Quantification Analysis (RQA). RQA is a nonlinear data analysis method that provides information about patterns of repeated behavior in a continuous or categorical time series. Like many techniques used in the dynamic systems theory framework, this methodology has been used in a variety of domains, both within and outside the realm of human behavior (Shockley, Santana, & Fowler, 2003; Dale & Spivey, 2005).

The initial outcome of the RQA approach is the recurrence plot, which is a visualization of a matrix where the elements in the matrix represent particular time points in a temporal series that are visited more than once (i.e., they recur). In other words, this plot represents the times in which the dynamical system visits the same area in the phase space (Marwan, 2008). Take for example, the sentence “The ice cream man brought ice cream on Friday”. This sentence generates the recurrence plot below:
A point in the plot indicates where a word appears in both the X and Y axes. Because the sentence is plotted against itself, the plot is symmetrical with the diagonal line of identity (LOI) through the center. The points of interest in a recurrence plot are those outside of the LOI. In this example, the words “ice” and “cream” appear twice and, notably, appear next to each other in the same order both times. Beyond the visualization of the recurrence plot, RQA can provide important information about the patterns of recurrence in a dynamic system, such as the rate of recurrence (probability a state will recur), the determinism (predictability) of the system, the number of lines in the dynamic system, the average length of these lines, and the maximum line length (see Coco & Dale, 2014, for a more thorough explanation of this methodology). Within the context of reading comprehension assessment, RQA may be easily used in a variety of ways, such as examining fine-grained behavioral measures (e.g., eye tracking or reading times) to high-level aspects of the reading processes (e.g., affective states and linguistic properties.)

At the fine-grained level, keystroke analyses present a particularly fruitful area in which RQA might be applied. As described previously in this chapter, researchers have
recently begun to leverage the power of computer logging tools to extract fine-grained information about students’ writing processes (Leijten & van Waes, 2016). Patterns in the pauses and bursts of keystroke logs have provided important information about affect, fluency, and proficiency of individual writers (Allen et al., 2016a; Bixler & D'Mello, 2013; Leijten & van Waes, 2016). One problem with these analyses, however, is that there is a tendency to focus primarily on aggregated features of the keystrokes (e.g., total number of pauses), rather than taking the temporal structure of the time series into account. This aggregation process can often cause researchers to miss out on important nuances in the time series that they are analyzing. RQA might be a particularly useful technique to apply to this form of data set. Information about the number and time of recurrent patterns might, for example, provide information about different behavioral states. For example, high determinism (i.e., a high number of consecutive keystroke intervals that recur throughout the writing task) during typing may relate to high degree of “flow” during writing, such that writers engage in highly predictable patterns. Low determinism, on the other hand, might be characteristic of writers who are less engaged in the task or having problems getting into a “rhythm” with their writing (Allen, Likens, & McNamara, 2017).

Beyond keystroke indices, RQA measures have the potential to provide important information about recurrence in a number of processing measures, such as eye-tracking and reading times. Rather than examining how long a particular passage took for an individual to read, for example, researchers could examine whether there were systematic patterns in these reading times, and whether this variability relates to an individual’s successful comprehension of the text. An additional strength of RQA is
that it can be calculated on continuous or categorical data. Thus, RQA can be applied to
categorical data, such as words or parts of speech, to examine the dynamics of the text
itself (Allen, Perret, Likens, & McNamara, 2017). The flexibility of this RQA technique
(i.e., the fact that it can be applied to both continuous and categorical data sets) may be
particularly salient for the study of reading comprehension. Using this methodology, the
dynamics of text comprehension could be modeled at multiple levels, rather than relying
solely on one level of analysis.

CONCLUDING THOUGHTS

Text comprehension is a widespread activity in our current society and is
important for our ability to successfully engage with our community. Like most learning
processes, the ability to learn from text is a complex process that requires individuals to
leverage their knowledge across a variety of domains to actively construct a coherent
representation of the concepts they are learning. Given this complexity, text
comprehension researchers must rely on multiple forms of assessment to develop a
robust understanding of this process. In this chapter, we provided a brief overview of
research in this domain and described the most common methods for assessing text
comprehension. In particular, we described assessments that focus on the “offline” or
final understanding of a text that a reader develops, as well as assessments that tap
into the processes that contribute to this understanding.

An important conclusion of this overview is that text comprehension cannot be
thoroughly explained with any one particular metric. In our lab, for example, we have
relied on a wide variety of the assessments described in this chapter. We have used
recall tasks to determine the information that readers recall from text, as well as think-
aloud protocols to tap into their comprehension processes. In parallel with these measures, we have analyzed the properties of the texts that students are asked to read, as well as the linguistic properties of the think-alouds they have produced. We have even analyzed their keystrokes and reading times to identify potential areas of confusion or engagement. On their own, none of these metrics would be able to provide a sophisticated understanding of comprehension. Rather, the power of these assessments lies in the information that is discovered when they are combined.

Our objective in this chapter is to encourage researchers to consider this wide breadth of assessment options when conducting their own experiments. Our view is that it is crucial that the complexity of reading comprehension processes is highlighted and assessed more thoroughly so that we can develop a more complete understanding of these processes. Such an approach will be essential to integrating notions of lower- and higher-level reading, as well as learning more broadly.
References


Kendeou, P. Bohn-Gettler, C.M. Fulton, S. (2011). What we have been missing: The role of goals in reading comprehension. In M. T. McCrudden, J. P. Magliano, &
G. Schraw (Eds.), *Text relevance and learning from text*. Greenwich, CT: Information Age Publishing.


van den Broek, P., Young, M., Yuhtsuen, T., & Linderholm, T. (1999). The Landscape Model of reading: Inferences and the online construction of memory representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction*
of mental representations during reading (pp. 71–98). Mahwah, NJ: LEA.


