Running head: INFERENCES IN NATURALISTIC CONTEXTS

Constructing Inferences in Naturalistic Reading Contexts

Arthur C. Graesser, Haiying Li, and Shi Feng

University of Memphis

Graesser, A.C., Li, H., Feng, S. (in press).Constructing inferences in naturalistic reading contexts.In E. O'Brien, A. Cook, and R. Lorch, (Eds.), *Inferences during Reading*. Cambridge: Cambridge University Press.

Send correspondence to:

Art Graesser
Psychology Department
202 Psychology Building
University of Memphis
Memphis, TN, 38152-3230
901-678-4857
901-678-2579 (fax)
a-graesser@memphis.edu

KEYWORDS: Inferences, comprehension, naturalistic texts

Constructing Inferences in Naturalistic Reading Contexts

Twenty-five years ago my colleagues and I (the first author of this chapter) were in an intense debate over what inferences are generated during text comprehension. It was a lively debate. At the one end there was the *minimalist* position that predicted that the only inferences that were encoded were those that were activated quickly by associations in long-term memory and those that were triggered by obstacles in text cohesion that forced more deliberate thought (McKoon & Ratcliff, 1992).At the other end was the *promiscuous* position, typically by researchers outside of psychology (ranging from literature to computer science) that postulated that a very large and unconstrained set of inferences were generated virtue of background knowledge and experiences. My research team advocated an intermediate *constructionist* position (Graesser, Singer, & Trabasso, 1994; Singer, Graesser, & Trabasso, 1994) that attempted to predict what subset of inferences are generated on the basis of what we know about social, discourse, and cognitive mechanisms. These mechanisms stretched beyond the memory-based models of the minimalist position and into concerns of the goals and emotions of people and the explanations of events that occur in our everyday worlds. A book edited by Graesser and Bower (1990) articulated the various theoretical positions, methods to study inference generation, and of course empirical data.

A flurry of models and empirical findings evolved in the 1990's in an effort to pin down what inferences were generated under what conditions. The models also aspired to more general goals of specifying the discourse representations that are encoded during comprehension, the processes of generating these representations, and performance on different tasks. The research efforts are captured in several edited volumes in the mid to late 1990's that covered psychological research on text comprehension, inference generation, coherence, and other components of deeper levels of understanding (Britton & Graesser, 1996; Goldman, Graesser, &

van den Broek, 1999; Goldman & van Oostendorpo, 1999; Lorch & O'Brien, 1995; Weaver, Mannes, & Fletcher, 1995; Zwaan & van Oostendorp, 1993).Below are a sample of these models and their signature claims.

(1) *Construction-integration model* (Kintsch, 1998). Comprehension is guided by the bottom-up activation of knowledge in long-term memory from textual input and contents in working memory (the *construction* phase) followed by the integration of activated ideas in working memory (the *integration* phase). As each sentence or clause in a text is comprehended, there is a construction phase followed by an integration phase.

(2) *Structure building framework* (Gernsbacher, 1990).Information structures are built during comprehension but cohesion gaps force the reader to build new structures. Inferences are needed to conceptually relate structures that are weakly related.

(3) *Memory-resonance models* (Cook, Halleran, & O'Brien, 1998; Myers & O'Brien, 1998; O'Brien, Rizzella, Albrecht, & Halleran, 1998). Inferences are generated to the extent that there is a resonance between world knowledge and text cues plus contents of working memory.

(4) *Event indexing model* (Zwaan, Magliano & Graesser, 1995; Zwaan & Radvansky, 1998). Inferences are generated when there is a break in causal, intentional (goals), spatial, and/or temporal cohesion.

(5) *Landscape model* (van den Broek, Risden, Fletcher, & Thurlow, 1996). Inferences are activated by text and world knowledge in order to establish referential and causal cohesion.

(6) *Embodied and indexical models* (Glenberg, 1997; Glenberg & Robertson, 1999). Inferences are needed to elaborate the perceptions, actions, and emotions of characters in the situation model.

It appeared that each research team had its own sketch of inference mechanisms as they attempted to sort out the contributions of text characteristics, world knowledge, working memory, comprehension strategies, and task constraints.

In the midst of this blossoming of models and empirical work on inferences and discourse representations in the late 1990's, an unexpected series of events unfolded that shined the spotlight on studies of inference generation and text comprehension in the national arena. There was substantial funding behind the momentum. The Institute of Education Sciences of the US Department of Education was launched at the beginning of the new millennium. Substantial funding was influenced by some research panel reports that emphasized the need to better understand comprehension mechanisms (such as the *Reading for Understanding* report of the National Research Council, Snow, 2002) and better understand learning at deeper levels of mastery (Bransford, Brown, & Cocking, 2000). The National Science Foundation and the Office of Naval Research had a series of interdisciplinary initiatives that focused on the learning sciences, discourse comprehension, automated natural language processing, intelligent tutoring systems, and other computer technologies that promote deeper learning, comprehension, problem solving, and reasoning. Revolutionary advances in computer science, computational linguistics, and information retrieval changed the landscape of research avenues for some of us who had been investigating discourse comprehension. For the first time in history, we could get electronic access to a large repository of thousands (if not millions) of documents for computer analyses. We could analyze the texts with lexicons, syntactic parsers, and semantic analyzers developed in computational linguistics (Jurafsky & Martin, 2008) and statistical models of world knowledge (Landauer, McNamara, Dennis, & Kintsch (2007). The world had changed.

These tends fundamentally changed the direction of research for many of us. Our focus shifted from systematic experimental studies designed to discriminate models of comprehension (including inferences) to more practical but equally scientific directions. By the year 2000, the research teams at the University of Memphis were building automated computer systems with conversational agents (talking heads), such as AutoTutor and iSTART, that helped students better understand science texts by holding conversations in natural language (Graesser et al., 2001, 2004, 2012; Graesser, Jeon, & Dufty, 2008; McNamara, O'Reilly, Best, & Ozuru, 2006). Developing such systems required an interdisciplinary intersection of computer science, psychology, education, and linguistics. We were building automated text analysis systems, such as Coh-Metrix, that analyzed discourse automatically at multiple levels: words, syntax, discourse cohesion, and genre (Graesser et al., 2004, Graesser, McNamara, & Kulikowich, 2011; McNamara, Graesser, Cai, & McCarthy, 2014).We were using eye tracking methodologies to analyze college students comprehension of naturalistic materials, such as illustrated texts from Maccauly's (1988) *The Way Things Work* (Graesser, Olde, Cooper-Pye, & Whitten, 2005) and survey questions on the US Census (Graesser, Cai, Louwerse, & Daniels, 2006). Our research world was shifting from the tight experimental paradigms that investigated *textoids* (texts created by experimental psychologists and linguists) and short discourse segments to automated systems, complex learning, and lengthy naturalistic texts. Some of our colleagues may have viewed us as going down the primrose path but others viewed our efforts as having a good balance between science in the lab and science in the real world.

This chapter has two major goals. Our first goal is to articulate the constructionist model of inference generation (Graesser, Singer, & Trabasso, 1994) and to reflect on where it stands today, over two decades later. We will particularly focus on its three distinctive components:

reader goals, coherence, and explanation. These three components are receiving considerable research attention today, just as they did decades ago. Our second goal is to briefly identify some ways that automated computer analyses can help researchers solve some of the theoretical and empirical challenges in investigations of inferences during comprehension.

The Constructionist Model: What does it Claim and Where does it Stand Today?

The constructionist model of Graesser, Singer and Trabasso (1994) was originally designed to account for the inferences that readers generate during narrative comprehension. Readers generally have a sufficient body of experiences and background world knowledge to understand the episodes and supporting content of narrative texts so many inferences are expected to be constructed (Graesser, 1981; Hiebert & Mesmer, 2013).The model was extended to science texts (Graesser & Bertus, 1998; Millis & Graesser, 1994) with some modest success, but inferences are limited because of the lack of background world knowledge on most science topics. As with virtually all other psychological models of inference generation, the constructionist model assumes that readers have a rich background of declarative and experiential world knowledge in long-term memory that gets activated during comprehension and that gets recruited to fill in inferences. The model also assumes, along with other models, that there is a working memory that holds a limited amount of information and a discourse focus that holds prominent words or ideas in the mind's eye.

Nevertheless, memory activations are not sufficient according to the constructionist model. Comprehension also is to some extent directed and strategic. The distinctive strategies of this model are reflected in its three principal assumptions: reader goals, coherence, and explanation. The *reader goal* assumption states that readers attend to content in the text that is relevant to the goals of the reader. For example, advertisements in a newspaper are written and

read for very different purposes than factual news reports. The *coherence* assumption states that readers attempt to construct meaning representations that are coherent at both local and global levels. Cohesion gaps in the text will stimulate the reader to actively think, generate inferences, and reinterpret the text in an effort to fill in, repair, or acknowledge the cohesion gap. The *explanation* assumption states that good comprehenders tend to generate explanations of *why* events and actions in the text occur, *why* states exist, and *why* the author bothers expressing particular ideas. Why-questions encourage analysis of goals/plan of characters, of causal mechanisms, of justifications of claims, and other deeper levels of understanding.

The constructionist model set the bar higher on what it means to comprehend text than the minimalist position, memory-based models, and the construction integration model. It also was more attentive to the social, biological, and developmental worlds of humans. We do live in a world where people are trying to achieve goals, are experiencing emotions when the goals are blocked versus achieved, are explaining what is happening in the world to ensure survival and satisfaction, and when they want to be entertained. Designers of memory-based models have never understood the depth of this wisdom that the rest of psychology has profoundly embraced. However, the constructionist model was limited by the fact that the deeper inferences relied so much on world knowledge, and also the fact that the reader's goals were potentially so ad hoc that a science could never emerge from the reader-goal assumption. So this raises the question of where the science is. This is the question to which we turn.

Reader Goal Assumption

This assumption is trivial and nonscientific if the underlying claim is that readers generate inferences that address their goals. The obvious prediction would be that explicit information and inferences that are relevant to a person's goals would have a privileged status

compared to irrelevancies. If we take orienting questions as a type of goal for reading, then eye tracking studies confirm that relevant information is inspected more prevalently than irrelevant information (Lewis & Mensink, 2012; Graesser & Lehman, 2012; McCrudden, Schraw, & Kambe, 2005; Reynolds & Anderson, 1982; Rothkopf & Billington, 1979; Wiley et al., 2009) and inferences follow this direction as well (Graesser, Baggett, & Williams, 1996; Narvaes, van den Broek, & Ruiz, 1999). Such results confirm theory and intuitions. Although that is the case, the claim lacks teeth unless there are systematic mechanisms that generate the goals and questions. A theory of goal and question generation is needed before this claim is theoretically interesting.

We contend that two mechanisms are at the heart of generating questions and goals: cognitive disequilibrium and genre. We believe that these two mechanisms will fortify the constructionist theory's reader goal assumption over and above the ad hoc and circular prediction that inferences are generated if they are relevant to the idiosyncratic goals of the reader.

A cognitive disequilibrium framework has been sketched to integrate a number of psychological processes: confusion (and other learning-centered emotions), question asking (inquiry), deliberative thought, inferences, and deeper learning. Cognitive disequilibrium is a state that occurs when people face obstacles to goals, interruptions, contradictions, incongruities, anomalies, impasses, uncertainty, and salient contrasts (Barth & Funke, 2010; D'Mello & Graesser, 2012; Festinger, 1957; Graesser, Lu, Olde, Cooper-Pye, & Witten, 2005). Initially the person experiences various emotions when beset with cognitive disequilibrium, but notably confusion, surprise, or curiosity (D'Mello & Graesser, 2012; Graesser & D'Mello, 2012; Lehman, D'Mello & Graesser, 2012). This elicits question asking and other forms of inquiry (Graesser & McMahen, 1993; Graesser, Lu et al., 2005; Otero & Graesser, 2001), such as social

interaction, physical exploration of the environment, the monitoring of focal attention, and inference generation. Why does the impasse occur? What can be done? The person engages in inference generation, problem solving, reasoning, and other thoughtful cognitive activities in an attempt to resolve the impasse and restore cognitive equilibrium.

Breaks in cohesion breaks, contradictions, and misinformation trigger such questions and inferences during text comprehension for proficient readers. Discourse processing studies have confirmed that additional time for inference generation and elaboration occurs when an event, action, or state in the text contradicts earlier information (Blanc, Kendeou, van den Broek, & Brouillet, 2008; Cook, Halleran, & O'Brien, 1998; Hyona, Lorch, & Rinck, 2003; Myers & O'Brien, 1998; O'Brien et al., 1998) or when the text statement is incompatible with prior knowledge of a knowledgeable reader (Maier & Richter, 2013; van Oostendorp, 2003; Rapp, 2008). This mechanism is also compatible with the coherence assumption.

The cognitive disequilibrium construct would ideally extend to the reading of multiple texts. That is, readers would discover when the claims or inferences in different texts are compatible (drawing a generalization) or incompatible (creating cognitive disequilibrium). Researchers have recently investigated comparisons among texts in multiple documents when processing the plausibility of claim (Braasch, Rouet, Vibert, & Britt, 2012; Braten & Stromso, 2006; Britt & Rouet, 2012; Goldman et al., 2012; Perfetti, Rouet, & Britt, 1999; Stadtler, Scharrer, Brummernhenrich, & Bromme, 2013; Wiley et al., 2009).At this point in history, many readers have difficulty integrating content from multiple documents and often miss contradictory information. This may change when students are trained to take a more critical stance and acquire a more fine-tuned palate on intertextuality. Contradictions in text and conversation are most likely to be detected when they are contiguous in time, discourse, and working memory.

Text genre is the other component that gives the reader goal assumption some teeth in making discriminating predictions. There are salient questions that are associated with different text genres and these questions go a long way in predicting the inferences that readers generate (Graesser & Lehman, 2012). In a narrative text, the relevant questions during the setting are Who?, What?, Where?, and When? but the questions shift to Why? and So What? when the plot occurs. In persuasive texts, the underlying questions: What does the writer believe?, Why is the writer telling me this?, What is the argument?, and Do I agree with the writer? When the text has the function of giving directions to a destination (i.e., instructions in a map to a party), the relevant questions are: Where is the destination? and How do I get there? When the text is a procedure or recipe, the questions are: How do I enact the procedure?, What can go wrong?, and How do I know if I succeed? A text on purchasing a car would have questions like How is car X similar or different than car B? A text in a claim+evidence frame would have questions such as: What is the claim? How does X support the claim? Why do people doubt the claim? There are of course other questions for other genre. We propose that there are a limited number of genres in a culture, there are distinctive questions associated with each genre, and these questions drive the goals and inferences that readers encode during comprehension. Our conclusions are of course quite plausible, but empirical evidence is needed to support them.

Coherence and Explanation Assumptions

The constructionist model offered some discriminating predictions about the classes of inferences that are consistently, if not automatically and unconsciously, encoded during comprehension. Table 1 lists the predictions for 12 categories of inferences and the status of whether the inference is generated. Of course, any of these inferences could be generated if the reader adopts idiosyncratic goals that target a particular class of inferences (e.g., the reader is

tracking the personality of one of the characters) or if the inference is highly activated by idiosyncratic prior experiences (e.g., the reader's occupation matches that of a particular character). The predictions in Table 1 reflect the coherence and explanation assumptions of the constructionist model.

Table 1 specifies that only 5 out of the 12 classes of inferences are encoded during comprehension. The referential (class 1) and causal antecedent inferences (class 2) are constructed by virtue of the coherence assumption. The causal antecedent (class 2), character emotional reactions (class 4), superordinate goals (class 5), and thematic inferences (class 10) are constructed by virtue of the explanation assumption. The latter are answers to why-questions:

Why did the event occur? (class 2, causal antecedent)

Why did the character do something? (class 5, superordinate goal)

Why did the author write this? (class 10, thematic)

The emotional reactions (class 4) are predicted to be constructed by virtue of the intuition that emotional reactions of characters (e.g., happiness, anger, fear) are motivated by the goals of other characters who try to elicit the emotions.

The remaining classes of inferences are not predicted to be routinely generated during the comprehension of narrative texts. Some are mere elaborations of nouns (classes 8 and 9) and character actions (classes 6 and 7) that are not sufficiently constrained by context. Causal consequences were not generated because most predictions about the future plot do not end up being correct because they are insufficiently constrained by prior context (Graesser, 1981).The author's intent (class 11) and emotions of the reader (class 12) tap the communicative exchange between the author and reader, once again being insufficiently constrained by context. To most readers, the author is invisible.

It is noteworthy that other models of inference generation make rather different predictions than the constructionist model, although there is some overlap in the predictions. The minimalist position predicts classes 1 and 2, but not 3 through 12.Interestingly, 9 of 12 of the predictions overlap between the constructionist and minimalist positions. The embodiment and indexical position predict that the manner in which actions are executed are generated during comprehension (e.g., the style and path in which a gangster drives to the hideout in a mystery), classes 6 and 7, but that is irrelevant information according to the constructionist model. It is difficult to know how many of the 12 predictions match with the constructionist model for the latter model because the embodiment and indexical hypothesis is silent on so many predictions. Quite clearly, it is difficult to compare the models when the scope and decisiveness of the predictions vary.

There are challenges in comparing models when the models are progressively more statistical and complex. The construction-integration model (Kintsch, 1998) and the landscape model (Van den Broek et al., 1996) are the most quantitatively sophisticated models for comparison. They have a suite of parameters in the models that generate the activation levels of potential inferences at various points in time as texts are read, sentence by sentence. That quantitative infrastructure presents a foundation for generating quantitative predictions, but the down side is that the predictions are dependent on the parameters and any judgments that the human modelers smuggle into the formulae. These limitations compromise the decisiveness of predictions. To compare models, a researcher would need to model the degree to which a set of inferences (I_1, I_2,… I_n) had corresponding encoding values (A_1, A_2,… A_n) that fit a set of performance measures on the inferences (M_1, M_2,… M_n), such as word naming latencies or recall proportions. The predictions of each model would need to be compared on an even playing field.

A large space of parameters would need to be tried out in order to give each model its fair test. Such endeavors require quite a bit of work. The bar may simply be too high.

One criticism of the constructionist theory is that it gave discrete predictions of what inference classes were versus were not generated rather than giving a continuous set of values. Our reply has always been that our critics are correct, but we were at a pre-paradigmatic stage of research when we were trying to get an approximate handle on prospects of likely inferences. We were never extremely worried about this criticism because there is a long history of comparisons between discrete and continuous models in mathematics and computational sciences. Models in the two traditions are essentially interchangeable. One can test a discrete model in which inferences are encoded if they are activated to the point of meeting a distribution of thresholds; the researcher can vary the thresholds and inspect the output. Interpreting such data is comparable to interpreting the activation values of a continuous model.

One of the luxuries of developing a model is that you have the opportunity to be its most incisive critic. Doubts about our constructionist theory arose in our research teams at the turn of the millennium. The doubts were not prompted by details of experimental findings or quantitative modeling fits, both of which could be salvaged by creative interpretations of the data. There were two major sources of doubts. The first doubt and worry is that our findings were restricted to narrative texts and could not scale up to other text genres. For example, the minimalist position trumped the constructionist position when we investigated expository texts on physic (Graesser et al., 2012; Van Lehn et al., 2007) and computer literacy (Graesser et al., 2004). In essence, when students read textbooks on these subject matters and were later given tests on inferences, the performance on such tests were no different between a condition in which college students read text and a condition when they did nothing. As another example, the

embodiment and indexical model (Glenberg, 1997; Glenberg & Robertson, 1999) was likely to trump the constructionist position when the comprehender had to understand directions to execute a procedure, assemble a piece of equipment, or understand the directions to a destination. Subordinate goals (class 6), instruments (class 7), and visual-spatial states (class 9) were likely to play a more salient role in these context. We could of course salvage the constructionist model by appealing to reader goals, but that would open the door to an endless array of ad hoc predictions. We fundamentally recognized the importance of text genre and that the constructionist model could not explain the differences among genre.

These concerns never led us to doubt the prominence of coherence and explanation in inference mechanism, however. Coherence has a robust impact on constraining what inferences are encoded versus merely activated (Kintsch, 1998). Causal explanations play a fundamental role in constraining the inferences that are generated during comprehension (Briner, Virtue, & Kirby, in review; Graesser & Bertus, 1998; Millis & Graesser, 1994; Singer et al., 1992; van den Broek et al., 1996; van den Broek, Rapp, & Kendeou, 2005; Zwaan & Radvansky, 1998) and that inoculate the reader from accepting and remembering false claims in texts (Kendeou, Smith, & O'Brien, 2013).Coherence and explanation continue to be accepted as powerful constructs in contemporary research in discourse psychology.

The second doubt arose when we considered scaling up our findings to the real world. The NRC Reading for Understanding report (Snow, 2002) emphasized the importance of considering the texts, tasks, and reader in a sociological context when improving reading comprehension. Unfortunately, we found our field of discourse processing in need of improvement in considering all four of these components. There needed to be better

understanding of the variability of texts, tasks, and readers. Our approach to this challenge was to develop technologies to help us analyze the landscape.

Language and Discourse Technology

This is a unique point in history because there is widespread access to hundreds of computer tools that analyze specific texts and large text corpora. Advances in computational linguistics (Jurafsky & Martin, 2008), statistical representations of world knowledge (Landauer, McNamara, Dennis, & Kintsch, 2007), and corpus analyses (Biber, Conrad, & Reppen, 1998) have allowed us to analyze texts on objective criteria and thereby provide a broader and more precise analysis of text characteristics.

Our research team has devoted considerable effort in using technologies to analyze characteristics of texts and to develop tasks (interventions) that promote both learning and assessment of comprehension. This section provides highlights of these two efforts. Our efforts have implications for scaling readers on various abilities and for automatically analyzing sociocultural contexts, but these aspects will not be addressed in this chapter.

Automated Scaling Texts on Multiple Levels of Language and Discourse

We developed Coh-Metrix to scale texts on hundreds of dimensions of language and discourse (Graesser, McNamara, Louwerse, & Cai, 2004; Graesser, McNamara, & Kulikowich, 2011; McNamara, Graesser, McCarthy, & Cai, 2014). The original purpose of the Coh-Metrix project was to concentrate on the cohesion of the textbase, the coherence of the situation model, and discourse genre because those levels needed a more precise specification. However, we quickly discovered the need to also measure texts on characteristics of words and sentence syntax. Two versions of Coh-Metrix are available for the public for free on the web: The original

version with over 100 measures (http://www.cohmetrix.com) and a version with a handful of major dimensions called the *Text Easability Assessor* (http://tea.cohmetrix.com).

The original version of Coh-Metrix had nearly a thousand measures, 100 of which were put on the web site for colleagues to use. However, we were encouraged by researchers and practitioners to simplify the analysis and converge on a smaller number of factors. Therefore, a principle components analysis was performed on 37,520 texts in order to identify central constructs of text complexity (Graesser, McNamara, & Kulikowich, 2011). The PCA resulted in 8 dimensions that accounted for 67% of the variance in variations among texts. The top 5 of these dimensions were incorporated in TEA-Coh-Metrix. The five dimensions of TEA-Coh-Metrix have also been validated in a comprehensive analysis of texts in the Common Core of the National Governors Association (2010, http://www.corestandards.org) and various high-stakes assessments in the United States (Nelson, Perfetti, Liben, & Liben, 2012).The five major dimensions are listed and defined below.

1. **Narrativity.** Narrative text tells a story, with characters, events, places, and things that are familiar to the reader. Narrative is closely affiliated with everyday oral conversation.
2. **Referential cohesion**. High cohesion texts contain words and ideas that overlap across sentences and the entire text, forming threads that connect the explicit text together for the reader.
3. **Situation model cohesion.** Causal, intentional, and temporal connectives help the reader to form a more coherent and deeper understanding of the text.
4. **Syntactic simplicity**. Sentences with few words and simple, familiar syntactic structures are easier to process and understand. Complex sentences have structurally embedded syntax.

5. **Word concreteness**. Concrete words evoke mental images and are more meaningful to the reader than abstract words.

Each of the 5 dimensions above is expressed in terms of ease of comprehension. Text complexity is defined as the opposite of ease, so principal component scores are reversed in measures of text complexity.

The five Coh-Metrix dimensions have been correlated with unidimensional metrics of text complexity, such as Flesh-Kincaid, DRP, and Lexiles. If we use grade level and text genre (i.e., narrative versus informational texts) as a yardstick, several trends support the claim that researchers should consider multiple levels and resist the temptation to settle for a single dimension of text complexity (Graesser & Li, 2013; Graesser, McNamara, & Kulikowich, 2011). For example, narrativity and syntactic simplicity robustly decrease as a function of grade level, and word concreteness moderately decreases also. However, the correlation between grade level and cohesion is extremely small and sometimes not statistically significant. Apparently, cohesion is not on the radar of the standard readability metrics, even though discourse processing researchers have established that cohesion is an important predictor of reading time and comprehension (Goldman et al., 1999; Lorch & O'Brien, 1995; McNamara, Louwerse, McCarthy, & Graesser, 2010). There were also trade-offs among the different levels as we correlated the texts in different age groups and genres. For example, informational texts (non-narratives) typically are on topics that readers are less familiar with. These informational texts tend to have higher referential cohesion and simpler syntax than the narrative texts, perhaps because good writers compensate for the difficulty of the subject matter. Narrative texts tend to occur at earlier grade levels whereas informational texts at later grade levels. Therefore, any

analysis of texts at different grade levels needs to understand the tradeoffs among narrativity, cohesion, and syntax.

Coh-Metrix is a useful tool to sort out complex interactions among text constraints and data reported in laboratory experiments. As one example, McNamara, Louwerse, McCarthy, and Graesser (2010) analyzed the stimulus materials of experimental studies in discourse processing that investigated the impact text coherence/cohesion on measures such as reading time and recall. The researchers manipulated text cohesion by adding connectives or referring expressions to explicitly bridge text constituents rather than relying on the readers to fill in the connections inferentially. It is conceivable that there can be trade-offs in such manipulations. For example, adding connectives to link sentences can perhaps help cohesion, but it can also add to comprehension difficulty by virtue of sentence length and syntactic complexity. Replacing pronouns with nouns and rich referring expressions can perhaps help cohesion, but it adds to difficulty by lowering word frequency (i.e., nouns are less frequent than pronouns) and increasing noun-phrase density (i.e., with adjective modifiers).

Coh-Metrix allowed us to measure the impact of text manipulations on multiple levels of language/discourse and to track the fallout of such manipulations. We contend that these objective metrics from Coh-Metrix should be routinely used in experimental investigations of inferences in discourse.

Another example of the utility of Coh-Metrix to discourse researchers is to compare experimental texts to the norms of texts in different genres, such as narrative, science, social studies, and so forth (Graesser, McNamara, & Kulikowich, 2011; McNamara et al., 2014). A researcher would expect a large number of inferences to be generated in narrative texts with high frequency words, but fewer inferences in science texts with a technical vocabulary. Coh-Metrix

can be used to assess texts in experiments in order to judge whether readers are likely to draw inferences and to confirm that texts are in particular genres (for example, see Eason, Goldberg, Young, Geist, & Cutting, 2012).

It is beyond the scope of this chapter to describe the computational linguistics modules that were incorporated in the Coh-Metrix tool. This information is provided in other reports (Graesser, McNamara et al., 2004; McNamara et al., 2014) and there is a help system in the web facility. However, Table 2 presents a sample of example measures that went into analyzing texts at multiple levels of language and discourse, namely words, syntax, textbase referential cohesion, situation model coherence, and genre (Graesser & McNamara, 2011).

INSERT TABLE 2 ABOUT HERE

Technologies that Promote Learning and Assessment of Comprehension

Learning from texts for school and lifelong learning is one of the many authentic tasks to consider as one moves out of the experimental lab into the real world. This has been the direction of the research teams in Memphis during the last two decades. More specifically, we have investigated inferences during comprehension, reasoning, and problem solving activities that are associated with learning difficult subject matters, such as computer literacy, physics, biology, and scientific methods. Computer technologies have been developed to (a) track inferences, comprehension, and other psychological states while studying the material and (b) provide interventions to facilitate comprehension and learning of the difficult subject matter.

Coupling reading times with text difficulty. One simple technology collect self-paced reading times while students read texts and compares these times with the text difficulty metrics provided by Coh-Metrix (Vega, Feng, Lehman, Graesser, & D'Mello, 2013). Screens of text of approximately 100 words are scaled on Flesch-Kincaid scores and various dimensions of Coh-

Metrix, based on the TASA norms collected from 37,520 texts that are representative of the texts that college students experience throughout through lifetime (Graesser, McNamara, & Kulikowich, 2011; McNamara et al., 2014).Engagement during reading is manifested by a close correspondence between the self-paced reading times of screens of text (converted to z-scores for each individual reader) and the difficulty of the texts on the various metrics (converted to z-scores based on the TASA norms).An engaged reader should speed up on easy text and slow down on difficult text. Low engagement with the text is manifested by a decoupling between the time spent reading and the text difficulty profile (Franklin, Smallwood , & Schooler, 2011; Schooler et al., 2011; De Vega et al., 2013). More specifically, this decoupling may be explained by either mind wandering (Feng, D'Mello, & Graesser, 2013) or by thoughtful reflection about difficult ideas expressed in the text. For example, good comprehenders are expected to slow down and reason when they encounter contradictions between two or more ideas in the text (Baker, 1985; O'Brien, Rizzella, Albrecht, & Halleran, 1998) or by claims in the text that clash with world knowledge (Kendeou et al., 2012; Rapp, 2008). Reading times that are much higher than the projections of the Coh-Metrix metrics at these points of contradictions or inconsistencies in the text would be signals of deep comprehension whereas reading times shorter than the projections would be signals of shallow comprehension.

Tracking correct inferences and misconceptions from natural language. Automated essay scoring has now reached a level of accuracy that the scoring of many classes of written essays is as accurate as expert human raters (Attali & Burstein, 2006; Graesser & McNamara, 2012; Landauer, Laham, & Foltz, 2003; Shermis et al., 2010). These systems have had exact agreements with humans on a 5-point scale as high as the mid-80's, adjacent agreements in the high mid-90's, and correlations as high as the mid-80's. These performance measures are slightly higher than agreement between trained human raters. The *Intelligent Essay Assessor* (Landauer

et al., 2003) analyzes the words in the essay with latent semantic analysis (LSA, Landauer et al., 2007) and also sequences of words with an n-gram analysis (e.g., word pairs, word triplets). The algorithm computes the similarity of the words and word sequences between a new essay and the essays associated with each level of a scoring scale.

LSA is an important method of computing the conceptual similarity between words, sentences, paragraphs, or essays because it considers implicit knowledge. It is a mathematical, statistical technique for representing knowledge about words and the world on the basis of a large corpus of texts that attempts to capture the knowledge of a typical human. The central intuition of LSA is that the meaning of a word W is reflected in the company of other words that surround word W in naturalistic documents (imagine 37,520 texts and 11 million words). Two words are similar in meaning to the extent that they share similar surrounding words. For example, the word *cup* is highly associated with words in the same functional context, such as *glass*, *liquid*, *pour*, *handle*, *coffee*, and *heat*. These are not synonyms or antonyms that would occur in a dictionary or thesaurus. LSA uses a statistical technique that condenses a very large corpus of texts to 100-500 statistical dimensions (Landauer et al., 2007). The conceptual similarity between any two text excerpts (e.g., word, clause, sentence, entire essay) is computed as the geometric cosine between the values and weighted dimensions of the two text excerpts. The value of the cosine typically varies from approximately 0 to 1.

The accuracy of scoring short verbal responses has also improved with advances in computational linguistics and statistical representations of world knowledge like LSA. Short verbal responses by students can vary from one word to 2-3 sentences. In all of these automated assessments, the verbal input of the student is compared with an expected answer. The expected answer may be a word, a set of alternative words, a sentence, a set of sentences, or a pattern of

symbolic expressions (Cai et al., 2011; Lealock & Chadorow, 2003; Magliano & Graesser, 2012). The expected answer may be either correct (called an expectation) or incorrect (a bug or misconception). It may be either an explicit sentence in the text or an inference. A semantic or conceptual match is computed between the student input and anticipated answers, with match scores that vary between 0 and 1.Matches to single words are not particularly challenging, whereas matches to more complex expressions are computed by a variety of models. Technologies like *C-Rater* developed at Educational Testing System (Lealock & Chadorow, 2003) score answers to short-answer questions that extend beyond single words. AutoTutor (Graesser et al., 2012) and Operation ARA (Cai et al., 2011) track the verbalizations of students over many conversational turns with conversational computer agents (as discussed below) and compare this student input with sentential expectations (either inferences or explicit sentences in a text). The performance of these systems can be quite impressive. For example Cai et al. (2011) analyzed the student responses in conversational turns in Operation ARA, comparing student verbal responses and expectations. The match scores of trained judges correlated .69 whereas the match scores between the computer scores and human judges was .67.The Cai et al. algorithm used a combination of LSA and regular expressions (see Jurafsky & Martin, 2008 for a definition of regular expressions).

Tracking comprehension strategies. It is possible to track inferences and strategies during comprehension from verbal protocols while students comprehend text (Kurby & Zachs, 2013; Magliano, Trabasso & Graesser, 1999; Millis & Magliano, 2003).For example *RSAT (Reading Strategy Assessment Tool)* was developed to identify the comprehension strategies that are manifested in think-aloud protocols that students type in (or say aloud) while reading texts (Magliano, Millis, The RSAT Development Team, Levinstein, & Boonthum, 2011; Millis &

Magliano, 2013). One important comprehension strategy measured by RSAT is to identify content that reflects causal connections or *bridges* between clauses in the text. RSAT distinguishes between local and distal bridges. Local bridges occur between the target sentence and the immediately prior sentence. Distal bridges occur between the target sentences and sentences located two or more sentences back. Skilled readers are more likely to make distal bridges whereas less-skilled readers tend to focus more on the immediate context surrounding each sentence (Coté, & Goldman, & Saul, 1998). Another type of strategy is *elaboration*. Elaborative inferences are constructed in a fashion that caters to the constraints of the text but also recruits relevant world knowledge (McNamara & Magliano, 2009). Unlike bridges, elaborations do not connect sentences. A third strategy is *paraphrasing*. The student articulates explicit text information but in slightly different words. There is some evidence that the amount of paraphrasing in verbal protocols is negatively correlated with comprehension whereas bridging and elaborating is positively correlated (Magliano & Millis, 2003). The match scores that accumulate in the think aloud protocols allow the RSAT analyzer to infer the strategies of the reader.

Tracking affective states. Automated technologies are capable of tracking of emotions and other affective states when reading technical texts (Calvo & D'Mello, 2010; D'Mello, Dowell, & Graesser, in press) and when being tutored by humans or computers (D'Mello & Graesser, 2010; Graesser & D'Mello, 2012). The common affective states during learning are boredom, engagement/flow, frustration, confusion, delight, and surprise in a wide range of learning environments (Baker, D'Mello, Rodrigo, & Graesser, 2010; D'Mello & Graesser, 2012; Graesser & D'Mello, 2012). The learning-centered affective state that best predicts learning at deeper levels is confusion, a cognitive-affective state associated with thought and deliberation (D'Mello

& Graesser, 2012; D'Mello, Lehman, Pekrun, & Graesser, in press; Lehman et al., 2013). Confusion can be detected automatically by the patterns in tutorial dialog, body posture, and facial movements (D'Mello & Graesser, 2010) in addition to the decoupling profiles of reading time discussed above. Good comprehenders are expected to exhibit signals of confusion when encountering contradictions and false claims during reading (Lehman et al., 2013) whereas these signals should be less prevalent for shallow comprehenders.

Conversational pedagogical agents. Research teams at the University of Memphis have developed and tested conversational pedagogical agents (talking heads) to improve learning of technical material on subject matters in Science, Technology, Engineering, and Mathematics STEM). It is difficult for students to generate inferences while reading and studying texts in these STEM areas so the pedagogical agents scaffold them to promote deeper comprehension and learning. As discussed earlier, students acquire explicit information when reading texts on computer literacy, physics, and research methods but their performance on inference questions is no different than when they read nothing (Graesser, Lu et al., 2004; VanLehn et al., 2007). Learning environments with pedagogical agents have been developed to serve as substitutes for humans who range in expertise from peers to subject matter experts with training in tutoring (Graesser, Conley, & Olney, 2012).

Pedagogical agents can be designed to perform a large range of tasks that human tutors and peers can do.Agents can guide the interaction with the learner, tutor them with dialogs in natural language, instruct the learner what to do next, and interact with other agents to model ideal behavior, strategies, reflections, and social interactions. Some agents generate speech, gestures, facial expressions, and body movements in ways similar to people, as exemplified by *Betty's Brain* (Biswas, Jeong, Kinnebruw, Sulcer, & Roscoe, 2010), *Tactical Language and Culture*

System (Johnson & Valente, 2008), *iSTART* (McNamara, Boonthum, Levinstein, & Millis, 2007; McNamara, O'Reilly, Best, & Ozuru, 2006), *Crystal Island* (Rowe, Shores, Mott, & Lester, 2010), and *My Science Tutor* (Ward et al., 2011). Systems like *AutoTutor* and *Why-Atlas* can interpret the natural language of the human that is generated in spoken or typed channels and can respond adaptively to what the student expresses (Graesser et al., 2012; D'Mello, Dowell, & Graesser, 2011; VanLehn et al., 2007).These agent-based systems have frequently demonstrated value in improving students' learning and motivation (Graesser, Conley, & Olney, 2012; VanLehn, 2011) but it is beyond the scope of this chapter to review the broad body of research on agents in learning environments.

Training reading strategies with agents. Pedagogical agents can be used to directly train students how to implement comprehension strategies, including inferences. *iSTART (Interactive Strategy Trainer for Active Reading and Thinking)* is one of these systems that has successfully improved comprehension strategies (McNamara, O'Reilly, Rowe, Boonthum, & Levinstein, 2007; McNamara, O'Reilly, Best, & Ozuru, 2006). The iSTART trainer is designed to help students become deeper text comprehenders by constructing self-explanations of the text. The construction of self-explanations during reading is known to facilitate deep comprehension when there is some context-sensitive feedback on the explanations that get produced (Chi, de Leeuw, Chiu, & LaVancher, 1994; McNamara, 2004).The iSTART interventions focus on five reading strategies that are designed to enhance self-explanations, construction of inferences, and interpreting the explicit text: *monitoring comprehension* (i.e., recognizing comprehension failures and the need for remedial strategies), *paraphrasing* explicit text, making *bridging inferences* between the current sentence and prior text, making *predictions* about the subsequent text, and *elaborating* the text with links to what the reader already knows. The accuracy of

applying these inferential and metacognitive skills is measured and tracked throughout the tutorial session.

Groups of agents scaffold these strategies in three phases of iSTART training. In an *Introduction Module*, an instructor and two student agents collaboratively describe self-explanation strategies with each other. In a *Demonstration Module*, two agents in a trialog demonstrate the use of self-explanation in the context of a science passage and then the student identifies the strategies being used. The accuracy of the students' identifying the correct strategy is exhibited by the student agent serves as a measure of understanding the inference and metacognitive strategies. In a final *Practice* phase, an agent coaches and provides feedback to the student one-to-one while the student practices self-explanation reading strategies. For particular sentences in a text, the agent reads the sentence and asks the student to self-explain it by typing a self-explanation. The iSTART system then attempts to interpret the student's contributions, gives feedback, and asks the student to modify unsatisfactory self-explanations.

Training metacognition and self-regulated learning with agents. *MetaTutor* trains students on 13 strategies that are theoretically important for self-regulated learning (Azevedo, Moos, Johnson, & Chauncey, 2010).The process of self-regulated learning (SRL) involves the learners' constructing a plan, monitoring metacognitive activities, implementing learning strategies, and reflecting on their progress and achievements. Inferences are required in all of these processes. The MetaTutor system has a main agent (Gavin) that coordinates the overall learning environment and three satellite agents that handle three phases of SRL: Planning, monitoring, and applying learning strategies. Each of these phases can be decomposed further, under the guidance of the assigned conversational agent. For example, metacognitive monitoring is decomposed into judgments of learning, feeling of knowing, content evaluation, monitoring

the adequacy of a strategy, and monitoring progress towards goals. Learning strategies include searching for relevant information in a goal-directed fashion, taking notes, drawing tables or diagrams, re-reading, elaborating the material, and coordinating information sources (text and diagrams). Each of these metacognitive and SRL skills have associated measures that are based on the student's actions, decisions, ratings, and verbal input. The frequency and accuracy of each measured skill is collected throughout the tutoring session and hopefully increases as a function of direct training.

Trialogs in scientific reasoning. Trialogs involve two agents interacting with the human learner. The agents can take on different roles, such as tutor and student or two peer students. These trialogs can be implemented in junction with the human reading texts. Our research team has conducted several studies with trialogs in the context of critiquing case studies of scientific research with respect to scientific methodology. The design of the case studies and trialog critiques were adapted from an educational game called *Operation ARIES!* (Forsyth et al., 2013; Millis et al., 2011), which was subsequently commercialized by Pearson Education as *Operation ARA* (Halpern et al., 2012). Players learn how to critically evaluate research they read in magazines and newspapers. A series of cases were presented to the student that described experiments that have a number of flaws with respect to scientific methodology. For example, one case study described a new pill that purportedly helps people lose weight, but the sample size was small and there was no control group. The goal of the participants in the trialog is to identify the flaws and express them in natural language. The game was designed to teach high school or college students how to critically apply principles of good methodology in scientific investigations (e.g., the need for control groups, adequate samples of observations, operational

definitions, differentiating correlation from causation, etc.).These activities require inferences and reasoning in addition to the interpretation of explicit text.

A series of studies were conducted that planted false information and contradictions in the trialogs as case studies were critiqued (D'Mello, Lehman, Pekrun, & Graesser, in press; Lehman, D'Mello, & Graesser, 2012; Lehman et al., 2013). We attempted to induce cognitive disequilibrium by manipulating whether or not the tutor agent and the student agent contradicted each other during the trialog and expressed points that are incorrect. Each case study had a description of a research study that was to be critiqued during the trialogs. Then the tutor agent and student agent engaged in a short exchange about (a) whether there was a flaw in the study and (b) which part of the study was flawed if there was a flaw. In the *True-True* control condition, the tutor agent expressed a correct assertion and the student agent agreed with the tutor. In the *True-False* condition, the tutor expressed a correct assertion but the student agent disagreed by expressing an incorrect assertion. In the *False-True* condition it was the student agent who provided the correct assertion and the tutor agent who disagreed. In the *False-False* condition, the tutor agent and student agent agreed about an incorrect assertion. The human student was asked to intervene after particular points of possible contradiction in the conversation. For example, the agents turned to the human and asked "Do you agree with Chris (student agent) that the control group in this study was flawed?" The human's response was coded as correct if he/she agreed with the agent who had made the correct assertion about the flaw of the study. If the human experienced uncertainty and was confused, this should be reflected in the incorrectness and/or uncertainty of the human's answer. This uncertainty would ideally stimulate thinking and learning.

The data indeed confirmed that the contradictions and false information had an impact on the humans' answers to these Yes-No questions immediately following a contradiction. The proportion of correct student responses showed the following order: True-True > True-False > False-True > False-False conditions. These findings indicated that learners were occasionally confused when both agents agreed and were correct (True-True, no contradiction), became more confused when there was a contradiction between the two agents (True-False or False-True), and were either confused or simply accepted the incorrect information when the agents incorrectly agreed (False-False). Confusion was best operationally defined as occurring if both (a) the student manifested uncertainty/incorrectness in their decisions when asked by the agents and (b) the student either reported being confused or the computer automatically detected confusion (D'Mello & Graesser, 2010; Graesser & D'Mello, 2012).Interestingly, the results of these studies revealed that contradictions, confusion, and uncertainty caused more learning at deeper levels of mastery, as reflected in a delayed test on scientific reasoning and far-transfer case studies. There may be a causal relationship between contradictions (and the associated cognitive disequilibrium) and deep learning, with confusion playing either a mediating, moderating, or causal role in the process.

There was a need for the two agents to directly contradict each other in a conversation before the humans experienced an appreciable amount of uncertainty and confusion. We suspect that the contradiction would need to be contiguous in time before the contradiction would be detected. That is, the contradiction is likely to be missed if one agent makes a claim and then another agent makes a contradictory claim 10 minutes later. This is compatible with research in text comprehension that has shown that the contradictory claims must be co-present in working memory before they get noticed unless there is a high amount of world knowledge. It is also

compatible with the observation that it is difficult for many students to integrate information from multiple texts and spot contradictions (Bråten, Strømsø, Britt, 2009; Britt & Aglinskas, 2002; Goldman et al., 2012; Rouet, 2006) unless there is a high amount of world knowledge.A strategic attempt to integrate information from multiple texts would be needed to draw such connections unless the person is fortified with sufficient subject matter knowledge (Goldman et al., 2012; Graesser et al., 2007; Wiley et al., 2009).

Closing Comments

Research on inference generation has evolved considerably since we started conducting laboratory studies on short texts three decades ago. Discourse psychologists have a better understanding of the types of inferences that are routinely generated while comprehending different types of texts and of the psychological mechanisms that generate such inferences. All of the models postulate that subject matter knowledge and memory robustly guide these processes. However, such knowledge and memory are not enough. Competent readers in the 21st century also need strategies of language comprehension, discourse, critical thinking, and self-regulated learning. The three signature assumptions of the constructionist model, namely reader goals, coherence, and explanation, will no doubt be important components of the manifold of strategies.

Acknowledgements

This research was supported by the National Science Foundation (0834847, 0918409, 0904909, 1108845) and the Institute of Education Sciences (R305G020018, R305H050169, R305B070349, R305A080589, R305A100875, R305C120001). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF and IES.

References

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater R V.2. *Journal of Technology, Learning and Assessment, 4,* 1-30.

Azevedo, R., Moos, D., Johnson, A., Chauncey, A. (2010). Measuring cognitive and metacognitive regulatory processes used during hypermedia learning: Issues and challenges. *Educational Psychologist, 45,* 210-223.

Baker, L. (1985). Differences in standards used by college students to evaluate their comprehension of expository prose. *Reading Research Quarterly*, *20*, 298-313.

Baker, R.S., D'Mello, S.K., Rodrigo, M.T., & Graesser, A.C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68,* 223-241.

Barth, C.M., Funke, J. (2010). Negative affective environments improve complex solving performance. *Cognition and Emotion*, *24*, 1259-1268.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biswas, G., Jeong, H., Kinnebrew, J., Sulcer, B., & Roscoe, R. (2010). Measuring self-regulated learning skills through social interactions in a teachable agent environment. *Research and Practice in Technology-Enhanced Learning, 5*, 123-152.

Blanc, N., Kendeou, P., van den Broek, P., & Brouillet, D. (2008). Updating situation models during reading of news reports: Evidence from empirical data and simulations. *Discourse Processes*, *45*, 103-121.

Braasch, J. L., Rouet, J. F., Vibert, N., & Britt, M. A. (2012). Readers' use of source information in text comprehension. *Memory & cognition*, *40*, 450-465.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn* (expanded ed.). Washington, D.C.: National Academy Press.

Bråten, I., Strømsø, H. I., & Britt, M. A. (2009). Trust matters: Examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly, 44*, 6-28.

Braten, I., & Stromso, H. (2006). Constructing meaning from multiple information sources as a function of personal epistemology. *Information Design Journal*, *14*, 56-67.

Briner, S.W., Virtue, S., & Kurby, C. A. (in review). Forward and backward causal relations in narrative text. *Discourse Processes*

Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction, 20*, 485–522.

Britt, M. A., & Rouet, J. F. (2012). Learning with multiple documents: Component skills and their acquisition. In M. J. Lawson & J. R. Kirby (Eds), *The quality of learning: Dispositions, instruction, and mental structures* (pp. 385-404). Cambridge: Cambridge University Press.

Britton, B. K., & Graesser, A.C. (1996) (Eds.). *Models of understanding text*. Hillsdale, NJ: Erlbaum.

Cai, Z., Graesser, A.C., Forsyth, C., Burkett, C., Millis, K., Wallace, P., Halpern, D. & Butler, H. (2011). Trialog in ARIES: User input assessment in an intelligent tutoring system. In W. Chen & S. Li (Eds.), *Proceedings of the 3rd IEEE International Conference on Intelligent Computing and Intelligent Systems* (pp. 429-433). Guangzhou: IEEE Press.

Calvo, R. A., & D'Mello, S. K. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*, 18-37.

Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439-477.

Cook, A. E., Halleran, J. G., & O'Brien, E. J. (1998). What is readily available during reading? A memory-based view of text processing. *Discourse Processes*, *26*(2-3), 109-129.

Coté, N., Goldman, S.R., & Saul, E.U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes*, *25*, 1-53.

D'Mello, S. K., & Graesser, A. C. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*, 145-157.

D'Mello, S. K., & Graesser, A. C. (2012). Emotions during learning with AutoTutor. In P. J. Durlach and A. Lesgold (Eds.), *Adaptive technologies for training and education* (117-139). Cambridge: Cambridge University Press.

D'Mello, S. K., & Graesser, A.C. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-adapted Interaction, 20*, 147-187.

D'Mello, S. K., Dowell, N. & Graesser, A. C. (in press). Unimodal and multimodal human perception of naturalistic non-basic affective states during human-computer interactions. *IEEE Transactions on Affective Computing*.

D'Mello, S. K., Dowell, N., & Graesser, A. C. (2011). Does it really matter whether students' contributions are spoken versus typed in an intelligent tutoring system with natural language? *Journal of Experimental Psychology: Applied, 17,* 1-17.

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A.C. (in press). Confusion can be beneficial for learning. *Learning and Instruction*.

De Vega, B., Feng, S., Lehman, B., Graesser, A., & D'Mello, S. (2013). Reading into the text: Investigating the influence of text complexity on cognitive engagement. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 296-299).

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader–text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of educational psychology*, *104*, 515.

Feng, S., D'Mello, S. K., & Graesser, A. (2013). Mind wandering while reading easy and difficult texts, *Psychonomic Bulletin & Review*, *20*, 586-592.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Forsyth, C. M., Graesser, A. C. Pavlik, P., Cai, Z., Butler, H., Halpern, D.F., & Millis, K. (2013). Operation ARIES! methods, mystery and mixed models: Discourse features predict affect in a serious game. *Journal of Educational Data Mining, 5,* 147-189.

Franklin, M. S., Smallwood, J., & Schooler, J. W. (2011). Catching the mind in flight: Using behavioral indices to detect mind wandering in real time. *Psychonomic Bulletin & Review, 18*, 992-997.

Gernsbacher, M. A. (Ed.). (1990). *Language comprehension as structure building*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Glenberg, A. M. (1997). What memory is for? *Behavior and Brain Sciences, 20,* 1-55.

Glenberg, A. M., & Robertson, D. A. (1999). Indexical understanding of instructions. *Discourse Processes*, *28*, 1-26.

Goldman, S. R., & Van Oostendorp, H. (1999). Conclusions, conundrums and challenges for the future. In H. Van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 367-376). Mahwah, NJ: Lawrence Erlbaum Associates.

Goldman, S. R., Graesser, A. C., & Van den Broek, P. W. (1999). Reflections. In S. R. Goldman, A. C. Graesser, & P. W. Van den Broek (Eds.), *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (pp. 1-15). Mahwah, NJ: Erlbaum.

Goldman, S.R., Braasch, J.L.G., Wiley, J., Graesser, A.C., & Brodowinska, K. (2012). Comprehending and learning from internet sources: Processing patterns of better and poorer learners. *Reading Research Quarterly, 47*, 356-381.

Graesser, A. & D'Mello, S. K. (in press). *Emotions during the learning of difficult material*. In B. Ross (Ed.), Psychology of Learning and Motivation (Vol. 57): Elsevier.

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., & Bower, G. H. (Eds.). (1990). *The psychology of learning and motivation: Inferences and text comprehension*. New York: Academic Press.

Graesser, A. C., & D'Mello, S. K. (2011). Theoretical perspectives on affect and deep learning. In R. Calvo and S. D'Mello (Eds.). *New perspectives on affect and learning technologies* (pp. 11-21). New York: Springer.

Graesser, A. C., & D'Mello, S. K. (2012). Moment-to-moment emotions during reading. *The Reading Teacher, 66*, 238-242.

Graesser, A. C., & Lehman, B. (2012). Questions drive comprehension of text and multimedia. In M. T. McCrudden, J. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 53-74). Greenwich, CT: Information Age Publishing.

Graesser, A. C., & Li, H. (2013). How might comprehension deficits be explained by the constraints of text and multilevel discourse processes? In B. Miller, L.E. Cutting, & P. McCardle (eds.), *Unraveling reading comprehension: Behavioral, neurobiological, and genetic components* (pp, 33-42). Baltimore: Paul Brookes Publishing.

Graesser, A. C., & McMahen, C. L. (1993). Anomalous information triggers questions when adults solve problems and comprehend stories. *Journal of Educational Psychology*, *85*, 136-151.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, *3*, 371-398.

Graesser, A. C., & McNamara, D. S. (2012). Automated analysis of essays and open-ended verbal responses. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (pp. 307-325). Washington, DC: American Psychological Association.

Graesser, A. C., Baggett, W. B., & Williams, K. (1996). Question-driven explanatory reasoning. *Applied Cognitive Psychology, 10*, 17-31.

Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A web facility that tests question comprehensibility. *Public Opinion Quarterly, 70*(1), 3-22.

Graesser, A. C., Jeon, M., & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes, 45*, 298–322.

Graesser, A. C., Lu, S., Olde, B. A., Cooper-Pye, E., & Whitten, S. (2005). Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down. *Memory and Cognition*, *33*, 1235–1247.

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

Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z., & Hu, X. (2007). Using LSA in AutoTutor: Learning through mixed initiative dialogue in natural language. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 243–262). Mahwah, NJ: Erlbaum.

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

Graesser, A.C., Conley, M., & Olney, A. (2012). Intelligent tutoring systems. In K.R. Harris, S. Graham, and T. Urdan (Eds.), *APA Educational psychology handbook: Vol. 3. Applications to learning and teaching* (pp. 451-473). Washington, DC: American Psychological Association.

Graesser, A.C., Lu, S., Jackson, G.T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M.M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers,* 36, 180-193.

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

Halpern, D. F., Millis, K., Graesser, A. C., Butler, H., Forsyth, C., & Cai, Z. (2012). Operation ARA: A computerized learning game that teaches critical thinking and scientific reasoning. *Thinking Skills and Creativity*, *7*, 93-100.

Hiebert, E. H., & Mesmer, H. A. E. (2013). Upping the ante of text complexity in the Common Core State Standards examining its potential impact on young readers. *Educational Researcher*, *42*, 44-51.

Hyönä, J., Lorch, R. F., Jr., & Rinck, M. (2003). Eye movement measures to study global text processing. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 313-334). Amsterdam: Elsevier Science.

Johnson, L. W. & Valente, A. (2008). Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures. In M. Goker, & K. Haigh (Eds.) *Proceedings of the Twentieth Conference on Innovative Applications of Artificial Intelligence* (pp. 1632-1639). AAAI Press.

Jurafsky, D., & Martin, J.H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction*, *22*, 354-367.

Kendeou, P., Smith, E. R., & O'Brien, E. J. (2013). Updating during reading comprehension: Why causality matters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 854.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kurby, C. A., & Zacks, J. M. (2013). The activation of modality-specific representations during discourse processing. *Brain and language*, *126*, 338-349.

Landauer, T. K., Laham, R. D. & Foltz, P. W. (2003) Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor. In M. Shermis & J. Bernstein, (Eds.). *Automated Essay Scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (2007) (Eds.), *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*, 389-405.

Lehman, B., D'Mello, S. K., & Graesser, A. C. (2012). Confusion and complex learning during interactions with computer learning environments. *Internet and Higher Education*, *15*, 184-194.

Lehman, B., D'Mello, S. K., Strain, A., Mills, C., Gross, M., Dobbins, A., Wallace, P., Millis, K., & Graesser, A. C. (2013). Inducing and tracking confusion with contradictions during complex learning. *International Journal of Artificial Intelligence in Education*, *22*, 85-105.

Lewis, M. R., & Mensink, M. C. (2012). Prereading questions and online text comprehension. *Discourse Processes*, *49*, 367-390.

Lorch, R. F., Jr, & O'Brien, E. J. (Eds.). (1995). *Sources of coherence in reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Macaulay, D. (1988). *The way things work*. Boston: Houghton Mifflin.

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition and Instruction*, *21*, 251-283.

Magliano, J. P., Millis, K. K., Levinstein, I., & Boonthum, C. (2011). Assessing comprehension during reading with the reading strategy assessment Tool (RSAT). *Metacognition and Learning*, *6*, 131-154.

Magliano, J. P., Trabasso, T., & Graesser, A. C. (1999). Strategic processing during comprehension. *Journal of Educational Psychology*, *91*, 615–629.

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student constructed responses. *Behavioral Research Methods, 44*, 608-621.

Maier, J., & Richter, T. (2013). Text belief consistency effects in the comprehension of multiple texts with conflicting information. *Cognition and Instruction*, *31*, 151-175.

McCrudden, M. T., Schraw, G., & Kambe, G. (2005). The effect of relevance instructions on reading time and learning. *Journal of Educational Psychology*, *97*, 88.

McKoon, G., & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1155

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, *38*, 1-30.

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. In B. Ross (Ed.), The p*sychology of learning and motivation* (pp. 297-383). Oxford, UK: Elsevier.

McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: comparing word-based and LSA algorithms. In Landauer, T., D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 227-241). Mahwah, NJ: Erlbaum.

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

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*, 292-330.

McNamara, D. S., O'Reilly, T., Rowe, M., Boonthum, C., & Levinstein, I. B. (2007). iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. In D.S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 397–421). Mahwah, NJ: Erlbaum.

McNamara, D. S., O'Reilly, T., Best, R., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research, 34,* 147-171.

Millis, K., & Graesser, A. C. (1994). The time-course of constructing knowledge-based inferences for scientific texts. *Journal of Memory and Language*, *33*, 583–599.

Millis, K., Forsyth, C., Butler, H., Wallace, P., Graesser, A. C., & Halpern, D. (2011). Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, & J. Lakhmi (Eds.), *Serious games and edutainment applications* (pp.169-196). London: Springer-Verlag.

Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, *26*, 131-157.

Myers, J. L., O'Brien, E. J., Albrecht, J. E., & Mason, R. A. (1994). Maintaining global coherence during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 876-886.

Narvaez, D., van den Broek, P., & Ruiz, A. B. (1999). The influence of reading purpose on inference generation and comprehension in reading. *Journal of Educational Psychology*, *91*, 488.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. New York, NY: Student Achievement Partners.

O'Brien, E. J., Rizzella, M. L., Albrecht, J. E., & Halleran, J. G. (1998). Updating a situation model: A memory-based text processing view. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 24*, 1200-1210.

Otero, J., & Graesser, A. C. (2001). PREG: Elements of a model of question asking. *Cognition & Instruction*, *19*, 143–175.

Perfetti, C. A., Rouet, J. F., & Britt, M. A. (1999). Toward a theory of documents representation. *The construction of mental representations during reading*, 99-122.

Rapp, D. N. (2008). How do readers handle incorrect information during reading? *Memory & Cognition, 36*, 688-701.

Reynolds, R. E., & Anderson, R. C. (1982). Influence of questions on the allocation of attention during reading. *Journal of Educational Psychology*, *74*, 623.

Rothkopf, E. Z., & Billington, M. J. (1979). Goal-guided learning from text: Inferring a descriptive processing model from inspection times and eye movements. *Journal of Educational Psychology*, *71*, 310.

Rouet, J. (2006). *The skills of document use: From text comprehension to web-based learning*. Mahwah, NJ: Erlbaum.

Rowe, J., Shores, L., Mott B., & Lester, J. (2010). Integrating Learning and Engagement in Narrative-Centered Learning Environments. In V. Aleven, J. Kay, J. Mostow (Eds.), *Proceedings of the Tenth International Conference on Intelligent Tutoring Systems* (pp. 166-177). Pittsburgh, Pennsylvania.

Shermis, M.D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw & N. S. Petersen (Eds.), *International encyclopedia of education* (pp. 20-26). Oxford, UK: Elsevier.

Singer, M., Andruslak, P., Reisdorf, P., & Black, N. L. (1992). Individual differences in bridging inference processes. *Memory & Cognition*, *20*, 539-548.

Singer, M., Graesser, A. C., & Trabasso, T. (1994). Minimal or global inference during reading. *Journal of Memory and Language*, *33*, 421-441.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND Corporation.

Stadtler, M., Scharrer, L., Brummernhenrich, B., & Bromme, R. (2013). Dealing with uncertainty: Readers' memory for and use of conflicting information from science texts as function of presentation format and source expertise. *Cognition and Instruction*, *31*, 130-150.

Van den Broek, P., Rapp, D. N., & Kendeou, P. (2005). Integrating memory-based and constructionist processes in accounts of reading comprehension. *Discourse Processes*: 39, 299-316.

van den Broek, P., Risden, K., Fletcher, C.R., & Thurlow, R. (1996). A 'landscape' view of reading: Fluctuating patterns of activation and the construction of a stable memory representation. In B. K. Britton & A.C. Graesser (Eds.), *Models of understanding text* (pp. 165-187). Hillsdale, NJ: Erlbaum.

Van Oostendorp, H. (Ed.). (2003). *Cognition in a digital world*. Mahwah, NJ: Lawrence Erlbaum Associations.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems. *Educational Psychologist*, 46, 4, 197-221.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, *31*, 3-62.

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., Van Vuuren, S. Weston, T., Zheng, J., & Becker, L. (2011). My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Transactions of Speech and Language Processing, 13*, 4-16.

Weaver, C. A., Manners, S., & Fletcher, C. R. (eds) (1995) *Discourse comprehension: Strategies and processing revisited*. Hillsdale, NJ: Erlbaum.

Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in Internet science inquiry tasks. *American Educational Research Journal, 46*, 1060-1106.

Zwaan, R. A., & Van Oostendorp, H. (1993). Do readers construct spatial representations in naturalistic story comprehension?. *Discourse Processes*, *16*, 125-143.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185.

Zwaan, R. A., Magliano, J.P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 386-397.

Table 1: *Inference classes and their status of being generated during comprehension according to the Constructionist model of Graesser, Singer, and Trabasso (1994)*

	Inference Category	**Description of Inference**	**Constructed during comprehension?**
1	Referential	A word/phrase is referentially linked to a previous element or constituent in the text	Yes
2	Causal antecedent	The inference is on a causal chain between the current explicit action/event/state and the previous passage context	Yes
3	Causal consequence	The inference is on a forecasted causal chain into the future, including events and new plans of characters	No
4	Character emotional reaction	The inference is an emotion experienced by a character directly caused by an event or action	Yes
5	Superordinate goals	The inference is a goal that motivates a character's intentional action	Yes
6	Subordinate goal	The inference is a plan or action that specifies how a character's action is achieved	No
7	Instrument	The inference is an object, part of the body, or resource used when a character performs an action	No
8	Subcategory of noun	The inference is a subcategory or exemplar of an explicit noun	No
9	States of people and objects	Inferences about static properties of characters, objects, and spatial layout	No
10	Thematic	The main point or moral of the text	Yes
11	Author's intent	The inference is the author's motive in writing or attitude	No
12	Emotion of reader	The inference is the emotion that the reader is expected to experience	No

Table 2: *Example Coh-Metrix Measures and Indices*.

LEVEL OR CLASS	MEASURE (INDEX)
Words	Frequency, concreteness, imagery, age of acquisition, part of speech, content words, pronouns, negations, connectives (different categories), logical operators, polysemy, hypernym/hyponym (reflects abstractness); these counts per 1000 words.
Syntax	Syntactic complexity (words per noun-phrase, words before main verb of main clause)
Referential textbasecohesion	Cohesion of adjacent sentences as measured by overlapping nouns, pronouns, meaning stems (lemma, morpheme). Proportion of content words that overlap. Cohesion of all pairs of sentences in a paragraph.
Situation model cohesion	Cohesion of adjacent sentences with respect to causality, intentionality, temporality, spatiality, and latent semantic analysis (LSA).Cohesion among all sentences in paragraph and between paragraphs via LSA. Given versus new content.
Genre and rhetoric	Type of genre (narrative, science, other). Topic sentencehood
Other	Flesch-Kincaid grade level, type token ration, syllables per word, words per sentence, sentences and paragraphs per 1000 words.