Metacognition Learning (2011) 6:131–154
DOI 10.1007/s11409-010-9064-2

Assessing comprehension during reading with the Reading Strategy Assessment Tool (RSAT)

Joseph P. Magliano · Keith K. Millis · The RSAT Development Team · Irwin Levinstein · Chutima Boonthum

Received: 8 March 2010 / Accepted: 15 November 2010 /
Published online: 3 December 2010

Abstract Comprehension emerges as the results of inference and strategic processes that support the construction of a coherent mental model for a text. However, the vast majority of comprehension skills tests adopt a format that does not afford an assessment of these processes as they operate during reading. This study assessed the viability of the Reading Strategy Assessment Tool (RSAT), which is an automated computer-based reading assessment designed to measure readers' comprehension and spontaneous use of reading strategies while reading texts. In the tool, readers comprehend passages one sentence at a time, and are asked either an indirect ("What are your thoughts regarding your understanding of the sentence in the context of the passage?") or direct (e.g., why X?) question after reading each pre-selected target sentence. The answers to the indirect questions are analyzed on the extent that they contain words associated with comprehension processes. The answers to direct questions are coded for the number of content words in common with an ideal answer, which is intended to be an assessment of emerging comprehension. In the study, the RSAT approach was shown to predict measures of comprehension comparable to standardized tests. The RSAT variables were also shown to correlate with human ratings. The results of this study constitute a "proof of concept" and demonstrate that it is possible to develop a comprehension skills assessment tool that assesses both comprehension and comprehension strategies.

Keywords Comprehension assessment · Comprehension processes and strategies · Assessment

J. P. Magliano (✉) · K. K. Millis
Department of Psychology, Northern Illinois University, DeKalb, IL 60115, USA
e-mail: jmagliano@niu.edu

I. Levinstein
Old Dominion University, Norfolk, VA, USA

C. Boonthum
Hampton University, Hampton, VA, USA

Introduction

Educators and researchers interested in assessing reading comprehension have been in a bind. Comprehension occurs as one reads, but most available assessments rely on measuring comprehension *after* reading is completed. This has arisen because the vast majority of standardized comprehension assessment tools use a multiple-choice format where questions are answered after the text is read. Although there are advantages to this testing format (Freedle and Kostin 1994; Glover et al. 1979; Malak and Hegeman 1985; van den Bergh 1990), many of the instantiations of this approach do not adequately assess the products and processes specified by discourse theory to be critical for deep comprehension (Magliano and Millis 2003; Magliano et al. 2007). Part of the problem is that the multiple-choice testing format does not afford the assessment of these processes as they occur during reading because questions are answered only after the test passage is read. Perhaps more importantly, the items that comprise these tests are typically not constructed to assess comprehension products and processes specified in theories of discourse comprehension (Magliano et al. 2007; Snow 2002). For example, Magliano et al. (2007) showed that the Nelson-Denny test, a widely used reading test, primarily contained questions that verified word meanings rather than inferences required for comprehension.

The goal of the present study was to assess the viability of an assessment tool that (1) assesses comprehension online, that is, while students are reading a text, (2) is grounded in theory, and (3)assesses not only comprehension, but some of the processes that give rise to comprehension. We call this tool the *Reading Strategy Assessment Tool* (RSAT; Gilliam et al. 2007). RSAT requires the student to read texts on a computer and answer open-ended questions that are embedded within them. After reading pre-selected target sentences, readers are asked to produce responses to one of two types of open-ended questions: indirect and direct. Indirect questions were intended to tap comprehension processes and require readers to report thoughts regarding their understanding of the sentence in the context of the passage (Instructions include "What are your thoughts regarding your understanding of the sentence in the context of the passage?"). Direct questions were designed to assess comprehension level and required readers to answer specific "wh-"questions about the text at target sentences (e.g., "Why was the Union demoralized?" in a passage about the American Civil War). Based on the answers to the indirect and direct questions, RSAT provides a measure of information processing processes and overall comprehension, respectively. These will be described below.

We should be clear on the onset that RSAT provides no direct assessment of metacognition, although there is evidence linking metacomprehension and calibration skills to comprehension (Hacker et al. 2009; Pressley and Afflerbach 1995). One could develop automated procedures for detecting indications of statements that involve an assessment of one's comprehension (e.g., "I don't get this", "I get this"), but we doubt that this would be sufficient for capturing the nuances of metacognition. Metacomprehension is complicated; readers with a high degree of metacomprehension are able to dynamically adjust their reading strategies to the task, demands of the text, and an assessment of comprehension (e.g., Pressley and Afflerbach 1995; McNamara and Magliano 2009a). Clearly, RSAT would need much more intelligent algorithms than the present ones to correctly identify metacognitive thoughts that relate to comprehension.

We should also be clear on the onset that this approach does not measure many of the strategies that readers use when making meaning of text and educational researchers refer to as 'reading strategies' (e.g., See McNamara (2007) for an extensive review). Active readers will reread, question the text or the author, generate personal examples, look forward and

backwards through the text, take notes, create images, etc. (e.g., Pressley and Afflerbach 1995; Pressley et al. 1985). Moreover, skilled comprehenders change their reading strategies and behaviors in light of reading goals (Pressley and Afflerbach 1995; Taraban et al. 2000). It would be very ambitious to develop a computer-based assessment tool that directly tests the use of all of these strategies during reading, although there are tools that assess the self-reported use of many of them.

In the current study, we were interested in whether a computer-based system could detect a small number of reading processes that are known to contribute to comprehension. As will be described below, we focused on paraphrases, bridges, and elaborations, which together we refer to as information processing activities. Although bridges and elaborations are typically referred to as inferences (e.g., Singer 1988), in some cases, they can be under the strategic control of readers (Magliano et al. 1999). Nonetheless, if our efforts are successful, then this research would suggest that it would be worthwhile to pursue the development of a system that detects other important comprehension processes, as well.

The evidence-based approach for RSAT

The development of RSAT followed the evidence-based approach towards assessment development (Mislevy 1993; Pellegrino and Chudowsky 2003; Pellegrino et al. 2001). This framework specifies that test developers consider three primary components during test development: a model of the *student*, a model of the *task*, and principles for *interpreting data* provided by the task. The student model describes the types of mental processes and representations that the test purports to measure, which ideally are based on theories of student proficiency. The model of the task describes how the tasks or problems that the student will encounter in the assessment tool implicate the processes described in the student model. Principles for data interpretation refer to how performance on the task relates to the student model.

The student model: comprehension

Comprehension arises from a series of cognitive processes and activities that contribute to a reader's ability to connect the meaning of multiple sentences into a coherent mental representation of the overall meaning of text (e.g., Graesser et al. 1994; Kintsch 1988, 1998). The vast majority of theories of discourse comprehension delineate between two classes of information processing activities that support comprehension, namely bridging and elaborative inferences (McNamara and Magliano 2009b).

Bridging inferences provide conceptual links between explicitly mentioned ideas in the text. Bridging inferences include anaphoric and pronominal inferences as well as conceptual and causal-based inferences that require the application of world knowledge. RSAT was designed to detect the evidence for bridging inferences because there is ample evidence indicating that bridging inferences are required for coherence and are routinely generated during comprehension (e.g., Singer and Halldorson 1996) and that skilled and less skilled readers can be differentiated by the extent that they generate these inferences (e.g., Magliano and Millis 2003).

Elaborative inferences are based upon the reader's world knowledge of the concepts and events described by the text. Unlike bridges, elaborations do not provide connections between explicit units of text (e.g., sentences). They are presumably generated because the semantic context of the current sentence resonates with semantic knowledge of the world

and thus, becomes available for computations in working memory (e.g., Myers and O'Brien 1998). They embellish the text content with information from the reader's prior knowledge. This knowledge may be rooted in world (e.g., schematic knowledge), text/topic specific (e.g., knowledge of cancer or biology), or episodic knowledge (e.g., personal experiences). Moreover, some elaboration can be construed and entailments of the text, whereas others may require reasoning beyond the text (Wolfe and Goldman 2005). Unlike bridging inferences there is some controversy regarding the frequency and utility of elaborative inferences. For example, some research has shown that elaborative inferences are generated only when there is a strong semantic association between world knowledge and the text (McKoon and Ratcliff 1986, 1992) or that only a small class of these are routinely generated during reading such as those that support explanations (e.g., Graesser et al. 1994). On the other hand, others have argued that elaborative inferences help readers establish how their knowledge relates to the discourse context and are particularly important when understanding expository discourse (McNamara 2004). With respect to this later perspective, only those elaborations that are germane to the text and/or learning task are likely to support comprehension (e.g., Wolfe and Goldman 2005).

The task: answering questions

Indirect questions and strategies As mentioned earlier, readers answer two types of questions in RSAT. The indirect question—"What are you thinking now?"—was used to elicit responses similar to those produced when thinking aloud. In a think-aloud methodology, readers are asked to report thoughts that come to mind as they read a text. Answers to the indirect protocols should reveal the content of working memory while comprehending the target sentences (Ericsson and Simon 1993), which has been shown to be predictive of comprehension (Chi et al. 1989; Coté and Goldman 1999. Magliano and Millis 2003; Millis et al. 2006; Trabasso and Magliano 1996b). Traditionally, these thoughts are produced orally, transcribed, and then coded in order to make inferences regarding the strategies, information, and mental processes that occur during reading (e.g., Coté and Goldman 1999; Magliano 1999; Magliano et al. 1999; Pressley and Afflerbach 1995; Trabasso and Magliano 1996a, b). In the studies reported here, when this question appeared, participants were instructed to report thoughts regarding their understanding of the sentence in the context of the text (Magliano et al. 1999; Magliano and Millis 2003; Trabasso and Magliano 1996a, b).

It is important to consider some strengths and weaknesses of using verbal protocols for assessing comprehension. A strength of verbal protocols is that they provide a window on the products of comprehension and some of the processes that give rise to comprehension (e.g., Pressley and Afflerbach 1995; Trabasso and Magliano 1996a, b). However, distinguishing between the two can be complicated, and perhaps impossible with an automated scoring system. We are fairly certain that verbal protocols do reveal the emerging products of comprehension, such as inferences, because the thoughts corresponding to the inference would be included in the verbal protocol. But in many cases, the processes that led to the inference being generated might not be apparent from the protocol. For example, a reader might make a prediction about a future event, but whether that inference arose from personal experiences, the prior text or from particular schemata might be unanswerable. This points to a weakness of this approach, namely that all a researcher can analyze is the presence of content in a protocol and infer the processes that give rise to them (Ericsson and Simon 1993). Furthermore, a reader may not be able to write down or say everything that comes to mind, and may edit or omit thoughts that do

come to mind. It is important to validate verbal protocols against other independent behavioral measures, such as reading times or outcome measures of comprehension (Magliano and Graesser 1991). If the contents are correlated with these independent measures, then that provides some validation that their contents are indeed indicative of comprehension. RSAT reflects an application of this general approach for validating verbal protocols.

RSAT was designed to detect three activities that have been revealed in the context of thinking aloud: paraphrasing, bridging, and elaborative inferences (Magliano and Millis 2003; Millis et al. 2006). Paraphrasing reflects an activity in which a reader is rephrasing all or part of the current sentence. Consider the example verbal protocols presented in Table 1 produced at Sentence 11 of "How Cancer Develops" (see the Appendix A for the entire text). Clause 1 for Participant 1 is a paraphrase because the participant mentioned the event that was explicitly described in the current sentence. Paraphrasing provides readers the opportunity to describe the current discourse content with familiar language (McNamara 2004). However, college students who primarily paraphrase as opposed to using other strategies while thinking aloud may do so because they adopt a low standard of comprehension, although this is uncertain at the current time (Magliano and Millis 2003; Millis et al. 2006).

When readers produce bridges, they describe how the current discourse content is related to the prior discourse content. For Example, clause 1 for Participant 2 reflects a local bridge because that participant mentioned the event described in the immediately prior sentence (see the text in Appendix A). Conversely, clauses 1 and 2 for Participant 3 reflect distal bridges because they mentioned information contained in sentences 7, and 9, respectively. Sentences 7 and 9 would not be expected to be in working memory as sentence 11 is read. Elaborations are inferences, which are based on the reader's world knowledge and which do not link text units. Clauses 1 and 2 for Participant 4 are considered elaborations because the reader activated relevant world knowledge not presented in the text. In this case, the reader inferred that the text information could be potentially useful for developing a cure for cancer, a topic not discussed in the text.

There is a growing body of evidence that think aloud protocols reveal qualitative differences among skilled and less-skilled readers (Chi et al. 1989; Coté and Goldman 1999; Magliano and Millis 2003; Millis et al. 2006; Pressley and Afflerbach 1995; Trabasso and Magliano 1996b; Whitney et al. 1991). For example, Magliano and Millis conducted a series of studies in which they had students produce verbal protocols while reading simple

Table 1 Example indirect protocols for the sentence "A message within each receptor cell becomes activated.," From the text "How cancer develops"

Participant	Clause	Protocol	Strategy
1	1.	After it is activated it	Paraphrase
	2.	Becomes lethal to the human body	Elaboration
2	1.	The growth factor attaches to the cell and	Local Bridge
	2.	A message is activated	Paraphrase
3	1.	Cells influence their surroundings	Distal Bridge
	2.	They transmit signals: for example growth signals	Distal Bridge
4	1.	I am wondering how they can use this information	Elaboration
	2.	To find a cure for cancer	Elaboration

These sample protocols were parsed into idea units for ease of illustration, but the protocol were not parsed for the hand coding

narratives (Magliano and Millis 2003) and more challenging scientific texts (Millis et al. 2006). They determined reading comprehension proficiency from scores on the Nelson-Denny test of reading comprehension. Based on the verbal protocols, they found that skilled readers tended to bridge more than less skilled readers. More importantly, these studies showed that the extent to which readers engaged in bridging was correlated with the comprehension of both texts in which the protocols were produced and those that were read silently. This finding suggests that the verbal protocols revealed strategies that are related to comprehension and are used somewhat consistently across reading encounters.

Direct questions and comprehension The direct questions used in RSAT were designed to provide an assessment of comprehension at a particular point in the passage. The majority of the direct questions were why-questions based on the explicit content of the passage. For example, one direct question was "Why does a tumor develop?" and occurred immediately after the sentence "A tumor then develops." which was in a passage that described the development of cancer. Why-questions expose the lion's share of inferences that occur during the comprehension of actions and events (Graesser and Clark 1985; Graesser et al. 1987; Long et al. 1992; Magliano 1999; Millis and Graesser 1994). In particular, why-questions require readers to activate causal and explanatory information that provides a critical basis for deep comprehension (Graesser et al. 1996; Graesser and Clark 1985; Graesser and Franklin 1990; Graesser et al. 1994; Millis and Barker 1996). Therefore, why-questions should be ideal for assessing comprehension.

In RSAT, readers type in answers to both direct and indirect questions onto a keyboard. We are mindful that some researchers have expressed concern that readers adopt different strategies as a function of how the verbal protocols are produced (oral vs. written; Hausmann and Chi 2002). However, Muñoz et al. (2006) found negligible differences in reading inferences and strategies as a function of modality for both simple narratives and more difficult scientific texts. More specifically, they had participants produce verbal protocol both orally and by typing (in the context of a within-participants design) while reading science and narrative texts. They did not find differences in the magnitudes of the processes targeted in RSAT as a function of modality for the science texts. For narrative text, they found that participants produced slightly (albeit significantly) more paraphrases and bridges for the narrative texts when producing the protocols orally than when typing.

Data interpretation: word counts

RSAT relies on a computer-based assessment of the answers readers give to the questions. The presence of information processing activities is estimated by counting words in each answer to the indirect question which match words appearing in different sentences at the point in which the answer is given. Word counts are obtained for content words from the current sentence, local sentence (immediately prior sentence), and distal sentences (two or more sentences back) in the prior discourse context. In addition, content words in each answer that do not appear in the text up to that point are also counted (new words). Words from the current sentences are indicative of paraphrasing, words from the local sentences are indicative of local bridging, words from the distal sentences are indicative of distal bridging, and new words are indicative of elaborations (Magliano et al. 2002; Millis et al. 2007). Answers to the direct questions are assessed by counting the number of content words contained in an "ideal" answer. For each reader, the word counts in each category and question type are averaged.

We are making the assumption that the sources of words used in answers to indirect questions reflect the amount of paraphrasing, local and distal bridging, and elaborations done by a reader (Magliano and Millis 2003; Magliano et al. 2002; Millis et al. 2004, 2006, 2007). As mentioned above, we do acknowledge that there is an inherent ambiguity of what processing activities are being indicated by classifying words on whether they had appeared in the text or not. Given that the prior discourse activates only a small portion of a reader's world knowledge, when a person uses a word from the prior discourse, it is likely that he or she is referring to the prior discourse. Of course, there are no guarantees. A more significant problem arises in the scoring procedure when a person uses a word not found in the prior text. Because the word counting algorithms are unable to detect when a student uses a synonym of a word in the text, synonyms are classified as new words even though they may actually reflect a bridge or paraphrase. Therefore, our measure of elaboration contains words that reflect true elaborations and possibly other processing activities. The extent that this synonym problem poses interpretational limitations is an empirical issue, and we will be mindful of this fact when we interpret the findings.

Overview of the study

The goal of this study is to provide evidence that RSAT measures comprehension and the information processing activities of paraphrasing, elaboration, and bridging. Consequently, the goal of this study is to provide evidence that these measures have construct, convergent, and discriminant validities. Construct validity is established when an instrument truly measures what it intends to measure. Convergent validity helps to establish construct validity by correlating with other measures thought to measure the construct of interest (i.e., comprehension). Discriminant (divergent) validity occurs when instruments that purport to measure different constructs, do not correlate.

To this end, three comprehension assessments were administered in this study. Participants were administered RSAT, the Gates-McGinitie (GM; Level 10/12, Form T) test of reading comprehension, which is a paper multiple-choice standardized test, and also an experimenter-generated test which required participants to answer open-ended questions to expository passages. We also gained access to ACT composite scores as a measure of academic achievement. The ACT (American College Testing) is a high-stakes, standardized test for high school students in the United States that is used by many colleges and universities when making decisions regarding admissions. RSAT's ability to account for overall comprehension (convergent validity) was assessed by correlating RSAT's measure derived from the direct questions with the GM, the experimenter-generated test, and the ACT scores. In addition, convergent validity of RSAT's measure of the information processing strategies was assessed by correlating scores derived from the answers to the indirect questions to blind expert judges who also rated the answers on the same reading strategies. We were also able to assess the discriminant validity of the RSAT processing measures. Specifically, the measure for one process (e.g., paraphrasing) should be more highly correlated with human judgments of that process (e.g., paraphrasing) than human judgments of another process (e.g., elaboration).

It is important to understand the nature of the GM and ACT tests. The comprehension section of GM requires students to read short text segments and answer 3–5 questions for each segment. The texts are available when the participants answer the questions. There is no information provided by the test publishers regarding the underling student model of comprehension that GM is designed to test. However, Magliano et al. (2007) conducted an

analysis of the items in the GM test and specifically classified them as requiring (1) extraction from a local segment (one or two sentences), (2) extraction from a global segment (e.g. paragraph), or (3) inference. These categories roughly correspond to extraction and integration items used in PISA (OECD 2002). Magliano et al. found that 56%, 13%, and 31% of the items could be classified as local extraction, global extraction, or inference items. As one can see, the test primarily contains local extraction items.

The ACT has subtests for reading, English, science, and mathematics. The reading subtest adopts the traditional multiple-choice format, but no information is provided regarding the types of items on this test. Although it would have been optimal to use scores on this subtest, we had to rely on a composite score, which is average score on English, mathematics, reading, science, and an optional writing test subscores. (We did not have access to scores on the reading subtest.) The composite score can be conceptualized as a measure of general academic achievement at the end of a student's junior year in high school, of which certainly reading comprehension ability would contribute.

Methods

Participants One hundred and ninety undergraduates participated for course credit associated with an introductory to psychology course taught at Northern Illinois University. Sixty-one, 58, and 71 participants received RSAT lists A, B, and C, respectively. One hundred and fifty six participants were able to provide their ACT scores. ACT scores were not available for 34 students. The mean composite ACT score was 21.92 (Median score = 22; SD = 3.28). The minimum score was 12 and the maximum score was 30 (36 is the highest possible score).

Procedure There were two phases to the study. In phase 1, participants took the GM test of reading comprehension meant for eleventh and twelfth grade. The test is a standardized test, and took 35 min to administer. Participants were also administered a short-answer (SA) test of comprehension created by the experimenters. The SA test required the student to read two texts and then answer open-ended questions about them. Some of the text questions measured the textbase, which is the propositional representation of the explicit ideas in the text, and others measured the situation model, a representation which includes inferences based on world knowledge (van Dijk and Kintsch 1983). Answers to the textbase questions could be found in one sentence or two adjacent sentences, whereas answers to situation model questions would require the reader's world knowledge or the integration of several sentences across the span of the test. One text was a historical narrative and the other was a science text. These texts were approximately a single page in length (single spaced) and had Flesch-Kincaid Grade Levels of 10.5 and 12 for the science and historical texts, respectively. Ten short-answer questions were constructed for each text.

Two days after phase 1, participants completed phase 2 of the study. In this session, participants took RSAT administered on personal computers in a web-based environment (Gilliam et al. 2007). The texts were presented in black font in a gray field left justified near the top of the computer screen. The title of each text remained centered at the top of the screen while participants read the entire text. In the current study, only one sentence of a text was shown on the screen during reading because this presentation has been shown to be a good predictor of comprehension skill (Gilliam et al. 2007). Participants navigated forward through the text by clicking on a "next" button, which is located near the bottom left portion of the computer screen. "NEW PARAGRAPH" markers appeared when there is

a shift to a new paragraph. After participants clicked the "next" button, the next sentence appeared, provided it was a non-target sentence. The text sentences were not present on the screen when there was a question prompt and nor could the participants navigate back and reread the texts in response to the questions. For target sentences, a response box appeared to the right of the "next" button with a prompt above the box. The prompt for an indirect question was "*What are you thinking now*?" For direct questions, the target sentence was removed from the screen when the question and response box appeared. Participants typed their answers to the question in the response box. They clicked the next button when they were finished, after which the response box disappeared and the next sentence was presented. Responses were recorded on a computer server. The order of the texts was randomly presented to the participants.

Materials Three stimulus lists of passages were used in RSAT. Each stimulus set contained six texts: two science texts, two history texts, and two narratives for a total of 18 texts. The texts and type of question at the target sentences were empirically determined (Magliano et al. 2010). They had participants read and answer either direct or indirect questions in the context of the RSAT tool at pre-selected target sentences (sentences which immediately preceded the questions), which were chosen based on a causal network analysis (Trabasso et al. 1989) of each passage. Target sentences had a relatively high number of causal connections to prior sentences and therefore, an ideal reader would theoretically be able to make bridging inferences at these locations. The texts in the three stimulus sets were chosen because they had a high proportion of target sentences in which the computer-based assessments of the answer were correlated with independent outcome measures (e.g., performance on the GM). Within any given text in the current study, participants answered the indirect or a direct question after reading each pre-selected target sentence. The type of probe (i.e., direct or indirect) was also empirically determined based on the strength of the correlations between the automated assessments and the independent outcome measures. The stimulus sets were created such that the strength of these correlations was as equal as possible. Information regarding the characteristics of the text in the stimulus sets is shown in Table 2. As can be seen in Table 2, the three lists had roughly the same number of direct and indirect target sentence. These scores indicate that the texts were suitable for eighth graders, and should be understandable by university students. However, it is notable that the Lexile score for stimulus set B is higher than the other two sets, but that occurred because one text had an outlier score of 2420L. The Appendices A and B shows a sample text along with direct and indirect questions for the target sentences. The order of presentation of the texts was randomized for each participant. .

RSAT coding of the answers Each answer to the target sentences was automatically scored by identifying the number of content words in the answer that was also in the text or in an ideal answer (Millis et al. 2007). Content words included nouns, adverbs, adjectives and

Table 2 Summary characteristics of the three stimulus lists

Stimulus list	Direct target	Indirect targets	Number of sentences	Number of words	F-K grade level	Lexile score
A	19	17	29.00	364.67	8.00	911.67
B	18	16	22.50	347.33	8.20	1225.00
C	19	15	30.17	404.83	7.40	960.00

verbs (semantically depleted verbs, such as *is*, *are*, were omitted). Word matching was accomplished by literal word matching and Soundex matching (McNamara et al. 2004), which detects misspellings and changes in verb forms (Birtwisle 2002; Christian 1998). For answers to the indirect question, four scores were computed. The *paraphrase* score was the number of content words from the target sentence. The *local bridging* score was the number of content words from the sentence immediately prior to the target sentence. The *distal bridging* score was the number of content words from sentences that were more than two sentences prior to the target sentence. The *elaboration* score was the number of content words in the answer that were not present in the text.

In the case when the same content word appeared in more than one category, it was omitted from the category according to the following order: distal, local, and current sentence. That is, if the same word appeared in both a distal and current sentence, it was excluded from distal and retained in the current sentence. This was done to address an ambiguity that arises when an answer contains a word that occurred in both the current sentence and prior text. Was the reader referring to the current sentence or to the prior text? It is impossible to tell with certainty. Therefore, we assumed that when this occurred, the source of the thought was the current sentence because of its heightened state in working memory, despite the fact that it may have been mentioned earlier. For the direct questions, there was only one score computed: the number of content words in the answer that was in the ideal answer.

For each participant, we computed mean scores by averaging over the individual scores obtained for each target sentence. Therefore, we calculated mean scores for paraphrases, local bridges, distal bridges, elaborations from the answers to the indirect questions, and mean comprehension scores from the answers to the direct questions.

Human coding of the answers The indirect answers were scored by trained human judges using a coding system designed to identify the presence of at least one in the following information processing activities: paraphrases, bridges (local and distal), and elaborations. Think aloud protocols can contain other kids of responses (e.g., Pressley and Afflerbach 1995) and our decision to focus on these stemmed from the fact that the automated system was designed to detect these processes. The unit of analysis was the entire answer to a question. It is important to note that the human judges did not simply count words which appeared or did not appear in the answer or texts. Rather, they were trained to detect the conceptual features of paraphrasing, bridging and elaboration. Moreover, they were trained to identify the use of synonyms, which the word count algorithms cannot detect.

Judges were instructed that paraphrases occurred when the participant restated or summarized information contained in the target sentence. There were three levels for analyzing paraphrases. A “0” indicated that no paraphrase was present. A “1” indicated that the answer contained a noun or noun phrase from the current sentence. A “2” indicated that the answer contained a verb phrase that had its basis in the current sentence. It is important to note that judges included synonyms to both nouns and verbs in their ratings.

Bridges were instances where people mentioned concepts (i.e., content words) and clauses from the prior text. Both local and distal bridges were scored via the same criteria. A “0” indicated that the answer did not contain a local bridge. A “1” indicated that the answer contained a noun or noun phrase from a prior sentence. A “2” indicated that the answer contained a verb clause that had its basis in a clause from a prior sentence. Local bridges occurred when the answer contained information from the immediately prior sentence and distal bridges contained information from all other prior sentences. Again, judges were trained to detect synonymous expressions for local and distal bridges.

Elaborations were instances that contained concepts and inferences not mentioned in the text. A "0" indicated that no elaboration was present. A "1" indicated that the answer contained a noun or noun phrase not present in the text. A "2" indicated that the answer contained a main idea containing a verb clause from world knowledge. Protocols given a two contained a statement that was entailed by the text or were likely the results of reasoning (see Table 1 for examples). Personal recollections (e.g., "There was a thunderstorm last night.", "My grandma died of cancer") and evaluative statements (e.g., "I hate thunderstorms.", "Cancer is scary.") were not considered elaborations. Trained judges worked in pairs and there were two groups of trained judges. Inter-rater reliability for assessing the presence of each category of processing was acceptable (r ranged from .81 to .93).

Given that the unit of analysis was the entire protocol, each protocols could contain evidence for multiple processes, which is illustrated in the first two examples in Table 1. In fact, it is often the case that protocols contain any combination of the processes targeted by RSAT (e.g., Trabasso and Magliano 1996a, b; McNamara 2004). As such, these processes are not to be viewed as mutual exclusive. Moreover, the protocols could contain categories of responses that that were not part of the coding system (e.g., evaluative statements, recollections from episodic knowledge).

Another scoring system was developed to assess the quality of responses to the direct questions and the answers to the SA questions (i.e., the experimenter-generated test). The system identified the ideal parts of ideal answers for each question. Responses were scored on a four point scale (0–3). A three indicated that the answer was completed; a 2 indicated that it was almost complete; a 1 indicated that the answer was vague, but largely correct; finally a 0 indicated that the answer was incorrect. Rules for assigning these numbers were established by the coders for each question-answer pair. Inter-rater reliability was high (r=.89). The SA questions were scored in a similar fashion. The inter-item reliability of the SA test was adequate (r=.92).

To summarize, both RSAT and human coding used the same unit of analysis, which was the entire answer to an indirect or direct question. RSAT counts the number of words that fall into different categories that corresponded to the constructs that were intended to be assessed by RSAT. For indirect questions, RSAT counted the number of content words which were also present in the target sentence, the sentence immediately prior to the target sentence, other prior sentences, and the number of words which did not appear in the text. These corresponded to paraphrasing, local bridging, distal bridging and elaborations. All of these had a lower bound of 0 and no upper bound. The human scoring was based on the conceptual presence of these processing activities, and had a lower bound of zero and an upper bound of 2. For direct questions, RSAT counted the number of words in the participant's answer that was included in an 'ideal' answer, whereas the human judges used a 3-point scale of answer completion (i.e., incorrect, partially complete, complete).

Results

There were four sets of analyses conducted to assess the validity of RSAT. The first set correlated our on-line measure of comprehension (word counts based on the answers to direct questions) the GM test of comprehension, the experimenter-generated comprehension test (i.e., the SA test), and the ACT. If RSAT shows convergent validity for measuring comprehension, then its measure of comprehension should correlate with these other measures to about the same extent they correlate with one another. The second set assessed

RSAT's construct validity for its measure of comprehension, as well as its measures of information processing activities (paraphrasing, bridging, elaborations) which were word counts based on the answers to the indirect questions. We correlated RSAT-generated word count measures of comprehension and information processing activities with expert human judgments of these same variables. The third set assessed the construct validity of the RSAT measures of information processing activities by testing whether they predict performance on the comprehension measures. According to theory (e.g., Graesser et al., 1994), the measures should be correlated with comprehension. Finally, we tested whether the set of RSAT's measures (comprehension, paraphrasing, local bridging, distal bridging, elaborations) predict performance on the SA test of comprehension over and above that accounted for by the GM comprehension test.

The mean and standard errors for all measures of comprehension and information processing activities are presented in Table 3. The means derived from RSAT were based on the answers to the embedded indirect and direct questions, and were calculated for each participant separately.

Analysis 1: Comparing RSAT's measure of comprehension with other measures of comprehension We computed the bivariate correlations between the word counts for embedded direct questions (RSAT's measure of comprehension), GM performance, and performance on the SA questions (see Table 3). The correlations among the measures were all comparable and statistically significant ($p<.001$). The mean word counts for direct questions were significantly positively correlated with the ACT ($r=.54$), GM ($r=.53$) and the experimenter-generated SA test ($r=.45$) scores. The correlations between RSAT and the two multiple choice tests were similar in magnitude that they correlated with each other ($r=.59$). The experimenter-generated SA questions correlated with ACT ($r=.56$), GM ($r=.52$). The similar magnitudes of these correlations provide some degree of construct validity of RSAT's assessment of comprehension.

Analysis 2: Comparing RSAT's word counts to human raters The next research questions pertained to establishing that the RSAT measures (as revealed by the embedded direct and

Table 3 Means and standard errors for the measures of comprehension and comprehension processes

Measures	Mean	SE
ACT	21.92	3.28
Gate-McGinitie (GM)	34.81	0.46
Short-Answer (SA)	0.34	0.01
Direct question—word counts	1.53	0.04
Direct question—human judgments	0.38	0.02
Paraphrase - word counts	0.95	0.03
Local bridging—word counts	0.35	0.02
Distal bridging—word counts	1.15	0.05
Elaboration—word counts	3.07	0.11
Paraphrase—human judgments	0.91	0.02
Local bridging—human judgments	0.31	0.01
Distal bridging—human judgments	0.70	0.02
Elaboration—human judgments	1.30	0.02

indirect questions) of comprehension and information processing activities actually measure what they were intended to measure. The correlations between RSAT measures and expert human raters were very encouraging, and again were all statistically significant (r<.001). The correlations for paraphrases, local bridges, distal bridges and elaborations were .70, .70, .64, and .44, respectively. As can been seen, word counts correlated better for information in the discourse context (paraphrases, local bridges, and distal bridges) than for elaborations. Overall, these correlations help to establish the convergent validity of the information processing activities. That is, RSAT and humans show relatively high correlations for each processing activity.

It is also helpful to address discriminant validity. Discriminant validity refers to when a measure is uncorrelated with measures of theoretically unrelated concepts. That is, RSAT's measure of paraphrasing should be correlated with human ratings of paraphrasing (convergent validity) but relatively uncorrelated with human ratings of local bridges, distal bridges, and elaborations (discriminant validity) since these later relate to other processing activities. A simple way to address this is to compare the correlations in the preceding paragraph to the off-diagonal correlations when the correlations are plotted in a 4 (RSAT) × 4 (Human) matrix. For example, the RSAT measure of paraphrase correlated with the human rating of paraphrases at .70, and this correlation should be, and in fact was higher, than its correlation with the human ratings of local bridging (.43), distal bridging (.27) and elaboration (.19). Hence, the on-diagonal and average off-diagonal correlation for paraphrases was .70 and .29, respectively. The corresponding on- and off-diagonal correlations for local bridges, distal bridges and elaborations were .70 vs. .41, .64 vs. .41, and .44 vs. .32, respectively. Overall, the pattern of correlations indicate the following order of convergent and discriminant validity: paraphrasing > local bridges > distal bridges > elaborations.

In reference to RSAT's measure of overall comprehension, the word counts from the direct answer were significantly correlated with the human judgments with a correlation of .70 (p<.001). Overall, the pattern of correlations suggests that most of the word counts are correlated with human experts, and therefore, can serve as a proxy for human judgments.

Analysis 3: Predicting comprehension from RSAT's measures of information processing activities According to theory, measures of paraphrasing, bridging and elaboration should predict comprehension (Graesser et al. 1994). First, we predicted performance on the direct questions from mean scores of paraphrasing, local bridging, distal bridging, and elaboration using multiple regression (Dummy-coded variables were entered for each of the three RSAT stimulus lists.). These variables were simultaneously forced entered. Table 4 contains the resulting coefficients. The regression equation on the embedded direct questions accounted for a significant 38% of the variance, F (6, 144)=14.40, p<.001. Paraphrase, distal bridging, and elaborations scores were all significant positive predictors of performance on the embedded direct questions.

Second, we predicted the performance on the SA questions from the same measures. The equation accounted for a significant 21% of the variance, F (6, 144)=6.35, p<0.001. As can be seen in Table 5, distal bridging, and elaborations scores were both significant positive predictors. Paraphrasing scores was a significant negative predictor of performance on the SA test, consistent with prior findings (Magliano and Millis 2003; Millis et al. 2006).

Overall, these results illustrate that RSAT's measures of information processing activities predict measures of comprehension provided by the RSAT tool (direct questions) and by independent measures (SA questions), thus establishing their construct validity.

Table 4 Bivariate Pearson correlations between measures of comprehension and comprehension processes

Measure	2	3	4	5	6	7	8	9	10	11	12	13
1. ACT	0.59	0.56	0.54	0.54	0.11	0.19	0.37	0.35	0.16	0.30	0.24	0.12
2. GM		0.52	0.53	0.50	0.16	0.17	0.28	0.30	0.11	0.26	0.18	0.10
3. Short answer			0.45	0.49	−0.05	0.16	0.38	0.34	−0.05	0.24	0.17	0.18
4. Direct Question—word counts				0.70	0.32	0.17	0.45	0.50	0.28	0.30	0.32	0.32
5. Direct Question—human judgments					0.12	0.15	0.24	0.38	0.18	0.17	0.32	0.32
6. Paraphrase—word counts						0.37	0.35	0.17	0.70	0.33	0.32	0.12
7. Local bridging—word counts							0.44	0.22	0.43	0.70	0.51	0.25
8. Distal bridging—word counts								0.50	0.27	0.50	0.64	0.29
9. Elaboration—word counts									0.19	0.31	0.48	0.46
10. Paraphrase—human judgments										0.43	0.37	0.20
11. Local bridging—human judgments											0.53	0.26
12. Distal bridging—human judgments												0.44
13. Elaboration—human judgments												

Table 5 Predicting performance on comprehension measures

Predictor variables		Direct questions	SA questions
Paraphrase score	*Beta*	0.20	−0.17
	SE	0.09	0.03
	t	2.83	2.16
	p	0.01	0.03
Local bridging score	*Beta*	0.05	0.12
	SE	0.19	0.07
	t	0.65	1.31
	p	0.59	0.19
Distal bridging score	*Beta*	0.20	0.26
	SE	0.09	0.03
	t	2.04	2.41
	p	0.04	0.02
Elaboration score	*Beta*	0.30	0.21
	SE	0.03	0.10
	t	3.85	2.29
	p	0.00	0.02

Analysis 4: Comparing RSAT to standardized tests The final question pertained to whether the RSAT measures can account for comprehension performance on a level comparable to standardized tests. Our "gold" standard measure of comprehension performance was the answers to the SA test. The standardized tests were the GM and ACT scores. A series of regression analyses were conducted for these analyses. The first involved simultaneously entering RSAT scores for comprehension (direct questions), paraphrasing, local bridging, distal bridging, and elaboration into a regression equation predicting performance on the SA questions. This analysis revealed that the measures provided by RSAT accounted for a significant 33% of variance, F (7, 143)=10.21, $p<.001$. A comparison of the multiple correlation coefficient provided by the regression equation to the bivariate correlations between the standardized measures and SA performance provide one basis for comparing the measures. The correlation coefficients for RSAT, GM, and ACT were .58, .52, and .56, respectively. As such, each of the assessment approaches were comparably correlated with performance on the SA questions.

A set of two-step, hierarchical regression analyses was also conducted. This analysis compared RSAT to the GM and ACT on the amount of unique variance they share with the SA performance. In the first step of each analysis, the RSAT measures were simultaneously force entered into the equation, and in the second step, the standardized measure was entered (GM or ACT). Next, another hierarchical analysis was conducted, but the order of entry of the assessment measures was reversed (e.g., GM entered in the first step and RSAT measures entered in the second step). R^2 changes for the second steps of these analyses were extracted to assess the unique variance of each measure. With respect to the comparison of the RSAT measures with GM, RSAT scores accounted for a significant 14% of the variance. (F (5,142)=6.94, $p<.001$) and the GM scores accounted for a significant 8% of the variance, (F (1,142)=20.54, $p<.001$). With respect to the comparison of the RSAT measures with ACT, RSAT scores accounted for a significant 10% of the variance. (F (5,142)=4.97, $p<.001$) and the ACT scores accounted for a significant 9% of the variance, (F (1,142)=23.19, $p<.001$). The results of this last set of analyses suggest that the RSAT measures of comprehension and comprehension processes were as predictive of comprehension (SA questions) as the two standardized tests (GM, ACT).

Discussion

Traditional comprehension assessment tools are typically not designed to assess comprehension as it emerges during reading or the processes that give rise to comprehension (Magliano et al. 2007). Developing assessment tools that provide valid and reliable assessments of these dimensions of comprehension could provide a boon to educational practices because they could provide a basis for giving feedback to students regarding *how* they approach reading for comprehension. The present study assessed the viability of assessing comprehension by using a computer-based scoring system of open-ended responses to questions embedded in text, which we believe could be the basis of such an assessment tool. In this study, we used RSAT, which, unlike the vast majority of commercially published comprehension skills assessments, was developed based on the evidence-based approach for test development (Mislevy 1993; Pellegrino and Chudowsky 2003; Pellegrino et al. 2001). Theory guided the construction of the questions and where they were placed within the passages. The research we presented here can be viewed as

providing a "proof of concept" for RSAT and more generally an assessment tool designed not only to assess comprehension skill, but some of the processes that support it.

What evidence do we have for the proof of concept? First, we have evidence of convergent validity of the RSAT measure of comprehension and standardized and experimenter-generated measures of comprehension. That is, the RSAT comprehension score was correlated with performance on the GM and ACT tests as performance to the same extent that these were correlated with each other. Moreover, the RSAT, GM and ACT scores were all comparably correlated with performance on the SA test. When we assessed the unique variance accounted by the RSAT measures and the standardized measures (GM and ACT), we found that the RSAT measures accounted for slightly more unique variance than the GM test and a comparable amount as the ACT. One advantage of RSAT over these standardized tests is that it provides an assessment of some of the processes that support comprehension.

Second, the data indicate that the RSAT comprehension and information processing activities were correlated with human judgments. The pattern of correlations showed convergent validity and some degree of discriminant validity. The correlations indicative of convergent validity were high and statistically significant. These data are impressive given that human judges were trained to detect the strategies, rather than particular words or word counts. These data indicate that the word counts were indeed valid measures of paraphrasing, local and distal bridges, and answers to the direct questions. The measures of elaboration had the lowest indicators of validity.

Third, we showed that the RSAT measures of information processing activities were predictive of performance on both the RSAT measure of comprehension (direct questions) and performance on the independent SA test. The measures of integration (distal bridging scores) and elaboration (elaboration score) were all significant positive predictors of both measures of comprehension. However there was discrepancy with respect to the paraphrase score. The paraphrase score was a significant positive predictor for the direct questions, but a significant negative predictor for the SA test. In our prior research, we have documented similar negative correlations with comprehension tests similar to the SA test used here (Magliano and Millis 2003; Millis et al. 2006). We have interpreted this finding as indicating that readers who paraphrase excessively tend not to integrate the current sentence to prior sentences and therefore do not construct globally coherent representations (see also Wolfe and Goldman 2005). That is, they focus on understanding individual sentences rather than linking each sentence to the existing passage representation. One reason for finding a positive correlation between paraphrasing and performance on the direct questions is that paraphrasing probably strengthens memory for the explicit content of the text, and in many cases, the explicit text provided the correct answer to the direct answers.

Although we believe we have met the 'proof of concept' requirement, it is premature to conclude we are ready to develop a test of comprehension skill based on RSAT that would be ready to be used on a large scale. Comprehension emerges from a complex interaction between the reader, text, and task (Snow 2002). We believe that the development of an assessment tool would require a deeper understanding of this interaction. For example, some strategies would be more appropriate for some text than others. If readers are comprehending a texts that describes a causal sequence of events, then causal bridging inferences are important for comprehension, but this would not be the cause for text that describe biological nomenclature. We do believe that the data show that tests based on open-ended responses are plausible to construct and as RSAT could provide a valuable research tools to develop assessment tools of this nature.

It is important to acknowledge that RSAT only assesses paraphrasing, bridging, and elaboration, which only comprise a small subset of the processes that support comprehension. Skilled comprehenders use a number of strategies for meeting their reading goals (Gaskins et al. 2007; Pressley and Afflerbach 1995; Pressley et al. 1985; Pressely and Woloshyn 1995). RSAT cannot determine the processes (strategic or otherwise) that give rise to the answers that readers produce. For example, RSAT cannot determine if bridging words were produced as part of a strategy to summarize or self explain, or whether it was under the strategic control of the reader. The cognitive processes that give rise to the content of verbal protocols can be induced by human judges in some instances (Trabasso and Magliano 1996a, b), but it would be challenging to implement them in an automated system with our current approach of essentially taking "snap shots" of comprehension and processes across a text. Despite the limitations of the current version of RSAT, its framework allows for flexibility in that researchers and educators can compose their own direct questions to suit particular goals. For example, one could construct questions that align with well established assessment frameworks, such as what is adopted by OECD (2002). Specifically, one could develop questions that require the extraction, integration, or evaluation of the materials.

Additionally, one would need to be able to assess the relationship between metacognitive awareness and the processes revealed by RSAT. Certainly, the effective use of these strategies requires metacognitive awareness of their appropriateness given the dynamically changing demands of a text and a reader's level of comprehension (McNamara and Magliano 2009a). We have conducted subsequent research that had shown that measures associated with self-regulation (including self-reported awareness and use of metacognitive strategies) partially mediate the relationship between the reading processes measured by RSAT and comprehension (Magliano et al. 2010). It is important to note that this study relied on human judgments of the answers and one would want to assess if RSAT word counts are sensitive to this relationship as well.

Another factor that requires improvement in RSAT is the detection of elaborations. As we discussed above, the synonym problem creates a situation such that our measure of new words will contain the use of synonyms for words in the discourse. As such, our measure of elaborations can be conceptualized as a continuum of relevant knowledge ranging from synonyms of text words to true knowledge-based elaborative inferences. The low to moderate correlations between new words and human judgments of elaboration can be explained by the fact that our human judges were trained to detect true elaborations rather than synonyms of text content. The synonym problem may also explain why 'new words' was not significantly correlated with outcome measure of comprehension, although they were in the analyses involving the entire data set. In our past attempts to detect elaborations via computer-based coding, we created a list of words from elaborations that readers tend to use when thinking aloud while reading the target sentences and then used LSA (Landauer and Dumais 1997) to compare those words to the protocols (Millis et al. 2004, 2007). LSA stands for latent semantic analysis and is a statistical method for assessing the semantic similarity between two units (words, sentences, paragraphs, texts) of text. However, we have obtained higher correlations with human judgments with the current approach. Yet the 'elaboration problem' remains—it is difficult to anticipate the variety of elaborations that readers can produce, which is different from text-based inferences, which are constrained by the content of the text.

Moreover, not all elaborations are relevant and support deep comprehension (Wolfe and Goldman 2005). For example, connecting the text with personal episodic knowledge may not be useful in supporting deep comprehension that reflects the underlying meaning of the

text. The human coding system was designed to focus on elaborations that were based on general or text specific prior knowledge or the result of "on the fly" reasoning. However, the RSAT word count algorithms currently have no basis for distinguishing these processes from other kinds of responses that would involve words outside the textual context (e.g., valianced evaluations or personal recollections). As such, unlike the word counts for paraphrasing and bridging, the word counts for elaborations are likely reflective of processes that go beyond those that were targeted in the coding scheme. This is clearly an aspect of the current word count algorithms that warrants improvement.

Additionally, we should point out some concerns regarding the ecological validity of RSAT, although some have argued that overzealous concerns regarding ecological validity can stifle innovations in education (Dunlosky et al. 2009). The most pressing, in our opinion, is how embedded questions affect the profile of responses and the comprehension of RSAT texts. Readers are not typically asked questions as they read. There has been an impressive literature over the last three decades on the effect that adjunct questions have on increasing comprehension (e.g., Callender and McDaniel 2007; Peverly and Wood 1999; Rothkopf 1970; Sagerman and Mayer 1987). It is likely that the indirect question had less of an impact than the direct questions on comprehension because think aloud instructions in which the indirect question was modeled on do not interfere with complex cognitive activities (Ericsson and Simon 1993). Future research is needed to indicate the extent that answering indirect and direct questions affect the comprehension of the texts, if they do so at all. Another ecological issue arises from the fact that readers were only able to see one sentence at a time and were unable to go back to prior sentences. Normally, readers can go back to read prior text when they wish. An alternative version of RSAT would allow readers to regress, but not during question-answering because the reader might use the prior text to answer the question, and that would be antithetical to the goal of assessing comprehension as it happens.

In conclusion, RSAT provides a new approach to reading assessment. We believe that its greatest promise is as a formative assessment that could be used to inform and guide interventions designed to help struggling readers. For example, it could be used in the context of tailoring strategy training to the specific needs of students in the context of computer based strategy training (e.g., McNamara et al. 2004). For example, if RSAT demonstrates that a student does not bridge, then that strategy could be emphasized during training. Additionally, RSAT could be valuable in developmental reading programs in post secondary education. Most of these rely on the Nelson-Denny Test of reading comprehension to diagnose students with comprehension problems (Boylan 1983; Wang 2006). RSAT could be valuable for this population of readers because we have demonstrated that RSAT outperformed the GM on accounting for comprehension, a multiple-choice test similar to the Nelson-Denny. In addition, RSAT is able to identify subcomponents of comprehension, namely paraphrasing, bridging and (to some extent) elaborating, in addition to an overall measure of comprehension. The Nelson-Denny and GM give an overall account of comprehension, but no measure of strategies or inferences.

Acknowledgments The research was supported by Institute for Education Sciences Grant (IES) R305G040055. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the IES. The RSAT development team consists of, in alphabetical order, Srikanth Dandotkar, Sara Gilliam, Chris Kurby, P. J. Perry, Stacey Todaro, and James Woehrle. The authors would like to thank the following individuals for all of their hard work. We would like to thank Mary Hecht, Laura Goetten, Siva Kancherla, Megan Miller, Brent Munoz, and Pratap Verbenelli for their valuable contributions on this project.

Appendix A

Example RSAT Text

Bolded sentences are locations for indirect questions and italicized sentences are locations for direct questions. The direct questions follows these sentences

How Cancer Develops

NEW PARAGRAPH.

1 Cancer begins in genes, bits of biochemical instructions composed of individual segments of the long, coiled molecule deoxyribonucleic acid (DNA).
2 Genes contain the instructions to make proteins, molecular laborers that serve as building blocks of cells, control chemical reactions, or transport materials to and from cells.
3 In a cancerous cell, permanent gene alterations, or mutations, cause the cell to malfunction.
4 For a cell to become cancerous, usually three to seven different mutations must occur in a single cell.

5 Cancer may take many years to accumulate.

NEW PARAGRAPH.

6 Understanding how cells communicate with one another is an important part of the story.
7 While each human cell performs its own specialized function, it also exerts influence on the cells around it.
8 Cells communicate with one another via receptors, protein molecules on the cell surface.
9 A cell may instruct other cells in its neighborhood to divide, for example, by releasing a growth-promoting signal, or growth factor.
10 The growth factor binds to receptors on adjacent cells.

11 A message within each receptor cell becomes activated.

NEW PARAGRAPH.

12 When the growth factor message reaches the cell nucleus, it activates genes called proto-oncogenes.
13 These genes produce proteins that stimulate the cell to divide.
14 In cancerous cells, mutations in proto-oncogenes cause these genes to malfunction.
15 When a proto-oncogene mutates, it becomes an oncogene—a gene that instructs the cell to grow and divide repeatedly without stimulation from neighboring cells.
16 Some oncogenes overproduce growth factors.
17 The cell may then divide too often.

Why would this happen?

NEW PARAGRAPH.

18 When severely mutated, a cell will continue to divide, even with damaged DNA.
19 The faulty DNA cannot duplicate properly during cell division.
20 The cells can become even further mutated.

Why would the cells become further mutated?

NEW PARAGRAPH.

21 Mutated cells that continue to reproduce can cause further mutations.
22 A tumor can then develop.

How can a tumor develop?

23 A tumor is a mass of cells not dependent upon an extracellular matrix.
24 These cells can grow on top of each other, creating a mass of abnormal cells.
25 Often, a tumor develops its own network of tiny blood vessels to supply itself with nutrient-rich blood, a process called angiogenesis.

26 Tumors can spread to other parts of the body.

27 Tumors can severely hinder the functioning of vital organs and biological processes.

Appendix B

Example text and questions for the short-answer test.

Franco Dictatorship

The Franco dictatorship lasting from 1936 to 1975 was one of the most oppressive periods in modern Spanish History. Franco took power in Spain after the Spanish Civil War in 1936. Supporters of the prior government, known as Republicans, included most workers, liberals, socialists, communists, and Basque and Catalan separatists. The Franco government labeled all political opposition as communists and used that to justify their harsh actions. In the first 4 years after the war, the government imprisoned hundreds of thousands of people and executed many thousands of others. The Franco government tracked people suspected of Republican sympathies and persecuted them for decades.

The dictatorship's main source of political support included the army, the Catholic Church, and the Falange, the Spanish National Movement. The common enemies were the socialist and communist movements in Spain. The army provided the dictatorship with security, while the Catholic Church and the National Movement gave Franco's rule a measure of legitimacy. As long as Franco openly opposed communism, the Church turned a blind eye to the dictatorship. To this day, many Spanish citizens who lived under the dictatorship have a distrust of the Catholic Church.

Franco, who sympathized with fascist ideas, was a great admirer of Adolf Hitler. Spanish industries were inefficient and the transportation system was largely in ruins, making mobilization

for war difficult. Thus, Spain was unable to offer assistance to Germany. Spain was forced to adopt an official policy of neutrality during the war. Despite this, Spain sold valuable raw materials, such as steel, to some of the Axis powers. Spain emerged from the war politically and economically isolated. Many countries cut off diplomatic relations with Spain also.

Domestically, Franco's economic policies further isolated Spain and led to a disastrous period of economic stagnation. Franco believed that Spain could achieve economic recovery and growth through rigorous state regulation of the economy. Franco's government made few investments to rebuild the nation's shattered infrastructure, as well as his policies effectively deprived Spain of foreign investment. Agricultural output and industrial production languished, wages plummeted, and the black market flourished. High inflation and low wages defined the Spanish economic landscape. To make matters worse, Franco refused to seriously open the Spanish economy to foreign trade and investment.

Franco was forced to institute changes that ultimately weakened his government's grip on the country. The cabinet was reorganized in order to increase labor and business representation in the government. Industrial production boomed. Impoverished agricultural workers left the fields for better paying jobs in the city. Labor agitation increased, workers were dissatisfied and organized into unofficial trade unions to press for better pay, benefits, and working conditions. By the late 1960's and early 1970's, Spain was a society at odds with the aging Franco dictatorship. The dictatorship finally lost power in 1975.

Short Answer (SA) Questions

1.) When did Franco take power in Spain? TEXTBASE
2.) Identify at least two enemies and supporters of Franco's government. TEXTBASE
3.) Why would some people living in Spain today distrust the Catholic Church? SITUATION MODEL
4.) Was Spain neutral during World War II? Why or why not? SITUATION MODEL
5.) What did Spain sell to its allies during World War II? TEXTBASE
6.) What did most countries do to Spain after World War II? TEXTBASE
7.) What were the causes of the great period of economic stagnation that followed World War II? SITUATION MODEL
8.) Why did Franco re-organize his Cabinet and what were the results of that reorganization? SITUATION MODEL
9.) Near the end of Franco's rule, why did agricultural workers leave their fields and what were the consequences? SITUATION MODEL
10.) When did the Franco Dictatorship loose power? TEXTBASE

References

Birtwisle, M. (2002). The soundex algorithm. Retrieved from: http://www.comp.leeds.ac.uk/matthewb/ar32/basic_soundex.htm.

Boylan, H. R. (1983) A review of diagnostic reading tests. *Research in Developmental Education newsletter, 1*(5), np.

Callender, A. A., & McDaniel, M. A. (2007). The benefits of embedded question adjuncts for low and high structure builders. *Journal of Educational Psychology, 99*, 339–348.

Chi, M., Bassok, M., Lewis, M. W., Reimann, R., & Glaser, R. (1989). Self-explanation: how students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182.

Christian, P. (1998). Soundex—can it be improved? *Computers in Genealogy, 6*, 215–221.

Coté, N., & Goldman, S. R. (1999). Building representations of informational text: Evidence from children's think-aloud protocols. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 169–193). Mahwah: Erlbaum.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: verbal reports as data*. Cambridge: MIT.

Dunlosky, J., Bottiroli, S., & Hartwig, M. (2009). Sins committed in the name of ecological validity: A call for representative design in education research. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Handbook of metacognition in education* (pp. 430–440). NY: Psychology Press.

Freedle, R., & Kostin, I. (1994). Can multiple-choice reading tests be construct valid? *Psychological Sciences, 5*, 107–110.

Gaskins, I. W., Satlow, E., & Pressley, M. (2007). Executive control of reading comprehension in the elementary school. In L. Meltzer (Ed.), *Executive function in education: From theory to practice* (pp. 194–215). New York: Guilford.

Gilliam, S., Magliano, J. P., Millis, K. K., Levinstein, I., & Boonthum, C. (2007). Assessing the format of the presentation of text in developing a Reading Strategy Assessment Tool (RSAT). *Behavior Research Methods, Instruments, & Computers, 39*, 199–204.

Glover, J. A., Zimmer, J. W., & Bruning, R. H. (1979). Utility of the Nelson-Denny as a predictor of structure and themanticity in memory for prose. *Psychological Reports, 45*, 44–46.

Graesser, A. C., & Clark, L. C. (1985). *Structures and procedures of implicit knowledge*. Norwood: Ablex.

Graesser, A. C., & Franklin, S. P. (1990). QUEST: a cognitive model of question answering. *Discourse Processes, 13*, 279–303.

Graesser, A. C., Haberlandt, K., & Koizumi, D. (1987). How is reading time influenced by knowledge-based inferences and world knowledge? In B. Britton (Ed.), *Executive control processes in reading*. Hillsdale: Erlbaum.

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

Graesser, A. C., Baggett, W., & Williams, K. (1996). Question-driven explanatory reasoning. *Applied Cognitive Psychology, 10*, S17–S32.

Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.). (2009). *Handbook of metacognition in education*. Mahwah, NJ: Lawrence Erlbaum and Associates.

Hausmann, R., & Chi, M. T. (2002). Can a computer interface support self-explaining? *International Journal of Cognitive Technology, 7*, 4–14.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review, 95*, 163–182.

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

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Long, D. L., Golding, J. M., & Graesser, A. C. (1992). Test on the on-line status of goal-related inferences. *Journal of Memory and Language, 31*, 634–647.

Magliano, J. P. (1999). Revealing inference processes during text comprehension. In S. R. Goldman, A. C. Graesser, & P. van den Broek (Eds.), *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (pp. 55–75). Mahwah: ErIbaum.

Magliano, J. P., & Graesser, A. C. (1991). A three-pronged method for studying inference generation in literary text. *Poetics, 20*, 193–232.

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition and Instruction, 21*, 251–284.

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

Magliano, J. P., Wiemer-Hastings, K., Millis, K. K., Muñoz, B. D., & McNamara, D. S. (2002). Using latent semantic analysis to assess reader strategies. *Behavior Research Methods, Instruments, & Computers, 34*, 181–188.

Magliano, J. P., Millis, K. K., Ozuru, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 107–136). Mahwah: Erlbaum.

Magliano, J. P., Perry, P. J., Millis, K. K., & Parker, C. (2010). *The co-influence of reader resources and inference processes on comprehension*. Presented at the 21st Annual Meeting of the Society for Text and Discourse, Chicago, IL.

Malak, J., & Hegeman, J. N. (1985). Using verbal SAT scores to predict Nelson-Denny scores for reading placement. *Journal of Reading, 28*, 301–304.

McKoon, G., & Ratcliff, R. (1986). The automatic activation of episodic information in a semantic memory task. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 12*, 108–115.

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*, 440–466.

McNamara, D. S. (2004). SERT: self-explanation reading training. *Discourses Processes, 38*, 1–30.
McNamara, D. S. (Ed.). (2007). *Reading comprehension strategies: Theories, interventions, and technologies*. Mahwah: Erlbaum.
McNamara, D. S., & Magliano, J. P. (2009a). Self-explanation and metacognition: The dynamics of reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 60–81). Mahwah: Erlbaum.
McNamara, D. S., & Magliano, J. P. (2009b). Towards a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation, vol. 51* (pp. 297–384). New York: Elsevier Science.
McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments & Computers, 36*, 222–233.
Millis, K. K., & Barker, G. (1996). Question answering for expository texts. *Discourse Processes, 21*, 57–84.
Millis, K. 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. K., Kim, H. J., Todaro, S., Magliano, J. P., Wiemer-Hastings, K., & McNamara, D. S. (2004). Identifying reading strategies using latent semantic analysis: Comparing semantic benchmarks. *Behavior Research Methods, Instruments, & Computers, 36*, 213–231.
Millis, K. K., Magliano, J. P., & Todaro, S. (2006). Measuring discourse-level processes with verbal protocols and latent semantic analysis. *Scientific Studies of Reading, 10*, 251–283.
Millis, K. K., Magliano, J. P., Todaro, S., & McNamara, D. S. (2007). Assessing and improving comprehension with latent semantic analysis. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *LSA: A road to meaning*. Mahwah: Erlbaum.
Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederikson, R. J. Mislevy, & I. I. Bejar (Eds.), *Tests theory for a new generation of tests*. Hillsdale: Erlbaum.
Muñoz, B., Magliano, J. P., Sheridan, R., & McNamara, D. S. (2006). Typing versus thinking aloud when reading: implications for computer-based assessment and training tools. *Behavior Research Methods, Instruments, & Computers, 38*, 211–217.
Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes, 26*, 131–157.
Organisation for Economic Co-Operation and Development (OECD). (2002). *Reading for change. Performance and engagement across countries*. Paris: OECD Publications.
Pellegrino, J. W., & Chudowsky, N. (2003). The foundations of assessment. *Interdisciplinary Research and Perspectives, 1*, 103–148.
Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science of design of educational assessment*. Washington: National Academy of Sciences.
Peverly, S. T., & Wood, R. (1999). The effects of adjunct questions and feedback on improving the reading comprehension skills of learning-disabled adolescents. *Contemporary Educational Psychology, 26*, 25–43.
Pressely, M., & Woloshyn, V. (1995). *Cognitive strategy instructions that really improves children's academic performance*. Cambridge: Brookline Books.
Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale: Erlbaum.
Pressley, M., Forrest-Pressley, D. L., Elliot-Faust, D., & Miller, G. (1985). Children's use of cognitive strategies, hot to teach strategies, and what to do if they can't be taught. In M. Pressley & C. J. Brainer (Eds.), *Cognitive learning and memory in children: Progress in cognitive development research* (pp. 1–37). New York: Springer-Verlag.
Rothkopf, E. Z. (1970). The concept of mathemagenic activities. *Review of Educational Research, 40*, 325–336.
Sagerman, N., & Mayer, R. E. (1987). Forward transfer of different reading strategies evoked by adjunct questions in science text. *Journal of Educational Psychology, 79*, 189–191.
Singer, M. (1988). Inferences in reading comprehension. In M. Daneman, G. E. MacKinnon, & T. G. Waller (Eds.), *Reading research: Advances in theory and practice (Vol. 6)* (pp. 177–215). San Diego: Academic.
Singer, M., & Halldorson, M. (1996). Constructing and validating motive bridging inferences. *Cognitive Psychology, 30*, 1–38.
Snow, C. E. (2002). *Reading for understanding: Toward a research and development program in reading comprehension*. Pittsburgh: RAND.
Taraban, R., Rynearson, K., & Kerr, M. (2000). College students' academic performance and self-reports of comprehension strategy use. *Reading Psychology, 21*, 283–308.
Trabasso, T., & Magliano, J. P. (1996a). Conscious understanding during comprehension. *Discourse Processes, 21*, 255–287.
Trabasso, T., & Magliano, J. P. (1996b). How do children understand what they read and what can we do to help them? In M. Graves, P. van den Broek, & B. Taylor (Eds.), *The first R: A right of all children* (pp. 158–181). New York: Teachers College, Columbia University Press.

Trabasso, T., van den Broek, P., & Suh, S. Y. (1989). Logical necessity and transitivity of causal relations in the representation of stories. *Discourse Processes, 12*, 1–25.

van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. *Applied Psychological Measurement, 14*, 1–12.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic.

Wang, D. (2006). What can standardized reading tests tell us? Question-answer relationships and students performance. *Journal of Reading and College Learning, 36*, 21–37.

Whitney, P., Ritchie, B. G., & Clark, M. B. (1991). Working memory capacity and the use of elaborative inferences. *Discourse Processes, 14*, 133–145.

Wolfe, M. B., & Goldman, S. R. (2005). Relationships between adolescents' text processing and reasoning. *Cognition & Instruction, 23*(4), 467–502.