

IN PRESS ---- PLEASE DO NOT QUOTE

A Multidimensional Framework to Evaluate Reading Assessment Tools

Joseph P. Magliano^a, Keith Millis^a, Yasuhiro Ozuru^b, & Danielle S. McNamara^b

Northern Illinois University^a

The University of Memphis^b

Please send correspondence to:

Joe Magliano

Department of Psychology

Northern Illinois University

DeKalb, IL 60115

Email: jmagliano@niu.edu

The assessment of reading comprehension is a critical part of designing and implementing programs that teach reading strategies. For example, assessing students' reading comprehension ability and skills prior to an intervention allows potential weaknesses of an individual reader to be diagnosed. Training can then be adjusted to meet the needs of that reader to maximize the impact of the intervention. Effective interventions should also assess and monitor students' progress in the development or improvement of reading comprehension skills throughout the program. Finally, it is usually necessary to assess the extent to which an intervention is effective in improving comprehension skills. For these reasons, evaluation of reading comprehension assessment tools is considered an important first step in designing and developing reading strategy interventions.

This chapter proposes a multi-dimensional framework for evaluating the appropriateness of reading comprehension assessment tools. Specifically, we claim that the effectiveness of a given reading comprehension assessment tool needs to be evaluated by taking various relevant factors into consideration, such as the assessment purpose, the processes or elements that the assessment is designed to assess, the target examinees, and the texts used in the assessment. A multidimensional framework can be used not only to evaluate existing assessment tools and methods, but also to guide the development of new assessment tools. The framework we present is then used to analyze three methods for assessing reading comprehension: 1) multiple-choice tests of comprehension; 2) short-answer questions designed to measure examinee understanding of the explicit content or the implied situation of a text; and 3) the Reading Skills Assessment Tool (R-SAT). R-SAT was developed by the first and second authors as a method to assess comprehension skills through analysis of "think-aloud" protocols produced by readers while reading texts (Magliano & Millis, 2003). After presenting an overview of each assessment

method, we describe the method and results of a correlation study we conducted to evaluate how strongly measures of selected comprehension skills evidenced in verbal protocols are associated with performance on different types of short answer questions. Through this analysis, we hope to show that the multidimensional framework will play a valuable role when developing new approaches to assess reading comprehension and the use of reading strategies.

Dimensions for Evaluating Assessment Tools

The multidimensional framework of reading comprehension presented here was inspired by the general framework of reading comprehension advocated by Snow (2002), which takes into consideration the reader, texts, and reading activities, all of which are bounded by a socio-cultural context. Similarly, we propose that reading comprehension assessment tools (called assessment tools hereafter) should be evaluated in the light of (1) the reading comprehension processes, products, and activities the assessment tool is designed to observe and measure, (2) the ability levels of the target readers, and (3) the types of texts the tool uses to structure and observe examinee reading performance. With respect to this latter dimension, we stress the importance of using a discourse analysis, such as a causal network analysis (Trabasso, van den Broek, & Suh, 1989) to explicate the underlying structure of the texts used in an assessment tool. These analyses can be invaluable for predicting comprehension processes and products that should reflect various levels of comprehension at specific points in a text (e.g., Magliano & Graesser, 1991, Trabasso & Suh, 1993) The importance of these dimensions is determined by one's assessment goals. By this, we refer to the reason why the assessment is being conducted as well as the aspect of comprehension targeted by the assessment. This may seem an obvious consideration, but we contend that assessment goals will be met to the extent that they are explicit and evaluated as to whether the tool meets those goals.

Processes, Products, and Activities of Comprehension

Comprehension arises from a series of cognitive processes and activities including word decoding, lexical access, syntactic processing, inference generation, reading strategies (e.g., self-explanation), and post-reading activities (e.g., summarization, question asking and answering, argumentation). These contribute to a reader's ability to connect the meaning of multiple sentences into a coherently connected mental representation of the overall meaning of text. These processes give rise to multiple levels of mental representations (Balota, Flores d'Arcais, & Rayner, 1990; Kintsch, 1988; Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983). Although many of these processes can be conceptualized as occurring sequentially on a temporal continuum (Ferreira & Clifton, 1986; Fodor, 1983), many are likely to occur in parallel (e.g., McClelland & Rumelhart, 1981; Wiley & Rayner, 2001), at least for proficient readers.

It is worth noting that theories of comprehension in discourse psychology over the past twenty years focused almost exclusively on processes that occur during reading (e.g., Balota et al., 1990) rather than on comprehension processes that continue after reading. However, comprehension may develop further even after one has finished reading a text (Bartlett, 1932). This is important when one considers educational settings in which students are asked to engage in activities that use knowledge gained from reading for purposes such as answering questions and/or writing essays drawing from multiple sources. These post-reading activities influence the reader's understanding of what was read and generally improve comprehension by helping the reader to reorganize and synthesize the information (see McNamara, O'Reilly, Best, & Ozuru, this volume).

Products of comprehension refer to mental representations resulting from comprehension processes. Theories of discourse processing assume that mental representations of texts contain

multiple levels of meaning (Fletcher, 1994; Graesser & Clark, 1985; Graesser, Millis, & Zwaan, 1997; Kintsch, 1988; Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). Readers construct a representation of the explicit text content, which is referred to as the propositional textbase. This representation contains a network of propositions that capture the explicit ideas contained in a text. The textbase is incrementally constructed in a network as the text is read. Relationships between the textbase propositions are often established when they share an argument (e.g., Kintsch & van Dijk, 1978). However, the textbase is not always sufficient to establish a coherent representation of a text (Giora, 1985). Rather, coherence emerges with the construction of a situation model (Graesser, Singer, & Trabasso, 1994; Magliano, Zwaan, & Graesser, 1999; van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). Readers generate inferences that are based on their world knowledge, which enables them to establish implied relationships between text constituents. As such, the situation model provides an index of text constituents along a number of dimensions, such as agents and objects, temporality, spatiality, causality, and intentionality (Magliano et al., 1999; Zwaan, Langston, & Graesser, 1995; Zwaan, Magliano, & Graesser, 1995; Zwaan & Radvansky, 1998). It is important to note that both the textbase and situation model representation are part of a highly integrated network that reflects the underlying meaning of a text (e.g., Graesser & Clark, 1985). Finally, readers may construct information about the rhetorical structure or agenda of an author (Graesser et al., 1997), but readers may not do so unless they have the explicit goal to do so.

Products and processes can be measured “online” and during reading or “offline” and after reading. For example, if a researcher wants to measure online inference processes involved in constructing the situation model while a student reads, then the researcher should choose measures that can be obtained during reading. These would include sentence reading times,

response latencies to probes presented during reading, and think aloud protocols, each of which have been shown to be valid measures of situation model construction (e.g., Magliano & Graesser, 1991).

In many educational contexts, an assessment of students' reading ability is often inferred on the basis of 'offline' measures, such as answering multiple-choice questions that are presented after the actual reading. It is important to emphasize that these questions are typically answered after an initial reading of the texts, as opposed to directly assessing processes during reading. The decision processes involved in responding to multiple choice questions introduce cognitive processing tasks that are not relevant to online text processing in a non-test context (VanderVeen et al., this volume; Gorin, 2005), thus causing complex interactions between the text, questions, and answers. As such, these tests assess processes and products of comprehension and of question answering, which is not exactly the same as reading comprehension (see Graesser & Clark, 1985, for a similar perspective regarding open-ended questions). Well-constructed questions and options can sometimes rule out knowledge utilization and directly tap what gets constructed on-line, but such designs of question composition are extremely difficult to engineer.

Alternatively, one may be able to engage in a theoretically motivated discourse analysis of items on existing tests in order to determine the extent to which they provide an indirect assessment of the products of online reading processes. Below we discuss the taxonomy for evaluating different types of questions that occur in the Gates McGinitie (G-M) and Nelson-Denny (N-D) tests of comprehension. For example, readers may be asked to evaluate the meaning of a word in the context of a sentence or text, which would measure lexical processing. On the other hand, a question may ask the student to evaluate an inference that relies on their situation model level of understanding. VanderVeen et al. (this volume) developed a similar

taxonomy of questions that occur in the Critical Reading Section of the SAT. They provide strong evidence that their classification of question types on this text can be used to create reader profiles that reflect proficiencies in different aspects of comprehension.

Ability Level of Readers

Reading comprehension is a product of complex interactions between the properties of the text and what readers bring to the reading situation. Proficient readers approach a text with relevant knowledge, word decoding ability, text-based and situation model-based inferencing skills, competency with a variety of reading strategies, metacognitive skills, and so on (Graesser 1997; McNamara & O'Reilly, in press; Oakhill, 1994; Perfetti, 1985, 1994, Snow, 2002). Each of these dimensions has a profound impact on comprehension and may hold implications for the assessment of individuals' reading comprehension ability (Hannon & Daneman, 2001; Oakhill, 1994; Perfetti, 1985, 1994).

Dimensions of an individual's reading ability are likely to vary as a function of literacy education or experience, and alter their contribution to overall reading ability. For example, whereas inadequate proficiencies in early and lower-level processes (e.g., phonological and lexical processes) are a primary reason why beginning readers struggle (Perfetti, 1985; 1988, 1994), there is some evidence that the reading abilities of older children are more closely related to differences in higher-level reading skills such as the ability to make text-based or situation-based inferences, to maintain coherence, to activate higher-order knowledge structures, or to monitor and manage comprehension processes (e.g., Oakhill, 1994; Oakhill & Cain, this volume; Perfetti, 1985; VanderVeen et al, this volume). As a result, assessing vocabulary knowledge and/or word decoding ability to identify at-risk readers may be particularly appropriate during the early stages of literacy training. However, the same assessment tool is likely to fail to

identify at-risk readers among older children. Rather, during later literacy training and secondary education, deficiencies in inference processes and strategic comprehension skills are the major roadblocks for students who are trying to learn new information through reading (Snow, 2002). Thus, the target processes or products of assessment need to be adjusted based on the developmental stage of the target students.

Influence of Text Characteristics

Students read text for different purposes, and reading purposes are closely associated with the text genre. For example, some goals for of reading narrative stories may be to understand the basic sequence of events described, be entertained, and extract some moral or point. On the other hand, the primary purpose of reading expository texts such as science or history texts is to learn or acquire new information about scientific or historical facts about natural/social events. In addition, these two types of texts differ in terms of the novelty of information contained in the text. Thus, the same reader may appear relatively strong or weak depending on the reading situations, which often involve different purposes that are largely associated with the text genres (Best, Rowe, Ozuru, & McNamara, 2005; McNamara, Floyd, Best, & Louwerse, 2004). In order to accurately detect the intra-individual differences in reading comprehension resulting from text/genre effect, it is important to take into consideration the notion that different goals are associated with these different types of texts.

Finally, there is evidence indicating that even within a given genre, text characteristics and individual differences interact in determining the reading comprehension performance of a given individual (McNamara, & Kintsch W., 1996; McNamara, Kintsch E., Songer, & Kintsch, 1996). For example, in the context of scientific texts, McNamara and colleagues have shown that low-knowledge readers comprehend high-coherence texts better than low-coherence texts,

whereas the opposite is true for high-knowledge readers. Hence, this line of research indicates the presence of intra-individual differences in reading comprehension performance as the function of matching characteristics of the texts (e.g., cohesion) and the individual's knowledge level. Readers appear to comprehend science texts optimally when reading a text that poses a moderate level of challenge.

Overall, the discussion presented in this section indicates a rather complex picture of reading comprehension assessment. One may get different pictures of comprehension depending on the combination of assessment tools, age groups, genre of the text used in the assessment, and text characteristics within a genre. We propose that assessment tools must be chosen and developed with the target reading situation in mind. Assessment tools should contain texts and activities that are representative of those that students actually encounter in the non-test context that the assessment is designed to measure.

As an example, the discourse technology group at Northern Illinois University was recently asked to evaluate the reading comprehension portion of the Law School Admissions Test (LSAT), which uses a multiple-choice format. Law students often encounter and produce argumentative texts. They must be able to construct coherent textbase and situation model representations for this type of text. They must also be able to reason beyond those texts and determine their relationship to other arguments that may occur in the context of a legal case. A careful analysis of sample LSAT problems revealed a relatively equal number of questions that assessed readers' ability to construct a textbase, to generate appropriate inferences in the context of a situation model, and to reason beyond the texts. As such, the LSAT measures a variety of processes and products of comprehension that a law school student is expected to master during the course of his or her training to comprehend, interpret, and argue based on legal documents.

A final point about the text materials used in assessment is that researchers should understand and be sensitive to the structural features (e.g., causal and rhetorical structures) of the texts included in the assessment tool because these structural features influence the extent to which readers engage in strategic processing of texts (e.g., Trabasso et al., 1989; Meyers & Wijekumar, this volume). Researchers should engage in some form of discourse analysis that provides an understanding of the features of the texts used in the assessment. These analyses can provide insight into the processes and products that a given text affords.

For example, if one wanted to assess the extent to which readers establish bridging inferences between important text constituents, a causal network analysis (Trabasso, van den Broek, & Suh, 1989) could be administered to determine causal relationships afforded by a text. Such an analysis could be used to identify potential causal inferences that skilled readers should generate while reading a particular text (e.g., Suh & Trabasso, 1993, Trabasso & Suh, 1993). As another example, a propositional network analysis advocated by Kintsch and van Dijk (1978) could also be used to identify the breaks in cohesion that skilled readers should be able to resolve (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara et al., 1996; O'Reilly & McNamara, in press). These analyses could then provide a basis for constructing test activities and items to assess readers' ability to establish coherence during reading.

Assessment Goals

The complexities suggest that selecting an assessment tool should be guided by the specific goal of the assessment. In this section, we discuss issues related to assessment goals in order to provide more specific guidelines for the selection process. Comprehension assessment may occur in any situation in which a researcher or educator is interested in understanding psychological processes or products of reading. Assessment may occur for a variety of purposes:

Evaluating the fluency of on-line processing of materials, assessing the nature of a memory representation, or determining how effectively a student can apply the knowledge gained from a text to a relevant task (e.g., law school students developing an argument based on legal materials). Assessment occurs in variety of contexts that range from laboratory to educational settings. The same assessment techniques are not appropriate in all settings. For example, in the context of discourse psychology research, the primary goal of the assessment may be to identify the nature of inference processes that occur online during reading (e.g. Graesser et al., 1994; Magliano & Graesser, 1991; McKoon & Ratcliff, 1992). As such, researchers have used a variety of tasks that provide measures of reading behavior, such as sentence reading times, eye movements, probe response methodologies (e.g., lexical decision, word naming), and verbal protocol methodologies (e.g., thinking aloud). Many of these methodologies could not be readily implemented in educational settings for both practical and institutional reasons. In addition, an emphasis on test-based accountability has resulted in individual state governments mandating the use of standardized assessment tools (Dwyer, 2005). As a result, alternative assessment approaches for evaluating student achievements used by discourse and school psychologists (Deno, 1985, 1986; Shinn, 1989) may not be readily adopted.

In the context of this volume, it is important to consider the extent to which assessment tools provide a basis for guiding reading strategy interventions. Reading comprehension assessment in the context of strategy interventions can roughly be classified into two categories based on the goals such assessment is designed to achieve: 1) general classification of readers, and 2) diagnosing readers' specific weakness or problems.

The first type of assessment is intended to provide a general classification of readers, rather than providing a detailed diagnosis of specific problems in reading comprehension. For

example, prior to an intervention, students are typically assessed in order to determine if they are at risk or experiencing reading problems. However, once at-risk students are identified, the second type of assessment needs to be administered to afford a more detailed diagnosis of the locus of students' problems within the reading comprehension process. Students may have difficulty at decoding, lexical access, or higher-level comprehension skills, such as inference making. Detailed diagnosis of the students' problems allows educators to determine the type of intervention that can specifically target the weaknesses or the problems exhibited by the students.

Thus, these two types of assessment goals become highly relevant constraints when selecting an appropriate assessment tool because assessment tools vary in terms of fulfilling the requirement associated with these two goals. Many standardized assessment tools have been designed to provide a general assessment of a reader's ability to comprehend text appear to be suited for classifying readers into skilled and unskilled readers who require intervention (VanderVeen et al., this volume). The standardized general reading ability tests, such as the G-M and N-D tests, for example, use a multiple-choice format in which readers comprehend a series of short texts and answer multiple-choice questions regarding different aspects of their understanding of the text. Although these tests are not without their shortcomings (Carver, 1992; Farr, Pritchard, & Smitten, 1990; Hanna & Oaster, 1980; Katz, Lautenschlager, Blackburn, & Harris, 1990), they are quite effective for classifying students because they are readily available, cheap to administer, and have been shown to be reliable and valid assessments of general reading skills (Freedle & Kostin, 1994; Glover, Zimmer, & Bruning, 1979; Malak, & Hageman, 1985; van den Bergh, 1990).

However, many of these standardized tests are not designed to provide a detailed picture of why less-skilled or at-risk students are comprehending texts poorly. Less-skilled students'

performance in these tests could result from problems or deficits within any phase of the reading comprehension processes. This shortcoming is in part related to the fact that the construct of reading comprehension on which these tests are based is not explicitly informed by a substantive psychological model of reading processes based on research in discourse theory. Consequently, these tools have, thus far, shed little light on specific reading deficits and their remediation. If an intervention attempts to tailor training to the needs of an individual reader, it may not be sufficient to merely identify at-risk students (meaning unclear). Rather, one must diagnose an individual reader's deficits, as there are multiple reasons why a student may struggle to read. To meet this goal, a battery of tests tapping lower and higher-level processes would most likely be needed. Nonetheless, it may be possible to make use of constructs in discourse theory to determine which aspects of comprehension are tapped by assessment tools (VanderVeen et al., this volume). In the next section, we describe a taxonomy that was developed to determine the processes and products of comprehension that are measured by the G-M and N-D tests.

Researchers often develop their own assessment tools in addition to these standardized tests. One tool commonly used among discourse researchers is short answer questions that assess memory for the propositional textbase and situation model representations for a text (e.g., Magliano, Todaro, Millis, Wiemer-Hastings, Kim, & McNamara, 2005; McNamara, 2004). For example, the question "What representations are assessed by short-answer questions?" could be used to assess the comprehension of this paragraph. Short answer questions require examinees to access specific aspects of their memory representation produced while reading. To the extent that accessibility of the specific information based on a given cue (i.e., question stem) is largely a function of the processing performed at the time of reading, memory-based short-answer questions may tap the representation formed at the time of reading the passage more directly than

multiple-choice reading comprehension questions. If the questions are presented to the readers after the reading, the questions do not influence the reading process. Short-answer questions require readers to generate the answer themselves based on the question stem, which makes a short-answer question distinct from multiple-choice questions, which can be answered partly on recognition memory, information search in the target passage, and reasoning. It is important to note, that short-answer questions also have limitations. As will be discussed in the following section, readers can sometimes provide correct answers to short-answer questions using strategies that are not always indicative of their comprehension ability. As a consequence, it may be difficult to identify a specific locus of a comprehension problem or deficit.

In the context of strategy interventions, it is important to identify which aspects of comprehension improve as a function of training is critical. It is particularly important to demonstrate that students adopt the reading strategies addressed in an intervention. However, neither multiple-choice reading comprehension questions nor short-answer questions provide a direct measure of online reading processes and strategies. This is especially important given growing evidence that different populations of readers differentially benefit from training (e.g., Magliano et al., 2005; McNamara, 2004; McNamara, O'Reilly, Rowe, Boonthum, & Levistein, this volume). As such, if the goal is to assess how reading behaviors changes as a function of an intervention, alternative measures that are sensitive to reading strategies, such as think-aloud protocols, would need to be adopted.

Think aloud protocols are well suited to assess of the nature of inferential processing and strategies that students use in an attempt to understand a given text. In the context of the second goal of the assessment, which is to diagnose specific weaknesses and problems that readers face in the temporal continuum of reading, we believe that think-aloud protocol analysis may provide

the valuable tool, in addition to other forms of assessment, in determining or designing specific interventions.

Evaluating Assessment Tools

In this section, we use a multidimensional approach to evaluate three different techniques for assessing comprehension. The first technique involves multiple-choice questions, the second, open-ended questions, and the third, think aloud protocols. In particular, we evaluate these three approaches in terms of (1) the processes and products measured by each approach, (2) the nature of the texts in relation to the target reader, and (3) the goals for implementing or developing the assessment tool. We present results of analyses performed on existing corpora of multiple-choice questions (G-M test and N-D test), short answer questions, and think aloud protocols, and compare them to assess relationships between these different approaches.

Multiple-choice Tests of Reading Comprehension

Multiple-choice tests of reading comprehension are arguably the most common of the three approaches. Although it is possible to construct these tests based on discourse theory, many tests are constructed based only on the psychometric properties of the items, and as a result one may be not be appropriate to assess many of the processes and products of comprehension as outlined by theories of discourse comprehension. As such, researchers and educators cannot assume that such tests adequately measure dimensions of interest in reading comprehension. In an effort to rectify this situation, we have developed a taxonomy that classifies assessment questions in terms of the nature of comprehension that the questions assess (i.e., type of processes and resulting representation such as textbase, situation model). We used the taxonomy to classify the questions in two commonly used assessments of general reading comprehension, the G-M and N-D reading comprehension tests. We used Form T of the G-M and Form F of the

N-D, which are both used to assess late adolescent readers' comprehension abilities (grade 12 and College Freshmen).

In an analysis of the N-D and G-M tests, we identified at least three general classes of questions. The question types differ on the processes and products of comprehension that they address. Example questions from the N-D and G-M for each type are shown in Table 1. The first class of questions is *local-textbase questions*. The processes involved in answering these questions are (1) searching and locating the explicit text content in a particular sentence and (2) verifying which answer most closely matches the text content. These questions require minimal if any inferential processes. The product of comprehension most closely associated with this question is the textbase. We consider these questions "local" because students only have to consider one or two adjacent sentences to answer the question.

Insert Table 1

The second class of questions is *global-textbase questions*. This question type differs from the local-textbase questions on the grain size of text that the reader is asked to consider. The answers to local-textbase questions are usually found in one sentence. With global-textbase questions, however, the reader is asked to determine if a phrase or word reflects the thematic meaning of a segment of text longer than one sentence (e.g., several sentences, paragraph, or entire texts). In this sense, the potential answers provide paraphrases or summary statements of the theme that are reflected in multiple propositions in the textbase representation. However, these questions may also require reference to the situation model to the extent that the reader may generate thematic inferences or generalizations of the text segments. In terms of processes, the reader must search and locate the appropriate segments of the texts, construct a summarization of that segment, and then assess which answer option best matches that

summarization. The example global question in Table 1 was classified as such because the reader must consider the entire text to identify the appropriate answer.

The third class of questions is *inference questions*. These questions require processes associated with generating inferences from the texts. The nature of the inference depends on the question. The inferences could be bridging, explanatory, predictive, or elaborative. Some of these questions may assess inferences that are considered to be normally generated while reading the texts (e.g., Graesser et al., 1994). As such, the comprehension product associated with these questions corresponds to the situation model (see Zwaan & Radvansky, 1998 for a review) that proficient readers are considered to form in the normal course of reading. The first example inference question in Table 1 was classified as such because it involves inferring the affective response of characters in a short narrative. Other inference questions may require readers to reason beyond the texts and generate inferences that are specific to the content of the questions. The second inference question in Table 1 is an example of this type of question because it requires the reader to reason beyond the text because it never discusses modern automobiles. It is important to note that not all inference questions in these tests reflect inferences that readers would normally construct when reading the texts. This is particularly the case for the second class of inference questions. Inference questions that tap situation model inferences would need to be carefully constructed in light of discourse comprehension theory. One may consider inference questions to be the most difficult because they require the reading to reason beyond the explicit text. However, the difficulty of these questions will be determined by the extent to which the text supports the inference and similarity of the alternative answer to the correct inference.

We analyzed the frequency of these different types of questions for the G-M and N-D Tests. With respect to the G-M, 56% (N = 27), 13% (N = 6), and 31% (N = 15) of the questions

were local textbase, global textbase, and inference questions, respectively. With respect to N-D, 58% (N = 22), 8% (N = 3), and 33% (N = 13) were local textbase, global textbase and inference questions, respectively. Clearly, both tests primarily measure processes associated with verifying the textbase. However, it is important to note that both tests contain a substantial percentage of questions that assess inferences associated with the situation model.

We also analyzed the passages used in the tests. As discussed earlier, text genre and text structure impact the nature of the examinee-text interaction and should be accounted for by assessment tools. The G-M contains a collection of eleven short texts, five of which were narrative texts and six were expository texts. All but two of the texts contained only one paragraph with a mean length of 122 words. We calculated reading grade level via the Flesch-Kincaid in order to determine if the texts are appropriate for late adolescent and early adult readers. The grade level for the texts ranged from 7.3 to 12.0 and the mean grade level was 10.4. With respect to the N-D test, there were eight texts, all of which were expositions. All texts but one had two to three paragraphs with a mean length of 266 words. The Flesch-Kincaid grade level ranged from 6.6 to 12.0 with a mean of 10.0. The texts seemed appropriate for late high school and college students. In fact, both tests contained outlier texts that lowered the average grade level. Because G-M contained both narrative and expository texts, it is more representative of the texts that students might encounter in academic reading situations in high school.

With respect to the assessment goals associated with these two tests, they are primarily used to identify skilled and less skilled readers (e.g., Magliano & Millis, 2003). However, the analyses of the different question types may allow one to identify proficiencies in constructing local textbase, global textbase, and situation model representations. We assessed the possibility

that performance on the different question categories is correlated with short-answer questions that require access of the textbase or situation model representations.

Short-answer Tests of Reading Comprehension

Another approach to assessing comprehension is using short answer questions (e.g., Magliano et al., 2005; McNamara & Kintsch, 1996). In a typical short-answer question, readers read text, and they answer the questions after they read a text without looking back at the text. Answers can range from a single word to several clauses. Answering questions in this type assessment requires a reader to access the memory representation for a text and retrieve and produce the relevant information.

The questions are typically designed to provide an assessment of the quality of either the explicit textbase or the situation model. The appendix contains an example text and questions. The textbase questions require the reader to retrieve information that can be found in a single sentence of the text. For example, question 10 for the text on the Franco Dictatorship is “In what year did the Franco dictatorship end?” Its answer can be found in the last sentence. The majority of the situation model questions assess the extent to which readers have inferred causal relationships between text events. For example, consider question 7, “What were the causes of the great period of economic stagnation that followed World War II?” This question can be answered by the content of several sentences in the fourth paragraph of the text. The extent to which readers can answer these questions should be related to the extent to which they generated the causal inferences. It is important to note, however, that these questions measure memory for both explicit content and inferred relationships.

Of course, there are some limitations of using short answer questions. First, readers might have known the answer before reading the text. In this case, the test question would not be

measuring their comprehension ability but rather their prior knowledge. Second, one cannot completely rule out the possibility that comprehension occurs as the readers search their memory representation as they try to answer the question. Their memory representations might be sufficient for the reader to successfully use reasoning and guessing strategies. The last limitation concerns the scoring process. Scoring the answer requires identification of the ideal answers that might contain several parts, as is the case with situation model questions. Each participant's answer must be classified with respect to the percentage of the important parts present within the answer. This type of practical limitation becomes a large factor when one needs to assess general reading skill for a large number of students in a short period of time prior to an intervention.

Think Aloud/Verbal Protocols

The first and second authors are developing an alternative to multiple-choice and short answers for assessing reading comprehension. The assessment tool is called the Reading Strategy and Assessment Tool (R-SAT) and uses verbal protocols. R-SAT is designed to measure comprehension strategies associated with different standards of coherence. Standards of coherence refer to a reader's criteria or general sense of the importance of forming a coherent representation, especially of how different parts of a text are related to one another (van den Broek, Risen, & Husebye-Hartman, 1995). As stated above, deep comprehension arises with the construction of coherent textbase and situation model representations. However, van den Broek et al. (1995) argued that readers differ in the extent to which they have a drive to achieve coherent representations. Some readers will accept disjointed representations of the explicit propositions contained in texts, whereas other readers attempt to construct a representation that contains coherent relationships between those propositions.

Our past research has consistently demonstrated that think-aloud protocols can reveal a reader's standards of coherence. Magliano and Millis (2003) had participants think aloud when reading selected sentences in simple narratives (i.e., Chinese folktales). Less-skilled readers, as identified by the N-D test, talked more about the target sentence than skilled readers, who talked more about how those sentences were related to the prior discourse context. Thus, skilled readers' protocols show their effort to maintain larger or more global coherence whereas less-skilled readers' protocols show their tendency to focus on target sentences in isolation. Magliano and Millis used latent semantic analysis (LSA: Landauer & Dumais, 1997) to provide a quantitative assessment of these strategies. LSA provides a measure of overlap between any two units of language by computing cosines between their vector representations in a high dimensional semantic space. These cosines typically range from 0 to 1.0 and represent the degree of conceptual relations between the linguistic elements. Magliano and Millis computed cosines between the think-aloud protocols of a target sentence and two semantic benchmarks (i.e., the target sentences and causally important prior text sentences), and analyzed participants' recall performance for different texts that were read silently in terms of these two types of LSA cosines. The analysis indicated that recall performance decreased as the function of the LSA cosine between the think aloud protocol and the target sentences that was just read (large conceptual overlap between the targets sentence and the protocol), whereas the recall increased as the function of the LSA cosine between the protocol and the causally important sentences (large overlap between causally important sentence and the protocol). Thus, two lines of results (i.e., relation between protocol and reading skill, and relation between protocol and subsequent recall) converge, establishing that R-SAT is capable of revealing readers' on-line processing involved in the maintenance of text coherence.

In R-SAT, readers are provided with questions about the content of the sentence (e.g., “Why did the battle fail?”) and questions to facilitate think-aloud (“What are you thinking right now?”) while reading. The questions appear immediately following the presentation of pre-selected sentences. These sentences were pre-selected based on the presence of strong causal connections to prior portion of the text. Correct and complete answers to both types of questions require that the reader generate the appropriate causal bridging inferences at that point in the text. With R-SAT, readers produce their answers and thinking aloud protocols by typing them into the computer. LSA and word matching algorithms are then used to (1) assess the completeness of the answers by comparing them to ideal answers, and (2) assess the extent to which the think-aloud protocols conceptually overlap with the current target sentence and causally important target sentences.

Overall, R-SAT provides important information to assess reading comprehension in terms of both process and product of comprehension. With respect to processes of comprehension, R-SAT assesses the extent to which readers are generating causal bridging inferences based on think-aloud protocols produced in response to a think-aloud question. With respect to products of comprehension, R-SAT assesses the quality of the textbase and situation model representations based on the readers’ answer to the think-aloud question.

Discourse analyses of the texts used in R-SAT are a central component to its development. Specifically, a causal network analysis (Trabasso et al., 1989) was used to identify local, distal direct, and distal indirect causal consequences in the prior discourse context. This analysis provides a basis for constructing the important prior text information that is compared to the think-aloud protocols in order to assess the extent to which readers are generated bridging inferences.

Finally, to be consistent with the multidimensional framework, it is important to assess the intended goal for R-SAT. We believe that R-SAT is appropriate for providing a general assessment of reading skill to the extent that reading skill is influenced by a reader's standard of coherence. Another goal of R-SAT is that it will be integrated into iSTART, a reading intervention designed to increase standards of coherence by teaching students how to self-explain as they read (see McNamara et al., this volume). We foresee that iSTART training will be tailored to individual readers' needs based on assessments provided by R-SAT. For example, students who show low standards of coherence as indicated by R-SAT will be taught the rudiments of self-explanation via iSTART. However, the training for students who show relatively high standards prior to training could focus on fine tuning self explanation, such as determining the appropriate sentences to self explain. The general approach used in R-SAT has also been used to assess ongoing progress during practice sessions in iSTART (e.g., Magliano et al., 2002). Specifically, the self-explanations produced during practice are compared to (1) the sentence just read, (2) prior discourse, and (3) concepts related to but not present in the discourse context. Success in training can be assessed by measuring the extent to which the protocols overlap with these "benchmarks". For example, students whose thoughts overlap with the sentence just read and not with either the prior discourse or concepts from world knowledge are most likely not explaining the text, but rather only paraphrasing or repeating the text. As a final goal of R-SAT, we have used this general approach to assess whether readers change their reading behaviors after training (Magliano et al., 2005). We have found that, in general, readers' thoughts overlap more with the discourse context after training than prior to training. More specifically, self-explanation protocols after training contain more concepts from the current sentence and prior discourse than prior to training. This suggests that after training readers were

explaining how the current sentence fit into the larger discourse context. R-SAT could provide a basis for making more detailed evaluations regarding changes in reading strategies as a function of iSTART.

Assessing Convergence between the Different Assessment Tools

In this section, we present two sets of analyses designed to evaluate relations between the different assessment approaches described in this chapter. The aim of these analyses is to determine the extent to which these assessment tools measure the intended processes and products of comprehension. The first analysis was conducted to determine convergences between the G-M and short-answer tests, whereas the second was designed to assess convergence between R-SAT, short-answer tests, and the N-D test

Convergence between G-M and short answer performance.

As described above, the G-M test is comprised of three general types of questions that address different aspects of comprehension. Local questions tap explicit ideas in the textbase representation, global questions tap thematic ideas in the textbase and perhaps situation model representations, and inference questions address the quality of situation model level representation. We had 223 college freshmen and sophomores take the G-M test of reading comprehension. They also read two texts and answered 10 short-answer questions for each immediately after reading the texts. Five of the 10 questions were textbase questions and the other five were situation model questions. The appendix contains one of the texts and questions that were used in the study. The textbase question could be answered via the content of a single sentence, whereas the situation model questions require readers to infer causal relationships across text sentences.

We assessed performance on the local, global, and inference questions on the G-M. One would expect that performance on the local questions to be the best, followed by global questions and inference questions based on theories of discourse comprehension that assume that the textbase reflects a more shallow level of understanding than the referential situation model. With respect to these latter two categories, one would expect that the inference questions would be harder to answer if these questions required one to consider the deeper, underlying meaning or implications of the discourse. However, the mean percentage of questions answered correctly was 66%, 60%, 76%, for local, global, and inference questions, respectively. A one-way ANOVA revealed that these means were significantly different from one another, $F(2, 610) = 94.92, p < .05$. Post hoc analysis (Tukey) revealed that inference questions were answered more accurately than local questions, which were in turn answered more accurately than global questions. These results suggest that these questions may not be measuring deep comprehension, contrary to common assumptions. It is important to note that this conclusion is limited to the G-M. For example, our discourse analyses of the inference questions on the LSAT suggest that these questions do require one to have a deep comprehension of the texts on that test, although we do not have data to empirically support this claim as of yet.

The short answer questions were scored based on the proportion of the key ideas present in each answer. We then added up the proportion scores for each participant, yielding the total number of questions answered completely out of a total possible score of 5. Finally, we calculated the average textbase and situation model questions answered correctly over the two texts. As one would expect, textbase questions ($M = 2.01, SD = .99$) were answered more accurately than situation model questions ($M = 1.51, SD = .88$), $t(1, 223) = 9.75, p < .05$.

One interesting question is how processes used to answer short answer questions (i.e., textbase and situation model questions) and multiple-choice questions are related. One reasonable expectation would be that answering local and global questions in the G-M test should be more closely related to performance on textbase questions because these questions require an intact textbase for them to be answered correctly. It may also be the case that the local questions, as opposed to global questions, on the G-M test better reflect the skills necessary to answer the textbase questions rather than the global questions because answering the local questions require test takers to identify specific propositions in a text and evaluate which answer best reflects that proposition. This process is conceptually similar to processes required for answering textbase short answer questions. The other important prediction is that performance in answering the G-M inference questions should account for most of the variance in performance answering the situation-model short answer questions, because inference questions are supposed to tap readers' ability to construct a situation model by generating inferences. Table 2 presents bivariate correlations between the short answer and multiple choice questions.

Insert Tables 2 and 3

In order to pursue these questions concerning the relations between these two types of assessment tools (i.e., multiple choice questions in the G-M test and short answer questions), multiple regression analyses were conducted in which the predictor variables for each regression analysis were the percentage of local, global, and inference questions in G-M test answered correctly. We performed two multiple regressions, with one using the textbase short-answer question performance and the other using the situation-model short-answer question performance as the criterion variables. The Beta weights and R^2 can be found in Table 3. The regression analysis on textbase short answer questions accounted for a significant 31% of the variance (F

(2, 219) = 33.40, $p < .05$). Performance on local ($t(222) = 4.27, p < .01$) and inference questions ($t(222) = 2.23, p < .05$) were significant predictors of performance on textbase questions. As one would expect, performance on local questions appeared to be the strongest predictor of performance on the textbase short answer questions. This is in line with our prediction that answering textbase questions involves accessing a specific proposition, essentially similar to the process underlying the local question answering in G-M test. Quite surprisingly, performance on the global questions was not a significant predictor of performance on the textbase questions. On the other hand, inference questions were indicative of performance on the textbase short answer questions. We will discuss the implications of these findings shortly.

The regression analysis on situation model questions accounted for a significant 41% of the variance ($F(2, 219) = 53.166, p < .05$). Performance on local ($t(222) = 3.77, p < .05$), global ($t(222) = 4.64, p < .05$), and inference questions ($t(222) = 2.97, p < .05$) were significant predictors of performance on situation model questions. The beta weight suggests that the different G-M questions were comparable predictors of performance on these questions. It is interesting to note that global questions were significant predictors of short-answer situation model questions, but not textbase questions. This suggests that there is indeed some overlap in the processes required to construct and evaluate a thematic representation of a discourse and constructing and accessing the referential situation model. Moreover, the situation model questions required readers to access relationships between distal texts sentences that are strongly implied by the texts. This is an aspect of comprehension that likely involves both constructing a coherent textbase and a referential situation model.

Relations between R-SAT, short answer, and N-D performance. Millis et al. (2005) conducted a study to assess the extent to which R-SAT is related to constructing textbase and

situation model representations. Specifically, they were interested in the extent to which overlap between the protocols and the different semantic benchmarks is related to answering short-answer questions that assess the textbase or situation model representation. Three semantic benchmarks were used in the study: current sentence, local sentences (immediately prior sentences) and distal causal sentences. As discussed earlier, overlap with the current sentence and local sentences should reflect processes involved in constructing a local textbase representation for explicit ideas whereas overlap with the distal causal sentences should be related to processes associated with constructing a coherent global textbase and building a situation model representation. They also assessed the relationship between R-SAT and the performance in N-D tests; that is, they examined which of the three types of R-SAT scores, as represented by the protocol's conceptual overlap with current sentence, immediate prior sentence, or distal causal sentence, predicted performance on the N-D test. Given that the majority of questions in N-D test tap local processes, one would expect that overlap with the benchmarks associated with local processing (current and local sentences) should be most indicative of performance on this test.

Participants thought aloud after every sentence while reading science texts that had an average grade level of 6.8. They calculated LSA cosines between think-aloud protocols for a given sentence and the three benchmarks, and computed an average cosine for each benchmark for each participant. Participants answered textbase and situation model short answer questions for the text and also took the N-D test of reading comprehension. The average cosines for the three benchmarks were used as predictor variables in three regression analyses predicting performance on the short-answer textbase questions, short-answer situation model questions, and the N-D multiple-choice reading ability questions. The bivariate correlations between the

variables are presented in Table 4 and the resulting standardized coefficients and R^2 s are presented in Table 5. Each equation was significant, indicating that the LSA cosines predict performance on these three types of reading comprehension questions. As in Magliano and Millis (2003), the cosines for the current sentence were negatively correlated with the question answering performance, whereas the cosines for prior causal antecedents were positively correlated with the question answering performance. This general pattern indicates that comprehension of the text was best when the verbal protocols contained information related to the causal antecedents as opposed to when the protocol primarily contained information related to the current sentence.

Insert Tables 4 and 5

The pattern of relations between the protocol and local or distal benchmarks is largely consistent with the expectation based on the theory of discourse processing. Specifically, overlap with the current sentence was negatively correlated with the two outcome measures that most closely reflect understanding of ideas in the explicit textbase, namely textbase questions and the N-D. Although this relationship is consistent with Magliano and Millis (2003), this may seem counterintuitive because one would expect that the explicit ideas would be better represented when readers talked about them while thinking aloud. However, this relation may exist because readers who tend to talk more about the current sentence may be doing so at the expense of constructing important bridging inferences to the prior discourse context. As a consequence, this may result in formation of isolated or fragmentary representations of the text. Because accessing specific textbase content in memory is partly a function of the connections between the textbase content and other related information in memory, failing to draw these bridging inferences is likely to cause retrieval problems at the time of answering the short answer textbase questions.

Indeed, overlap with the distal causal sentences was related not only to performance on the situation model questions which tapped the readers' understanding of causal relationships between text constituents but also to textbase question answering performance. These results indicate that the construction of globally coherent representation with the support of causal bridging inference is critical for retrieving a variety of textual information. The results also indicate, as expected, that overlap between the protocol and the local sentence is positively correlated with overall N-D test performance. This finding converges with our earlier observation that majority of questions in N-D test (58%) are local questions. It is however, important to note that the bivariate correlations between the benchmarks are high and could be causing suppression effects. We are currently exploring ways to minimize these correlations.

Finally, we also examined the unique variance accounted for by N-D and the LSA analysis of the verbal protocols in predicting the textbase and situation model scores. The LSA cosines predicted 15% and 16% of the variance of the textbase and situation model scores, respectively, whereas the N-D only predicted 3% and 6% percent. We believe that these differences indicate that R-SAT is more sensitive than N-D to the processes necessary to construct a coherent representation of a text, which primarily consist of questions that address a readers' ability to locate and verify explicit text content. Of course, one should note that the LSA values were computed from the same texts that the short answers were based upon, and therefore, do not constitute the strongest case for this claim.

Conclusions

In this chapter, we presented a multidimensional framework to evaluate assessment tools which is based on Snow's (2002) general framework of reading comprehension. This framework involves an assessment of the products and processes measured by a given tool, the intended

reader, text characteristics, and the overarching assessment goals. There are many tools that can be used to assess comprehension. However, not all tools will be sufficient to meet one's assessment goals. For example, if one wants to directly assess a reader's ability to engage in deep reasoning about a text, then an assessment tool that primarily taps processes associated with constructing a textbase representation would not be appropriate. Our goal was to provide a framework for evaluating the utility of an assessment tool given one's assessment goals.

We have illustrated the utility of this framework by evaluating existing assessment tools: multiple choice reading comprehension questions (e.g., G-M and N-D tests), short answer reading questions, and R-SAT based on think-aloud protocols. With respect to the multiple-choice reading comprehension questions, our corpus analysis of G-M and N-D tests of comprehension using a classification scheme based on the theory of discourse processing indicated that the tests contain questions that assess readers' proficiencies in constructing the local textbase, global macrostructure, and inferences. One concern with respect to the G-M was that participants answered inference questions more accurately than local or global questions with respect to the G-M test. If these questions did indeed tap inferences associated with deep comprehension, then one would have expected these questions to be the most difficult. Indeed, it appears that the global questions were the most challenging. It is important to note that G-M did account for an impressive amount of variance in performance on short answer questions that addressed the textbase and underlying situation model.

It remains to be seen as to whether our discourse analysis of the questions in the G-M could be used to diagnose specific comprehension problems. VanderVeen et al. (this volume) provides an overview of the kind of research necessary to pursue this endeavor with respect the Critical Reading Section of the SAT. Although these researchers did not explicitly employ the

multidimensional framework advocated in this chapter, their approach is consistent with this framework.

We are considerably more skeptical regarding the N-D test of reading comprehension. Millis et al. (2005) found that the N-D test accounted for relatively little variance in performance on short answer questions that involve both textbase and situation model representations. Furthermore, the extent to which readers established distal bridging inferences was not predictive of performance on the N-D. Based on these findings and the analysis of the type of questions contained in the N-D test, we are inclined to conclude that the N-D test may not be an ideal assessment tool if one wants to use a tool to assess a reader's ability to construct a representation that reflects deep meaning. Indeed, to find statistically significant differences between skilled and less skilled readers on various comprehension tasks, we have either used a quartile split (e.g., Magliano & Millis, 2003) or greatly shortened the length of allotted testing time from that recommended by the test publishers (e.g., Magliano et al., 2005; McNamara & McDaniel, 2004).

On the other hand, the results reported here from Millis et al. (2005) suggest that the reading strategies revealed in think-aloud protocols are indicative of inferential processes and products underlying deep level comprehension, at least to the extent that this measure correlates with one's ability to answer short answer questions in a systematic and theoretically predictable way. In general, we find that when readers primarily paraphrase the sentence (evidenced by a large overlap between protocol and current sentence), they tend to have more difficulty answering questions that require access to a textbase representation. The findings also indicated that readers who mention information about the prior text when thinking aloud tend to perform better not only on questions that tap the situation model representation but also on those that tap the textbase representation. We believe that this pattern of results supports the conclusion that

individual differences in strategies produced during thinking aloud reflect readers' standards of coherence (Magliano & Millis, 2003). Readers who tend to talk about how the current sentence is related to the prior texts have a higher standard of coherence than readers who tend to only talk about the current sentence; as a consequence, they can construct a more globally coherent representation of the text content that supports access to a variety of information contained in or implicated by the text.

Our evaluation of assessment tools based on the multi-dimensional framework suggest the R-SAT to be quite useful tool for assessing the readers' ability to engage in effective processing that results in coherent representation of the text content. By combining R-SAT with the analysis of text properties (e.g., causal network, referential cohesion, word level analysis), it is possible to obtain detailed pictures of the strengths and weaknesses of individual student's reading comprehension processes; whereas some students fail to draw inference when texts do not provide sufficient cues on causal structure, other students may fail when the target sentences include a unfamiliar word or have a complex syntactic structure. For these reasons, R-SAT shows promise as an approach for assessing reading comprehension, in particular when detailed assessments of the readers are needed to design an intervention or to assess the effect of a particular reading intervention program.

Certainly, the current form of R-SAT is still in its infancy stage, and requires further development. Several plans to improve R-SAT are in order. First, as seen in the Table 5, current R-SAT predicts approximately 20 % of variance ($R^2 = .17$ and $.22$) of the textbase and situation model questions answering performance, respectively. This performance is much poorer compared to the G-M test's ability to predict the short answer question performance ($R^2 = .31$ and $.42$) for textbase and situation model questions, which we obtained in a separate study

relating G-M test and short answer question answering performance. To improve the performance of R-SAT, we are currently adopting an approach of identifying specific sentences that are predictive of reading skills and strategies and using only those sentences in R-SAT. We are also exploring the extent to which we can identify specific comprehension strategies used by skilled and less skilled readers, which one could not readily do with traditional multiple-choice tests. It is our hope that we can use this information as a basis for guiding remediation.

R-SAT has only been tested with adult readers so far. We have tested it with texts well below the reading level of these readers (e.g., Magliano & Millis, 2003) and with more difficult scientific texts (Millis et al., 2005), and obtained similar results. Although, R-SAT could be implemented with younger readers, there is one limitation. Students must be reasonably proficient at typing their thoughts; typing could interfere with students' reading and think-aloud processes if they do not have proficient typing skill. If a version were to be developed for younger readers, it would likely have to be based on orally produced responses.

In summation, in the first half of chapter we outlined the multidimensional framework for evaluating reading comprehension assessment tool by drawing on recent developments in cognitive psychology and discourse processing research. Further, we evaluated three types of the existing reading comprehension assessment tools using the multi dimensional framework. In the second half of the chapter, we reported recent empirical studies on the relations between three types of the assessment tools: multiple choice reading comprehension questions, short answer questions, and R-SAT based on think-aloud protocols. Overall, empirical findings confirm the validity of theoretical inquiry of the strength and weakness of the three types of assessment tool within multidimensional framework. Although more research is required to refine the

framework, this framework appears to be useful for evaluating and improving tools for reading comprehension assessment.

References

- Balota, D.A., Flores d' Arcais, G. B., Rayner, K., (1990). Comprehension processes in reading. Lawrence Erlbaum Associates. Mahwah, NJ: Lawrence Erlbaum Associates.
- Bartlett, F. C. (1932). *Remembering*. Cambridge: Cambridge University Press.
- Best, R M., Rowe, M. P., Ozuru, Y., & McNamara, D.S. (2005). Deep-level comprehension of science texts: The role of the reader and the text. *Topics in Language Disorders, 25*, 65-83.
- Carver, R.P. (1992). What do standardized tests of reading comprehension measure in terms of efficiency, accuracy, and rate? *Reading Research Quarterly, 27*, 346-359.
- Deno, S.L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232.
- Deno, S.L. (1987). Formative evaluation of individual programs: A new role for school psychologists. *School Psychology Review, 15*, 358-374.
- Dwyer, C.A. (2005) (Ed.). Measurement and research in the accountability era. Mahwah, NJ: Erlbaum.
- Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27*, 209-226.
- Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory & Language, 25*, 348-368.
- Fletcher, C.R (1994). Levels of representation in memory for discourse. In M.A. Gernsbacher (Ed.), *Handbook of Psycholinguistics*, (589-607). San Diego, CA: Academic Press.
- Fodor, J.A. (1983). *Modularity of the mind*. Cambridge, MA: MIT Press.

- Freedle R., & Kostin, I. (1994). Can multiple-choice reading tests be construct valid? *Psychological Sciences*, 5, 107-110.
- Giora, R. (1985). Notes towards a theory of text coherence. *Poetics Today*, 6, 699-715.
- Glover, J.A., Zimmer, J.W., & Bruning, R.H. (1979). Utility of the Nelson-Denny as a predictor of structure and thematicity in memory for prose. *Psychological Reports*, 45, 44-46.
- Graesser, A. C., & Clark, L. F. (1985). *Structures and procedures of implicit knowledge*. Norwood, NJ: Ablex
- Graesser, AC., Millis, K.K., & Zwaan, R.A. (1997). Discourse Comprehension. *Annual Review of Psychology*, 48, 163-89.
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101, 371-395.
- Gorin, J. S. (2005) Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, 42, 351-373.
- Hanna. G.S., & Oaster, T.R. (1980). Studies of the seriousness of three threats to passage dependence. *Educational & Psychological Measurement*, 40, 405-411.
- Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in component processes of reading comprehension. *Journal of Educational Psychology*, 93, 103-128.
- Katz, S., Lautenschlager, G., Blackburn, A., & Harris, F. (1990). Answering reading comprehension items without passages on the SAT. *Psychological Sciences*, 1, 122-127.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163-182.

- Kintsch, W., & van Dijk, T.A. (1978). Towards a model of text comprehension. *Psychological Review*, 85, 363-394.
- 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. *Cognition and Instruction*, 3, 251-283.
- Magliano, J.P., Todaro, S. Millis, K.K., Wiemer-Hastings, K., Kim, H.J., & McNamara, D.S., (2005). Changes in reading strategies as a function of reading training: A comparison of live and computerized training. *Journal of Educational Computing Research*, 32, 185-208.
- Magliano, J. P., Zwaan, R. A., & Graesser, A. C. (1999). The role of situational continuity in narrative understanding. In S. R. Goldman & H. van Oostendorp (Eds.), *The construction of mental representation during reading*, (pp. 219-245). Mahwah, NJ: Erlbaum.
- Malak, J., & Hegeman, J.N. (1985). Using verbal SAT scores to predict Nelson-Denny scores for reading placement. *Journal of Reading*, 28, 301-304.
- McClelland, J.L., & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part 1., An account of basic findings. *Psychological Review*, 88, 375-407
- McKoon, G. & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, 99, 440-466.
- McNamara, D.S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, 38, 1-30.

- McNamara, D. S., Floyd, R. G., Best, R., & Louwse, M. (2004). World knowledge driving young readers' comprehension difficulties. In Y. B. Yasmin, W. A., Sandoval, N. Enyedy, A. S. Nixon, F. Herrera (Eds.), *Proceedings of the sixth international conference of the learning sciences: Embracing diversity in the learning sciences* (pp. 326-333). Mahwah, NJ: Erlbaum.
- McNamara, D.S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes, 22*, 247-287.
- McNamara, D.S., Kintsch, E., Songer, N.B., & Kintsch, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1-43.
- McNamara, D.S., & O'Reilly, T. (in press). Theories of comprehension skill: Knowledge and strategies versus capacity and suppression. In F. Columbus (Ed.), *Progress in Experimental Psychology Research*. Hauppauge, NY: Nova Science Publishers, Inc.
- McNamara, D.S., & McDaniel, M. (2004). Suppressing irrelevant information: Knowledge activation or inhibition? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*, 465-482.
- Millis, K.K., Magliano, J.P., & Todaro, S. (2005). Measuring discourse-level processes with verbal protocols and latent semantic analysis. Manuscript submitted for publication.
- Oakhill, J. (1994). Individual differences in children's text comprehension. In M.A. Gernsbacher (Ed.), *Handbook of psycholinguistics*, (pp. 821-848). New York: Academic Press.
- Oakhill, J., & Cain, K. (this volume). Issues of Causality in Children's Comprehension Development. To appear in D. S. McNamara (Ed.). *Reading Comprehension Strategies: Theory, Interventions, and Technologies*. Mahwah, NJ: Lawrence Erlbaum Associates.

- McNamara, D.S., O'Reilly, T., Best, R., Ozuru, Y. (this volume). A reading strategies framework. In D. S. McNamara (Ed.). *Reading Comprehension Strategies: Theory, Interventions, and Technologies*. Mahwah, NJ: Erlbaum.
- McNamara, D.S., O'Reilly, T., Rowe, M., Boonthum, C., Levinstein, I. (this volume). iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. In D. S. McNamara (Ed.). *Reading Comprehension Strategies: Theory, Interventions, and Technologies*. Mahwah, NJ: Erlbaum.
- Perfetti, C.A. (1985). *Reading Ability*. New York: Oxford Press.
- Perfetti, C.A. (1994). Psycholinguistics and reading ability. M.A. Gernsbacher (Ed.), *Handbook of psycholinguistics*, (pp. 849-894). New York: Academic Press.
- Shinn, M.R. (Ed), (1989). *Curriculum-based measurement: Assessing special children*. The Guilford Press: New York.
- Snow, C. (2002). *Reading for Understanding: Toward an R&D Program in Reading Comprehension*. RAND: Santa Monica, CA
- Suh, S., & Trabasso, T. (1993). Inferences during reading: Converging evidence from discourse analysis, talk-aloud protocols, and recognition priming. *Journal of Memory and Language*, 32, 279-301.
- Trabasso, T., & Suh, S. (1993). Understanding Text: Achieving explanatory coherence through online inferences and mental operations in working memory. *Discourse Processes*, 16, 3-34.
- Trabasso, T., van den Broek, P. & Suh, S. (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 den Broek, R., Ridsen, K., & Husebye-Hartman, E. (1995). The role of readers' standards for coherence in the generation of inferences during reading. In R.F. Lorch & E.J. O'Brien (Eds.), *Sources of coherence in reading*, (pp. 353-374). Hillsdale, NJ: Erlbaum.
- van Dijk, T. A., & Kintsch, W. (1983). *Strategies in Discourse Comprehension*. New York: Academic press.
- Wiley, J. & Rayner, K. (2000). Effects of titles on the processing of text lexically ambiguous words: Evidence from eye movements. *Memory and Cognition, 28*, 1011-1021.
- Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science, 6*, 292-297.
- 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.
- Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185.

Appendix

Francisco Dictatorship

The Franco dictatorship lasting from 1936-1975 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 four 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.

Francisco, 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.

- 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 lose power? TEXTBASE

Table 1.

Example questions for local, global, and inference questions.

TYPE	STEM	ANSWER OPTIONS
Local	The audience size mentioned was	<ul style="list-style-type: none"> a. five thousand b. eight thousand c. twelve thousand d. sixteen thousand e. twenty thousand
Global	This passage is mainly about how one kind of spider	<ul style="list-style-type: none"> a. excavates a tunnel b. traps its food c. fools it's enemies d. builds with silk
Inference	When Elizabeth's parents were watching the show, they were	<ul style="list-style-type: none"> a. impressed b. nervous c. ashamed d. proud of themselves
Inference	The passage suggests that the teacher would have thought that today's cars are	<ul style="list-style-type: none"> a. easier to drive than Model T's b. more fun than Model T's c. more like locomotives than Model T's d. more complicated to fix than Model T's

Table 2.

Correlation matrix for the textbase, situation model, local, global, and inference questions.

	TB	SM	L	G
Textbase(TB)				
Situation Model (SM)	0.68			
Local (L)	0.54	0.57		
Global (G)	0.31	0.49	0.44	
Inference	0.50	0.55	0.73	0.42

Table 3.

Regression Beta Weights Predicting Short Answer Comprehension Performance from the Local, Global, and Inference Questions on the G-M test of Reading Comprehension.

SHORT ANSWER		
QUESTION TYPE		
G-M PREDICTORS	Textbase	Situation Model
Local	.36**	.29**
Global	.07	.27**
Inference	.21*	.22**
R ²	.31**	.42**

Note: * $p < .05$; ** $p < .01$

Table 4.

Correlation matrix for the textbase questions, situation model questions, current sentence, benchmark, local causal benchmark, and distal causal benchmark.

	ND	TB	SM	CS	L
Nelson-Denny (ND)					
Textbase (TB)	0.28				
Situation Model (SM)	0.24	0.48			
Current Sentence (CS)	-0.33	0.00	0.11		
Local (L)	-0.02	0.21	0.17	0.67	
Distal	0.01	0.36	0.36	0.50	0.65

Table 5.

Regression Beta Weights Predicting Measures of Comprehension (Textbase and Situation ModelShort Answer, and Nelson Denny) from LSA Variables.

LSA Predictors	Measures of Comprehension		
	Textbase	Situation Model	Nelson Denny
Current sentence	-.36**	-.06	-.61**
Local cause	.09	.01	.29*
Distal cause	.45**	.41**	.11
R ²	.22**	.17**	.19**

Note: * $p < .05$; ** $p < .01$