Read Writ (2013) 26:349–379
DOI 10.1007/s11145-012-9372-9

Examining the structure of reading comprehension: do literal, inferential, and evaluative comprehension truly exist?

Deni Basaraba · Paul Yovanoff · Julie Alonzo · Gerald Tindal

Published online: 10 April 2012

Abstract Although the recent identification of the five critical components of early literacy has been a catalyst for modifications to the content of materials used to provide reading instruction and the tools used to examine student's acquisition of early literacy skills, these skills have not received equal attention from test developers and publishers. In particular, a review of early literacy available measures for screening and monitoring students reveals a dearth of tools for examining different facets of reading comprehension. The purposes of this study were twofold: (a) to examine the relative difficulty of items written to assess literal, inferential, and evaluative comprehension, and (b) to compare single factor and bifactor models of reading comprehension to determine if items written to assess students' literal, inferential, and evaluative comprehension abilities comprise unique measurement factors. Data from approximately 2,400 fifth grade students collected in the fall, winter, and spring of fifth grader were used to examine these questions. Findings indicated that (a) the relative difficulty of item types may be curvilinear, with literal items being significantly less challenging than inferential and evaluative items, and (b) literal, inferential, and evaluative comprehension measurement factors explained unique portions of variance in addition to a general reading comprehension factor. Instructional implications of the findings are discussed.

D. Basaraba (✉)
Center on Teaching and Learning, University of Oregon, 5292 University of Oregon, Eugene, OR 97403-5292, USA
e-mail: basaraba@uoregon.edu

P. Yovanoff
Annette Caldwell Simmons School of Education and Human Development, Southern Methodist University, 3101 University Blvd., Suite 345, Dallas, TX 75205, USA

J. Alonzo · G. Tindal
Behavioral Research and Teaching, University of Oregon, 5262 University of Oregon, Eugene, OR 97403-5262, USA

Keywords Reading comprehension · Assessment · Elementary reading

Introduction

The identification of five critical components of reading—phonological awareness, alphabetic understanding, fluency with connected text, vocabulary, and reading comprehension—by the National Reading Panel (National Institute of Child Health and Human Development [NICHD], 2000) has been a catalyst for change in the content of materials used to provide reading instruction and to examine students' acquisition of early literacy skills, as well as the educational policy that recommends the focus of instruction in the classroom. Examination of Reading First legislation, for example, reveals the expectation that sufficient reading instruction be focused on the following components: (a) providing students with the skills and knowledge that the sounds of spoken language (phonemes) correspond with printed letters (graphemes); (b) teaching the skills and strategies needed to decode unfamiliar words; (c) providing students with multiple opportunities to practice reading connected text fluently (e.g., with automaticity, accuracy, and prosody); (d) providing sufficient background information and vocabulary to foster reading comprehension; and (e) supporting the development of appropriate active strategies to construct meaning from print (No Child Left Behind Act, 2001). However, review of widely-used formative assessment systems (e.g., Dynamic Indicators of Basic Early Literacy Skills (DIBELS), AIMSweb, and easyCBM) reveals that the majority of these brief assessments focus primarily on the first three aforementioned critical components of early literacy.

Despite policy recommendations, not all five critical components of early literacy instruction have received attention by educational publishers and researchers. A review of the reading literature reveals that a wealth of research exists examining the contributions of phonological awareness (Adams, 1990; Bus & van Ijzendoorn, 1999; Smith, Simmons, & Kame'enui, 1998), alphabetic understanding (Chard, Simmons, & Kame'enui, 1998; Fien et al., 2008; Good, Baker, & Peyton, 2009), and reading fluency (Baker et al., 2008; Good, Simmons, & Kame'enui, 2001) to later reading success. Significantly less research, however, has been conducted on reading comprehension (Davey, 1988; Stothard & Hulme, 1996). Despite the fact that reading comprehension is viewed as "the essence of reading" (Durkin, 1989), a recent review of the literature conducted by the National Reading Panel shows that reading comprehension has only started to receive scientific attention in the past 30 years (NICHD, 2000). One potential explanation for this paucity of reading comprehension research (e.g., what it is, how to measure it, and instructional strategies to improve it) is that it is a more cognitively complex task than the precursor skills required to facilitate and support it. Foundational skills that have received much attention in the literature include the importance of being able to correctly identify sounds and blend them to form words (National Research Council [NRC], 1998), and the integral relationship between automaticity and fluent reading to be able to comprehend what is read (Perfetti, 1985). Reading comprehension not only depends on these skills, it also depends on an *interaction* between the reader and the text (NICHD, 2000; Sweet, 2005), something that is difficult to teach.

Changing perspectives on reading comprehension

Reading comprehension can be defined generally as the ability to extract meaning or learn from text (Rupley & Blair, 1983; Snow, 2002). This general definition, however, encourages acceptance of an outdated, simplistic view of reading comprehension as a skill that results from the independent, sequential development of hierarchically ordered lower-level skills, such as phonological awareness, alphabetic understanding, and fluency (i.e., students are able to comprehend only once they have developed accurate, automatic word reading skills and can read connected text with some degree of fluency) (Dole, Duffy, Roehler, & Pearson, 1991). These foundational reading skills, however, work in conjunction with other skills, such as (a) automaticity, (b) higher-level language comprehension processes, (c) background knowledge and schema construction, (d) knowledge of text structures, and (e) the capacity of different memory structures to support general reading comprehension.

Automaticity with decoding and word-reading, for example, is hypothesized to be a prerequisite for general reading comprehension because when students are able to decode words effortlessly and automatically (i.e., without devoting significant cognitive resources to identifying letter-sound correspondences), they have freed up additional cognitive resources that can be applied to understanding the meaning of words, phrases, and sentences within text (LaBerge & Samuels, 1974; Perfetti, 1985; Perfetti, Landi, & Oakhill, 2005) Understanding of words in isolation, however, is not sufficient for comprehension because the meaning of many words is often dependent upon the context within which they appear.

Thus, students also need to have a firm grasp on higher-level language comprehension processes, such as understanding the semantic, syntactic, and referential relationships among successive words to construct meaning from text (Hanon & Daneman, 2001). Moreover, students' understanding of a text is also influenced by their prior knowledge (Rupley & Willson, 1996), their ability to incorporate that prior knowledge to create a *schema*, or organized understanding of the world that can be readily applied to the text being read (Anderson, 2004; Gernsbacher, Robertson, Palladino, & Werner, 2004; Kintsch & Kintsch, 2005; Van Dijk & Kintsch, 1983; Zwaan, Radvansky, Hilliard, & Curiel, 1998), and their ability to use relevant prior knowledge to create a schema to support their understanding of that text (Cain, Oakhill, Barnes, & Bryant 2001; Kintsch, 1988; McNamara, 1997).

In addition, familiarity with text structures and the demands of texts from different genres, such as narrative and informational texts that utilize different text structures, are also useful in supporting reading comprehension (Gersten Fuchs, Williams, & Baker, 2001). Because narrative and informational texts differ in terms of their purpose (Duran, McCarthy, Graesser, & McNamara, 2007; Weaver & Kintsch, 1991), it is not surprising that they also differ in terms of their structure (Best, Floyd, & McNamara, 2008; Fox, 2002). With regard to purpose, narrative texts are crafted to tell a story and entertain while informational texts are designed to communicate information to the reader so that he or she might learn something;

different text structures are required to accomplish these different purposes. To accomplish their goal, narrative texts frequently use structures that are familiar to everyday life, such as setting-conflict-resolution, causal event chains, or the use of story grammar structures to describe episodes or events (Otero, León, & Graesser, 2002; Weaver & Kintsch, 1991). Informational texts, in contrast, typically have varied structures that align with the intended purpose of the text (e.g., classification, compare/contrast, procedural description; Weaver & Kintsch, 1991), to introduce new concepts and ideas with which the reader may not be familiar (Best et al., 2008), and are commonly organized in terms of a hierarchy of propositions (or idea units) that relate to a central subject (Tun, 1989). Research indicates that it is not only possible to teach students the basic elements of common text structures (e.g., setting, main character(s), plot, conflict, and resolution) but that knowledge of text structures supports comprehension of narrative (Boulineau, Fore, Hagan-Burke, & Burke, 2004; Dimino, Gersten, Carnine, & Blake, 1990; Short & Ryan, 1984;) and informational texts (Taylor & Beach, 1984).

More recently, the development of reading comprehension has been considered the result of emerging expertise with lower-level (e.g., decoding, fluency) *and* higher-level reading skills, taking into account the varying levels of complexity that comprise reading comprehension as students are expected to interact with a text to different degrees (Dole et al., 1991; Kintsch & Rawson, 2005). Each of the tasks involved in understanding a text—whether it is simply to recall what is stated in the text (literal comprehension), to interpret the authors' meaning through connecting information that is implicit in the text (inferential comprehension), or to go beyond the text by relating what is being read to prior experiences and knowledge (evaluative comprehension)—requires a different level of cognitive processing by the reader. This perception of comprehension not only has instructional implications related to enabling teachers to target the types of comprehension instruction they provide to students, it can also be used by test developers to increase the relevance of information provided by tests.

Increasingly complex levels of comprehension

The idea that there are different levels of reading comprehension, each of which imposes different cognitive demands on the reader and requires varying levels of interaction with the text, is not new (Herber, 1970; Snider, 1988; McCormick, 1992; Pearson & Johnson, 1978). Its prevalence is evidenced by the way in which the theory of levels of comprehension underlies numerous recommended teaching practices and instructional texts over the last four decades (Carnine, Silbert, Kame'enui & Tarver, 2010; Herber, 1970; Lapp & Flood, 1983; Vacca et al., 2009). This theory proposes a continuum of reading comprehension skills in which a student must first proficiently engage in tasks of literal comprehension before engaging in deeper interactions with the text, such as those prompted by inferential and evaluative understanding (Herber, 1970). More specifically, literal comprehension tasks require readers to simply retrieve information that has been explicitly stated in a passage (Carnine et al., 2010). Inferential comprehension tasks require readers to understand relationships that may not be explicitly stated in the passage but are essential for passage understanding, such as the connection between two events in a narrative or understanding a character's motive for a particular action (Applegate, Quinn, & Applegate, 2002). Evaluative comprehension tasks require readers to analyze and critically interpret the text based on their prior knowledge and experiences.

This theory, (Herber, 1970) however, is not merely based on the levels of interaction readers have with a text, but also on the type of information they are expected to contribute to the types of questions that are posed in assessments of reading comprehension (Leu & Kinzer, 1999; Rupley & Blair, 1983). Each of these levels of comprehension (literal, inferential, and evaluative) and the cognitive demands they place on the reader warrant further exploration and require detailed discussion.

Literal comprehension: the "bare bones" of reading comprehension

Literal comprehension, the first level of comprehension, requires that a student be able to extract information that is explicitly stated in a passage (Carnine et al., 2010; Lapp & Flood, 1983; McCormick, 1992). This level of understanding is dependent upon students' word-level processing skills, or their ability to accurately identify individual words and understand the meaning created by the combination of words into propositions and sentences (Perfetti, Landi, & Oakhill, 2005). Although these word-level processing abilities are requisite skills for understanding a text, they alone are not sufficient for facilitating comprehension (NRC, 1998); deeper interactions with the text are necessary. According to Rupley and Blair (1983), literal comprehension is composed of two strategies: recall, or the ability to provide an idea (e.g., main idea or detail) that was part of a passage; and recognition, or the ability to recognize whether specific information is provided in a passage. It is not sufficient, in other words, to simply remember a fact stated in the passage. The fact must also be recognized as existing within the context of a passage to determine whether or not comprehension has actually occurred; otherwise it is unclear if the reader comprehended what was read or relied on prior knowledge and understanding.

This focus on the textually-explicit information that even beginning readers can readily access helps explain why literal comprehension is the focus of the skills and strategies initially introduced to all readers, especially in the primary grades, when they are being taught to read with understanding (Carnine et al., 2010). Because literal comprehension tasks typically require only that a student locate information that is explicitly stated in the text (sometimes even using the same phrasing or wording that appeared in the text), the cognitive processing demands for proficient readers may be fairly minimal; students will need to be able to decode and understand the words and be able to locate words or phrases that appear in the text. Although literal comprehension is undoubtedly important (without surface-level understanding of a text, deeper interactions with the text are not possible), those designing and providing instruction and developing tests must also recognize that literal understanding is a stepping-stone to more advanced comprehension skills that must also be examined to continue to see growth in student performance (Kintsch & Rawson, 2005; Nation, 2005).

Inferential comprehension: making meaning from the text

Inferential comprehension can be viewed as a logical extension of the recognition step of literal comprehension proposed by Rupley and Blair (1983) in that readers are required to go beyond recognizing that facts are derived from a passage to actually *interacting* with a text to make inferences about meanings not explicitly stated in the text (Applegate, Quinn, & Applegate, 2002; Snider, 1988). At this stage, it is no longer sufficient for the reader to recognize and understand what the author has said. Instead, the reader is required to manipulate information in the text to search for relationships among the main idea and details and to use that information to interpret and draw conclusions about the author's intended meaning (Vacca et al., 2009), fill in omitted details, and/or elaborate upon what they have read (Dole et al., 1991). Each of these tasks, not surprisingly, places greater cognitive processing demands on readers, as they are required to hold some information presented in the text in working memory while searching for other information presented elsewhere in the text.

These relationships between objects, events, or details within the passage are more frequently than not implied in the text, thereby requiring readers to "read between the lines" to make their discovery (Carnine et al., 2010; Leu & Kinzer, 1999). This interaction with the text, in which readers bring to the text their own background information and draw connections between pieces of information presented in the text enable them to construct a situation model of what the text is about (Graesser, Singer, & Trabasso, 1994; Kintsch & Rawson, 2005; Perfetti, 1999). This situation model of the text is more complex than the text base, which is a simple representation of the propositions of the text, or a literal understanding of each word as it relates to those around it in the text (Perfetti, 1999) and therefore requires that students be able to (a) comprehend the text written on the page (literal understanding), (b) interpret meanings, arguments, or claims that are presented across the text (inferential understanding), and (c) apply their own background knowledge and prior to experience to the text to facilitate or enhance understanding.

Much of the research conducted thus far on reading comprehension has examined the role of inferences because they are at the "heart of the comprehension process" (Dole et al., 1991). Readers are required to make different types of inferences, such as text-based inferences (also known as text-connecting inferences) and knowledge-based (or gap-filling) inferences to understand the text. Text-based, or causal inferences, for example, are those that are required to establish coherence within a text (Perfetti, 1999; Perfetti et al., 2005). The type of coherence the reader needs to establish may be local coherence, or cohesion between elements, constituents, and references of adjacent clauses, or global coherence, which can be seen as cohesion between larger chunks of information within a text. Text-based inferences, especially those required to establish local coherence, are those that frequently are needed to keep the representation of the text base minimally coherent (Perfetti, 1999). These types of inferences, however, are not sufficient to fully understand the text being read. More complex inferences, such as knowledge-based (or gap-filling) inferences that draw on a reader's knowledge to help represent and understand the relationships between persons or events described in the text (Kintsch & Rawson, 2005; Oakhill & Cain, 2007) are also needed for understanding.

The role of working memory Because inferential comprehension tasks require more thorough processing and the integration of ideas presented in the text with prior background knowledge (Pearson & Fielding, 1991), some researchers have examined the role of working memory processes in reading comprehension. More specifically, researchers have proposed that reading comprehension ability is not only predicted by lower-level language skills (Cain, Oakhill, & Bryant, 2004; Masson & Miller, 1983; Perfetti, 1985), but is also the function of readers' working memory, or the ability to hold recent information in their memory, to integrate that information into coherent representations, and to integrate information acquired from the recently read text with information stored in long term memory (Cain et al., 2004; Swanson & O'Connor, 2009; van den Broek, Tzeng, Risden, Trabasso, & Basche, 2001b). Newly acquired information from the text must first be processed in working memory and held there while information in long term memory is accessed and compared (Masson & Miller, 1983).

Working memory, then, plays a central role in all three levels of comprehension. Literal comprehension tasks, for example, require readers to compare information contained in the words of the text with the words stored in their mental lexicon to ensure that the meaning of each word can be accessed and understood within the context of a sentence (or larger piece of text). Evaluative comprehension tasks pose demands on working memory by requiring readers to compare newly acquired information to prior knowledge or experiences in an effort to determine how this new information relates to their understanding of the topic being presented in the text. Researchers (Cain, Oakhill, Banes, & Bryant, 2001; Oakhill, Samols, & Hartt, 2005) have argued, however, that working memory capacity is particularly important to inferential comprehension because inferences can only be made when the general knowledge needed to make them is readily available (Cain et al., 2001; Oakhill et al., 2005). As students move from understanding text at a literal level to being able to make inferences, the demand placed on their working memory increases. This additional cognitive load associated with inferential comprehension helps explain why students might find questions targeting inferential comprehension more challenging than those targeting the more basic literal level of understanding.

Other factors influencing the ability to make inferences In addition to differences in working memory capacity, other factors may also affect one's ability to make any of the multiple types of inferences that are required for text comprehension. Among these factors are reading skill, accurate understanding of the demands and goals for the reading task, and background knowledge relevant to the topic of the text (van den Broek, Lorch, Linderholm, & Gustafson, 2001a). Just as processing ability and attentional resources are needed to literally understand what is being said in the text, so too are these resources required for accurate and appropriate inference generation. Research indicates, for example, that because the basic skills of younger readers are typically less automated than those of older readers, younger readers may experience more difficulty answering questions that require inferential thinking (van den Broek et al., 2001a). Language proficiency may also impact the number and kinds of inferences readers generate (van den Broek et al., 2001a;

Zwaan & Brown, 1996); Zwaan and Brown (1996), for example, compared reader's comprehension of narratives in their first language (English) with comprehension of narratives in their second language (Dutch) and found that readers not only engaged in fewer higher-level processes, such as inference generation, when reading texts in their second language, but also that they generated fewer associative and elaborative inferences to enhance their understanding of the text than they did when reading texts in their first language.

Evaluative comprehension: extending beyond the text

The third and most complex level of reading comprehension proposed by the levels of comprehension theory is evaluative comprehension (a.k.a. critical or applied understanding). Evaluative comprehension can be seen as an extension of the knowledge, skills, and strategies required of literal and inferential comprehension tasks. This extension is evidenced by the fact that the reader is required to understand the text written on the page (literal comprehension), make interpretations about the author's intended meaning and/or understand the relationships between the elements presented in the text (inferential comprehension), *and* subsequently analyze or evaluate the information acquired from the text in terms of prior knowledge or experiences (McCormick, 1992; Rupley & Blair, 1983) or knowledge that is imported from outside of the text (Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001). According to Herber (1970), at the evaluative level readers are required to juxtapose what they have read in the text with their own prior knowledge and experience, a juxtaposition that creates new meanings and/or relationships that extend beyond the scope of the text. The creation of these new meanings and relationships involves a myriad of different skills including divergent thinking, critical analysis, synthesis, and evaluation (Vacca et al., 2009), in addition to affective, or personal and emotional responses, when necessary (Rupley & Blair, 1983). No longer, in other words, is understanding dependent solely on information that is presented in the text, whether explicitly stated or appearing across multiple clauses in the text; instead, students are required to hold information that is presented in the text in their working memory and simultaneously access information, knowledge, or experience from their long term memory to analyze or evaluate what they have just read, thereby increasing the demands played on their cognitive processing. Drawing on the more basic levels of understanding, evaluative comprehension is posited to demand more of the reader and thus represents higher-order understanding of text.

Prior research examining the levels of comprehension theory

In addition to research examining the differences in the proposed levels of comprehension independently, research studies have also explicitly examined the relative degrees of difficulty of literal and inferential comprehension questions when administered together to groups of students. Snider (1988), for example, conducted a study with junior high students identified as Learning Disabled by state regulations to determine whether there were any observed differences in student performance on questions classified according to the taxonomy proposed by Pearson and Johnson (1978). Using this taxonomy, questions were classified as being: (a) textually explicit, where the answer was stated in the passage (i.e., literal); (b) textually implicit, requiring readers to use information that was stated in the paragraph to answer a question (i.e., inferential); or (c) scriptally implicit, requiring readers to rely on and integrate their prior background knowledge with the information presented in the text (i.e., evaluative; Chikalanga, 1992; Davey, 1988; Snider, 1988). Students' responses to the multiple-choice questions associated with each of the 24 passage sets (each passage set contained one textually explicit, textually implicit, and scriptally implicit passage) revealed that textually explicit (i.e., literal) questions were the easiest for students to answer, followed by textually implicit (i.e., inferential) and scriptally implicit (i.e., evaluative) questions; textually implicit questions were the most difficult for students to answer.

Others have obtained similar results regarding the relative difficulty of items designed to target literal, inferential, and evaluative comprehension (Davey, 1988; McCormick, 1992). Davey (1988) employed the taxonomy for question type proposed by Pearson and Johnson (1978) of textually explicit, textually implicit, and scriptally implicit questions to examine whether the location of response information and inference type could explain any of the observed variance in students' performance on a standardized reading comprehension measure. The location of response information was used as a proxy for question type (i.e., literal, inferential, and evaluative) in that question information and the correct answer were found within the same sentence for textually explicit questions, textually implicit questions required readers to combine information across sentences, and scriptally implicit questions required readers to integrate information from the text with their own background knowledge (Davey, 1988). Results of regression analyses revealed that the location of response information accounted for approximately 27 % of the unique variance observed in the performance of struggling readers and 12 % of the variance for proficient readers.

Although she did not employ the taxonomy of question types proposed by Pearson and Johnson (1978), McCormick (1992), in her work with fifth grade students identified as struggling readers, observed statistically significant differences in the amount of literal and inferential questions they were able to answer correctly. Specifically, while students in her study were able to answer, on average, 70 % of literal questions correctly, they were only able to answer 61 % of inferential questions correctly, implying that these questions were more difficult for them. Together, the findings from these studies provide empirical support for the idea that questions may be written to specifically target different levels, or types of reading comprehension and demonstrate clear differences in student performance on questions purposefully designed to assess different levels of reading comprehension.

Criticisms of the levels of comprehension theory

Given the complexity of reading comprehension, it is not surprising that criticisms have emerged regarding the levels of comprehension theory. Specifically, critics of this theory have argued that a serious flaw is its simplification of this complex process by assuming a linear progression of difficulty between the levels of comprehension. In addition, they argue with the premise that one skill is unquestionably a prerequisite for the next skill in the progression (Lapp & Flood, 1983; Vacca, Vacca, & Gove, 1987). It is also worth noting that much of the information available regarding the levels of comprehension theory can be found primarily in reading instruction textbooks and is not widely supported empirically. Although there appears to be agreement in the field that comprehension tasks differ in their levels of cognitive complexity and processing demands placed on the reader, questions remain as to whether the relationship among the levels is hierarchical or whether processes occur in conjunction with and support one another.

Questions about item format: are multiple choice questions appropriate?

Due to the complex nature of reading comprehension, questions have also been posed about which methods are the most appropriate for assessing such a multi-faceted construct and for distinguishing among *levels* of that construct (Allington, Chodos, Domaracki, & Trueax, 1977; Campbell, 2005; Freedle & Kostin, 1999; Sarroub & Pearson, 1998; Tal, Siegel, & Maraun, 1994). Numerous criticisms have been made, for example, about the frequent reliance on multiple-choice item formats on the basis that: (a) they do not promote student reflection or interactive learning (Sarroub & Pearson, 1998), (b) they do not require students to have read or comprehended the texts accompanying the test items to answer them correctly (Allington et al., 1977; Coleman, Lindstrom, Nelson, Lindstrom, & Gregg, 2010; Freedle & Kostin, 1999; Tuinman, 1973), (c) they tap primarily lower-level cognitive skills (Allington et al., 1977), and (d) their use can contribute to systematic differential test performance for different groups of learners (Tal et al., 1994).

Additional criticisms include the arguments that multiple-choice comprehension items reflect outdated perceptions about thinking and learning processes, the ability to select a correct response from provided options may have little to do with students' reading comprehension ability, and multiple-choice item formats assume there is a single, correct answer to a question that can be identified (Campbell, 2005). Others have noted, however, that not only do multiple-choice comprehension tests require students to use the same intellectual abilities as less-structured tests (van den Bergh, 1990), but also that valuable information about students' performance on items designed to target different *levels* of reading comprehension can be obtained from multiple-choice item formats (Rupp, Fern, & Choi, 2006). Moreover, examination of students' item-level performance on multiple-choice reading comprehension measures can provide valuable instructional information beyond that which can be obtained by an overall, global comprehension score (Alonzo, Basaraba, Carriveau, & Tindal, 2009; Tal et al., 1994).

Although, as Sarroub and Pearson (1998) point out, the underlying theoretical rationale for the reliance on standardized, multiple-choice comprehension items is lacking, this item format has had a long-standing history in education (Campbell, 2005; Pearson & Hamm, 2005). Despite this fact, researchers over the last several decades have been prompted to compare student performance on multiple-choice reading comprehension items to other item formats (e.g., cloze tasks, maze tasks, recall) to determine if multiple-choice items are sufficiently targeting the complex cognitive skills associated with reading comprehension (Fitzgerald & Fitzgerald, 1978; Kendall, Mason, & Hunter, 1980; van den Bergh, 1990). While some (Fitzgerald & Fitzgerald, 1978; Kendall et al., 1980) have found that multiple-choice items are consistently easier than other item formats for students, others (van den Bergh, 1990) have found that multiple-choice items require use of the same cognitive processes to answer correctly as other item formats and therefore may be both appropriate *and* efficient for assessing students' reading comprehension abilities.

In summary, review of the available literature about reading comprehension reveals inconclusive findings regarding which item format may be the most appropriate and whether item format fundamentally alters the construct being measured. Advocates for modifying existing reading comprehension assessments or designing new ones put forth a call for assessments that: (a) reflect the developmental and dynamic nature of comprehension (Sweet, 2005), (b) adequately represent the *interaction* between the reader and the text (Sweet, 2005), (c) provide information that has instructional utility for teachers (e.g., can be used for informed instructional planning and decision-making) (Pearson & Hamm, 2005; Sweet, 2005), and (d) are technically adequate and meet the requirements of psychometric theory (Pearson & Hamm, 2005; Sweet, 2005). As Campbell (2005) points out, however, perhaps what is most important is not selecting one item format over another but rather addressing the aforementioned issues and creating assessments that appropriately reflect this cognitively complex construct.

Purpose and research questions

The purpose of the current study was twofold: (a) to examine the structure of reading comprehension as represented by items written specifically to assess students' literal, inferential, and evaluative comprehension abilities, and (b) to examine the extent to which literal, inferential, and evaluative comprehension questions differ in terms of their difficulty. Corresponding to these objectives, we used formative assessment data from a commonly used curriculum based measurement (CBM) system, to explore the following research questions: (a) Do literal, inferential, and evaluative comprehension questions differ significantly in difficulty?; (b) Does a three-factor model composed of literal, inferential, and evaluative factors fit the data well?; and (c) Can specific item types (i.e., literal, inferential, and evaluative items) explain any additional variance observed in performance on formative, multiple-choice reading comprehension measures above and beyond a general reading comprehension factor? Additionally, we conducted a cross-validation study to determine the degree to which the findings replicated across multiple independent samples of students. Our exploration of the three-factor model as one suited for the data was based on the surface-level structure of the multiple-choice reading comprehension measure in which items were written to assess students' literal, inferential, and evaluative comprehension abilities. Although we had no a priori hypothesis regarding the amount of variance explained by literal, inferential, and evaluative comprehension items, our hypothesis regarding item difficulty was that items sampling literal comprehension would be the least difficult, items sampling evaluative comprehension would be the most difficult, and items sampling inferential comprehension would fall somewhere in between.

Method

Participants

We used two independent samples of fifth grade students attending schools in eight school districts in Oregon to examine our research questions. The first sample ($n = 1{,}210$) was used to estimate the models and determine the relative difficulty of the items, and the second sample ($n = 1{,}217$) was a replication sample used in an effort to validate the estimated model. In all, 47 % of students in the first sample were female, 14 % were English Learners, and 10 % were eligible to receive special education services. The second sample was composed of fifth grade students attending schools in the same eight school districts in Oregon. In this second sample, 52 % of students were female, 11 % were English Learners, and 10 % were eligible to receive special education services. Demographically, the samples are very similar because the larger sample was randomly split (female 47.3, 52 %; ELL 14.3, 11 %; special education eligibility 9.8, 10 %) into two separate samples. All students in both samples had scores for the benchmark measures administered in the fall, winter, and spring of fifth grade to determine whether students were on track for meeting benchmark goals for reading comprehension performance.

Measures

The data used in this study were collected from multiple-choice reading comprehension (MCRC) measures developed as part of the online easyCBM progress monitoring system (Alonzo, Tindal, Ulmer, & Glasgow, 2006). Students completed the comprehension assessment by first reading one fictional, narrative passage of approximately 1,500 words and then answering 20 selected response, multiple-choice questions that related to the passage. The stories and questions were written at a reading level appropriate for the middle of the fifth grade and students completed the assessment as part of a universal screening battery administered in the fall, winter, and spring of fifth grade. Alternate forms of the assessments were designed to be of equivalent difficulty (Alonzo & Tindal, 2008, 2009). Students took one of the equivalent forms in the fall, another in the winter, and a third in the spring.

All students who participated in this study read fictional narratives and answered the 20 multiple-choice questions written to accompany those passages. These computer-based assessments were untimed and students were asked to read a passage and then select a correct response from one of three possible response options that were designed to include: (a) the correct response, (b) a near distractor written to be enticing to students who might not have read the story closely or who might be basing their response on their prior experience or knowledge, rather than on information provided in the text, and (c) a far distractor that might contain some of the same words as the text but was intended to be clearly less correct than either of the other two answer options. Students had the option to refer to the passage as often as needed to answer the multiple-choice questions.

Researchers used three methods during instrument development to ensure that the reading levels of each passage and accompanying questions were appropriate: (a) the *Flesh Kincaid Readability Scale* in Microsoft Word, (b) the *EDL Core Vocabularies in Reading, Mathematics, Science, and Social Studies* (Taylor, Frackenpohl, & White, 1989) to ensure that vocabulary was grade-level appropriate, and (c) verification of grade-level appropriateness of the passage and questions by veteran teachers (Alonzo, Liu, & Tindal, 2007). In addition, the comparability of alternate forms and appropriateness for use with fifth-grade students has been studied through traditional test–retest and alternate form reliability studies, as well as through Generalizability Theory studies and Item Response Theory (IRT) test information curves (Alonzo & Tindal, 2008, 2009). Each passage was accompanied by 20 multiple-choice questions, seven of which were written to assess students' literal comprehension abilities, seven to assess inferential comprehension, and six to assess evaluative comprehension. Literal questions asked students to identify a specific event from the text, inferential questions required students to infer implicit meaning from the text, and evaluative questions required students to evaluate a situation presented in the fictional narrative and make a judgment; examples of these questions types are presented in Table 1.

Examination of available technical adequacy information for these fifth grade comprehension measures reveals moderate bivariate correlations ranging from $r = 0.55$ to 0.64 across the three time points (i.e., fall, winter, and spring) and Cronbach's alpha ranging from .70 to .75 (Saéz et al., 2011) More recent examinations of concurrent predictive criterion related evidence, however, with several state reading assessments (e.g., Oregon Assessment of Knowledge and Skills [OAKS] and the Washington Measures of Student Progress [MSP]) indicate that the multiple-choice reading comprehension measures may account for sizable portions of unique variance observed in scores on these outcomes measures (anywhere from 27.6 to 31.5 %; Anderson, Alonzo, & Tindal, 2011a, b; Saéz et al., 2011). Correlations with these same state assessments are moderate, ranging from 0.52 to 0.64, which is acceptable considering their intended purpose is to provide educators with a periodic snapshot of students' likelihood of meeting certain performance criteria on these state assessments.

Additionally, care was taken to ensure that each MCRC item met rigorous criteria associated with the one-parameter logistic Rasch model employed during the measurement design process, including: (a) a full range of item difficulties (e.g., challenging, moderate, and easy items), (b) small standard errors associated with item difficulties, (c) mean square outfit values within an acceptable range of 0.50–1.50, and (d) careful analysis of item distractors (Alonzo & Tindal, 2008).

Table 1 Examples of each question type and the text where the correct answer* is found

Comprehension level	Question stem	Answer options	Where in the text is the answer found
Literal	What was an example in the story of why Jeremy didn't like the rule about coming to dinner immediately when called?	A. He was playing a game at home with his friends and had to ask them to leave B. He had to leave a basketball game just when he started making shots* C. He wasn't always hungry when his parents decided to have dinner	No matter what Jeremy was doing, he had to stop immediately and run to the table. One time he was in the middle of a basketball game with his friends. They had only been playing for 15 min, and Jeremy's shot was just warming up. He had missed his first five shots, but had just made his last three. He was starting to get into a great rhythm. Then he heard the dreaded yell. His friends groaned as Jeremy dropped the ball and sprinted toward home, muttering under his breath the whole way
Inferential	Why did the dinner rule about everyone sharing events of their day drive Jeremy crazy?	A. He didn't like it when his sister would talk on and on about things that Jeremy thought were silly* B. His parents never gave Jeremy a chance to finish telling everything that had happened to him C. Jeremy's mother was only interested in what his sister had to say about all of her friends	The final rule related to dinner was that everyone must take a turn sharing the events of their day. This rule drove Jeremy crazy. His sister Lexi was quite the talker. She was only 7 years old, and every day she had some long story to share about her day. It always had to do with one of her friends, and she never seemed to tire of talking about the silliest things. Jeremy's parents would just sit there and smile, nodding their heads, listening to Lexi share her story
Evaluative	What will Jeremy probably do the next time his family has dinner at home?	A. Obey the family dinner rules* B. Refuse to eat everything on his plate C. Eat everything on his plate except vegetables	Before long, Jeremy started to compare his own life to the lives of these people. For the first time ever, he started to think about how much worse things could be if his parents were not able to make dinner for the family every night. He thought about the arguments he had with his parents over their rules about coming to dinner on time, eating everything on his plate, and sharing stories about their days. He looked around the room at the people sitting in long rows, eating their dinner with strangers. For the first time, Jeremy began to understand why his parents had their dinner time rules After the last person was served, it was time for Jeremy's family to eat. They went through the line, and Jeremy didn't even fuss about serving himself some vegetables. They took their plates and sat down at a table. His mother smiled at them and then said, "So tell me about your day, Jeremy." Jeremy couldn't help but smile back. For the first time in memory, he thought maybe he had something to say worth sharing

* Correct answer

Analyses

Comprehension item difficulties were estimated with a one-parameter logistic Rasch model. Characteristic of all logistic item response models, the item difficulties are estimated as the log-odds of an examinee with specific ability responding correctly. This model is most appropriate for this research. Among other attractive features of the Rasch model, item difficulties can be compared directly, irrespective of the examinee ability level (this is not true for proportion correct item difficulties, which depend on the ability level of examinees). More specifically, the probability of a correct response is modeled as a logistic function of the difference between the person (ability) and item (difficulty) parameters.

The Rasch item difficulties were estimated with a bifactor model in Testfact (Wood et al., 2003). The bifactor model is thought to be a natural alternative to the conditionally independent unidimensional IRT model in which all items are assumed to be measuring the same underlying latent construct. Instead, the bifactor model allows for conditional dependence among identified groups of items (Gibbons & Hedeker, 1992) or when items are hypothesized to have a two-level structure with a general, underlying latent factor that is represented by all items (Chen, West, & Sousa, 2006; Gibbons et al., 2006; Simms, Gröss, Watson, & O'Hara, 2008). We chose to estimate a bifactor model based on the theoretical rationale that a general reading comprehension factor (composed of such skills as automaticity, language comprehension, background knowledge, etc.) would account for the commonality of the items while several domain specific factors (a combination of literal, inferential, and evaluative comprehension) would account for significant amounts of the item covariances. Finally, central to this research, we were interested in examining the domain specific factors as well as the general reading comprehension factor (Chen et al., 2006). To fully evaluate the effectiveness and utility of the bifactor model, the results of each bifactor model were compared to a single-factor model that hypothesized a single underlying latent factor (i.e., general reading comprehension) accounted for variance observed in MCRC scores in the fall, winter, and spring of fifth grade. In addition, we compared the results of each bifactor model with three domain-specific factors to a three-factor second-order factor model estimated in Mplus (Muthén & Muthén, 2010) that was composed of three same domain-specific factors (i.e., literal, inferential, and evaluative comprehension) but did not include a separate general reading comprehension factor.

We conducted three separate analyses to answer our first research question regarding statistically significant differences in item difficulties for items written to assess students' literal, inferential, and evaluative comprehension. Treating the 20 items as the unit of analysis, each analysis used the Rasch item difficulties as the dependent variable. First, using item type as the independent variable, we conducted one-way Analyses of Variance (ANOVAS) for each of the three benchmark probes and each of the two samples to determine if there were statistically significant differences in the mean difficulties of each item type. This resulted in six independent analyses.

Then, due to the small number of items ($n = 20$) included in the analyses, we also conducted a series of non-parametric tests to test differences in the mean and median ranks of the three item types. After ranking all items with respect to item difficulty, the mean ranks for the three items were compared statistically using a Kruskal–Wallis test of the mean and median item ranks to examine significant relations between rank order and item type. The ordering of these statistical tests allowed us to test for overall differences in the mean ranks of the three item types, test for overall differences in the median ranks of the three item types, and then follow-up with pair-wise comparisons for tests of the mean and median ranks to determine which item types were significantly different.

Results

Descriptive statistics for the item difficulties obtained from the bifactor model for literal, inferential, and evaluative comprehension items on the fall, winter, and spring multiple-choice reading comprehension measures for two samples of students (independent and replication samples) are reported in Table 2. Examination of these data reveals that for both samples of students and all three measures literal items were less difficult than inferential items, and inferential items were easier than evaluative items. The relative difficulty of each of these item types provides preliminary empirical support for different levels of reading comprehension. Also worth noting is that the range of item difficulties across item types is the most narrow on the fall measure, with difficulties ranging from approximately -0.96 to -0.36 and is the largest in the spring, with difficulties ranging from -1.14 to -0.14.

Examining differences across levels of comprehension

To answer our first research question regarding significant variability in item difficulties for items written to assess students' literal, inferential, and evaluative comprehension abilities we first conducted a series of one-way ANOVAs with the Rasch item difficulties to test for statistically significant differences in the mean difficulties of the three item types. These results are reported in Table 3. No statistically significant differences in the mean item difficulties for literal, inferential, and evaluative items were observed in the fall for either the independent sample, $F(2,11.03) = 1.40$, $p > .05$, or the replication sample, $F(2, 11.11) = 1.53$, $p. > .05$. In contrast, statistically significant differences in mean item difficulties were obtained for the winter, $F(2, 10.81) = 4.12, p < .05$ and $F(2,10.90) = 4.23, p < .05$, and spring, $F(2, 8.90) = 6.46$, $p < .05$ and $F(2, 9.26) = 6.15$, $p < .05$ benchmark probes for the independent and replication samples, respectively. Games–Howell post hoc comparisons revealed statistically significant differences in the mean difficulties of the literal and evaluative items for the independent and replication samples in both the winter and the spring, but no significant differences between literal and inferential or inferential and evaluative items. Although statistically significant differences were not observed between the three item types, the significant differences observed between literal and evaluative items indicate that all items were not equally difficult for all students.

Because of the small number of items and because some of the data failed to meet the assumption of homogeneity of variance, we also used non-parametric tests to examine the mean and median ranks of item difficulties for significant differences

Table 2 Descriptive statistics for literal ($N = 7$), inferential ($N = 7$), and evaluative ($N = 6$) comprehension item difficulties in the fall, winter, and spring of fifth grade

Probe and item type	Independent sample				Replication sample			
	M	SD	Minimum	Maximum	*M*	SD	Minimum	Maximum
Fall								
Literal	−0.96	0.72	−1.82	0.40	−0.91	0.66	−1.68	0.35
Inferential	−0.43	0.54	−0.93	0.61	−0.40	0.54	−0.84	0.64
Evaluative	−0.39	0.60	−1.26	0.15	−0.36	0.58	−1.23	0.16
Winter								
Literal	−1.06	0.32	−1.34	−0.45	−1.02	0.30	−1.35	−0.46
Inferential	−0.81	0.26	−1.1	−0.56	−0.76	0.27	−1.16	−0.50
Evaluative	−0.52	0.33	−0.95	−0.04	−0.48	0.34	−0.91	−0.04
Spring								
Literal	−1.14	0.28	−1.54	−0.73	−1.09	0.30	−1.58	−0.69
Inferential	−0.66	0.85	−1.39	0.77	−0.64	0.83	−1.29	0.79
Evaluative	−0.14	0.63	−0.65	−0.97	−0.15	0.60	−0.65	0.90

by question type (literal, inferential, and evaluative) and by time of year (fall, winter, and spring). With the Kruskal–Wallis test of the means, the 20-multiple choice questions for each probe were rank ordered with respect to their item difficulties, from lowest to highest, to see if there were significant differences in the mean ranks of the three question types. The results of these tests, reported in Table 4, indicate statistically significant differences in the rank-ordering of literal, inferential, and evaluative items for the winter (independent sample) and the spring (independent and replication samples) of fifth grade, but again no differences for the fall measure. Follow-up pairwise comparisons of the rank ordering of the question types by their means revealed statistically significant differences of literal and evaluative questions for the winter χ^2 (1) = 5.22, $p < 0.05$ and χ^2 (1) = 4.59, $p < 0.05$, and for the spring, χ^2 (1) = 9.00, $p < 0.01$ and χ^2 (1) = 9.00, $p < 0.01$ for the independent and replication samples, respectively. Although statistically significant differences were not evident in pair-wise comparisons of the ranks for literal and inferential item types or inferential and evaluative item types, descriptive comparisons of the mean ranks revealed that the mean ranks of literal items were, on average, lower than the mean ranks for inferential items and the mean ranks inferential items were, on average, lower than the mean ranks for evaluative items. These findings are consistent with the claims of prior research (Alonzo et al. 2009) that literal items are, on average, less challenging than inferential items and that inferential items are, on average, less challenging than evalusative items. Similar trends in the relations across the item types were also observed in the Kruskal–Wallis omnibus tests and follow-up comparisons of the median ranks, with the only difference being statistically significant differences in literal and evaluative questions in the winter for *both* samples; for the sake of parsimony we present here only the results of the tests of the mean ranks, which are less sensitive to distributional issues.

Table 3 One-way analyses of variance for average Rasch item difficulties of literal, inferential, and evaluative comprehension items

Time of year	Sample	Literal items	Inferential items	Evaluative items	*F*	*df*
Fall	Independent	−0.95	−0.43	−0.39	1.73	2
	Replication	−0.91	−0.40	−0.37	1.80	2
Winter	Independent	−1.06	−0.81	−0.52	5.11*	2
	Replication	−1.02	−0.76	−0.48	5.04*	2
Spring	Independent	−1.14	−0.66	−0.14	4.03*	2
	Replication	−1.09	−0.64	−0.15	3.80*	2

* $p < .05$

One potential explanation for the lack of significance in the fall is that all of the items, on average, appear to be more difficult than they are in the winter and spring, as evidenced by the larger (more negative) values for the mean item difficulties. The mean difficulties for the literal items in the fall, for example, are anywhere from 0.07 to 0.23 higher on the logit scale than the mean difficulties for literal items in the winter or spring. Similarly, the mean difficulties for inferential items are anywhere from 0.21 to 0.41 higher on the logit scale for inferential items in the winter or spring, while evaluative items appear to be easiest in the winter but still approximately 0.20 higher on the logit scale in the fall compared to the spring. These observed differences in average difficulty are not negligible and may provide some explanation as to why significant differences in the question types were not observed in the fall. Overall, these findings are consistent with claims of prior research (Alonzo et al., 2009) and indicate that literal items were less challenging than inferential items, and that inferential items were less challenging than evaluative items.

Exploring the structure of the multiple-choice comprehension measure

To answer our second question regarding whether a three-factor model fit the data we conducted a bifactor model and a second-order (or confirmatory factor) model, each with three factors to represent the levels of comprehension the multiple-choice reading comprehensions were intended to assess: literal, inferential, and evaluative. The important difference between these two models is that the bifactor model, as described earlier, allows for conditional dependence among identified groups of items and is appropriate when item are hypothesized to have a two-level structure with a general, underlying latent factor that is represented by all items (Chen, West, & Sousa, 2006; Gibbons et al., 2006). In this study, use of the bifactor model allowed us to examine whether the three domain-specific factors explained any additional variance above and beyond the general reading comprehension factor, and to examine the magnitude of the relation between the domain-specific factors (i.e., literal, inferential, and evaluative comprehension) and their associated items. We were also able to compare the results from the multi-dimensional bifactor model to a uni-dimensional, single-order factor model with reading comprehension as the general factor *and* the second-order three-factor model with literal, inferential, and evaluative factors that did not include a general reading comprehension factor.

Table 4 Results from Kruskal–Wallis tests comparing the equality of population mean ranks across literal, inferential, and evaluative comprehension items

Time of year	Sample	Literal comprehension	Inferential comprehension	Evaluative comprehension	χ^2	*df*
Fall	Independent	7.29	12.00	12.50	3.20	2
	Replication	7.14	12.29	12.33	3.47	2
Winter	Independent	6.57	10.43	15.17	6.82*	2
	Replication	6.64	10.93	14.50	5.76	2
Spring	Independent	6.57	9.86	15.83	8.05*	2
	Replication	6.86	9.57	15.83	7.70*	2

* $p < .05$

We first compared the results of the uni-dimensional, single factor model to the multi-dimensional bifactor model based on the estimated log likelihood used to compute Akaike Information Criterion (AIC) statistics (Bozdogan, 2000). The AIC is an information statistic that is useful for judging the relative benefit of alternative models, where smaller AIC values are associated with the preferred model. The AIC values for the single factor and bifactor models for the Fall, Winter, and Spring were (1,697.78, 1,677.78), (1,560.79, 1,560.71), and (1,468.81, 1,529.81), respectively. These values suggest that the bifactor model was supported by the fall and winter data while the single factor model was supported by the spring data.

The fit indices obtained for comparisons of the three-factor bifactor and second-order factor models are reported in Table 5. Examination of these model comparisons reveals that no one model consistently fits the data better than the others; the AIC and BIC values, for example, indicate the bifactor model is a better fit for the independent sample in the fall and for both the independent and replication samples in the spring while the second order model is a better fit for the replication sample in the fall and both the independent and replication samples in the winter. It could be argued, for example, that the second-order model better represents the relationships among the comprehension factors because the literal, inferential, and evaluative comprehension factors were strongly correlated with one another (correlations ranged from 0.85 to 0.99) and a higher order factor, in this case general reading comprehension, could be hypothesized to account for the strong relationships observed among the domain specific factors (Chen et al., 2006). The model fit indices, however, indicate that this may not always be the case.

We chose to interpret the results of the bifactor model for four reasons. First, we were interested in examining the role of the domain specific factors—literal, inferential, and evaluative comprehension—independent of the general reading comprehension factor. The bifactor model allows this by providing information about the percent of variance explained by each of the domain-specific factors above and beyond the variance explained by the general factor. Second, because we were interested in examining the underlying structure of this multiple-choice reading comprehension measure we wanted to examine the relation between the domain-specific factors and their associated items. Results from the bifactor models

Table 5 Model fit comparisons for the three-factor bifactor and second-order factor models

Time of year	Sample	Bifactor model		Second-order factor	
		AIC	BIC	AIC	BIC
Fall	Independent	10,541.19	10,604.80	11,178.35	11,369.11
	Replication	11,451.87	11,516.67	10,166.19	10,356.32
Winter	Independent	9,865.09	9,930.32	9,740.26	9,929.83
	Replication	11,134.06	11,200.93	9,929.61	10,119.25
Spring	Independent	9,144.19	9,941.58	17,916.98	18,121.98
	Replication	9,210.30	10,008.86	19,161.49	19,376.42

for the three probes (fall, winter, and spring) and two samples (independent and replication) revealed low to moderate correlations between the domain-specific factors and their items, ranging from −0.27 to 0.54. Third, the empirical reliability estimates obtained from the bifactor models revealed moderate correlations (ranging from 0.70 to 0.75) between the observed true scores and the estimated latent scores, indicating that the models fit the data reasonably well. Finally, previous research indicates that the fit of the second-order structure can be statistically tested only when there are four or more first-order (or domain-specific) factors hypothesized (Chen et al., 2006); this was not the case in this study.

Examining variance explained by the levels of comprehension

To answer our third question regarding the amount of additional variance in students' reading comprehension scores explained by literal, inferential, and evaluative multiple-choice questions we employed a bifactor model. This model was used for the following reasons: (a) we hypothesized that a general factor—general reading comprehension—would account for commonality across the items that compose the fall, winter, and spring measures, (b) based on the results of our Confirmatory Factor Analyses, we hypothesized that three domain specific factors—literal, inferential, and evaluative comprehension—would account for unique variance above and beyond the general factor, and (c) we were interested in the domain-specific factors (i.e., levels of comprehension) in addition to the general factor of interest, reading comprehension (Chen, West, & Sousa, 2006).

Results of these model comparisons examining the amount of variance explained by each of the aforementioned factors for each multiple-choice reading comprehension benchmark probe (i.e., fall, winter, and spring) for two samples of students (the independent sample and replication sample) are reported in Table 6. For each assessment and sample the following information is reported: the percent of variance explained by general reading comprehension factor in the single factor model, the percent of variance explained by the general reading comprehension, literal inferential, and evaluative comprehension factors in the bifactor model, the variance explained by the domain-specific factors (i.e., literal inferential, and evaluative comprehension) in the bifactor model, the total percent of variance

Table 6 Percent of variance in fifth graders' performance on multiple-choice reading comprehension measures in the fall, winter, and spring explained by single factor and bifactor models

Sample	Single factor	Bifactor model						Model comparisons
	Percent of variance explained (%)	General reading comprehension (%)	Literal comprehension (%)	Inferential comprehension (%)	Evaluative comprehension (%)	Variance explained by specific factors (%)	Total percent variance explained (%)	Difference in percent variance explained (%)
Fall								
Independent	30.08	29.67	1.85	2.26	1.43	5.54	35.21	+5.13
Replication	29.63	29.46	1.59	2.45	2.14	6.18	35.64	+6.01
Winter								
Independent	32.15	32.49	1.82	1.09	1.12	4.03	36.52	+4.37
Replication	31.79	31.62	2.71	1.34	1.47	5.52	37.14	+5.35
Spring								
Independent	31.65	31.15	1.93	1.81	1.05	4.79	35.42	+3.48
Replication	32.33	31.84	1.79	0.37	1.70	3.88	35.70	+3.37

explained by the bifactor model, and the difference in the percent of variance explained by the bifactor model compared to the single factor model.

Examination of these data revealed that for the fall, winter, and spring benchmark probes questions written purposefully to assess students' literal, inferential, and evaluative comprehension abilities explained additional variance observed in students' performance above and beyond a general reading comprehension factor. Results indicated that, for the independent sample, these two domain-specific factors explained anywhere from 4.79 to 5.54 % of the variance above and beyond the general reading comprehension factor and, when compared to the single factor model, explained anywhere from 3.48 to 5.13 % *more* of the unique variance observed in students' reading comprehension scores. Similar results were obtained with the replication sample. In particular, the three domain-specific factors explained anywhere from 3.88 to 6.12 % of the variance above and beyond the general reading comprehension factor and, when compared to the single factor model, explained anywhere from 3.37 to 6.01 % more of the unique variance observed in students' reading comprehension scores. Moreover, the amount of variance explained by the domain-specific factors ranged from 1.59 to 2.71 % for literal comprehension items, 0.37–2.45 % for inferential comprehension items, and 1.05–2.14 % for evaluative comprehension items.

Discussion

The purpose of this study was twofold: (a) to determine whether there were statistically significant differences in the Rasch item difficulties estimated for literal, inferential, and evaluative comprehension items, and (b) to examine the structure of reading comprehension as represented by items written specifically to assess students' literal, inferential, and evaluative comprehension abilities. These questions were examined using data collected from two independent samples of students across the fall, winter, and spring of fifth grade. Our findings provide empirical support for the levels of comprehension theory in two distinct ways. First, results of the independent samples *t* tests and Mann–Whitney *U* tests comparing the mean ranks of the item types revealed, for the most part, statistically significant differences in the two groups of items, particularly in the winter and spring of fifth grade. The differences observed indicated that inferential and evaluative items were more challenging for students than literal items, findings that are consistent with previous research (Alonzo et al., 2009; McCormick, 1992; Snider, 1988). Second, results of the bifactor model indicated that the three domain-specific factors—literal, inferential, and evaluative comprehension—accounted for anywhere from 3.15 to 4.74 % of the variance observed in students' scores on multiple-choice comprehension measures over and above a general reading comprehension factor.

Evidence supporting a non-linear relation among the levels of comprehension

Also consistent with previous research was the finding that the relationship between the three levels of comprehension was not linear (Alonzo et al., 2009; Herber,

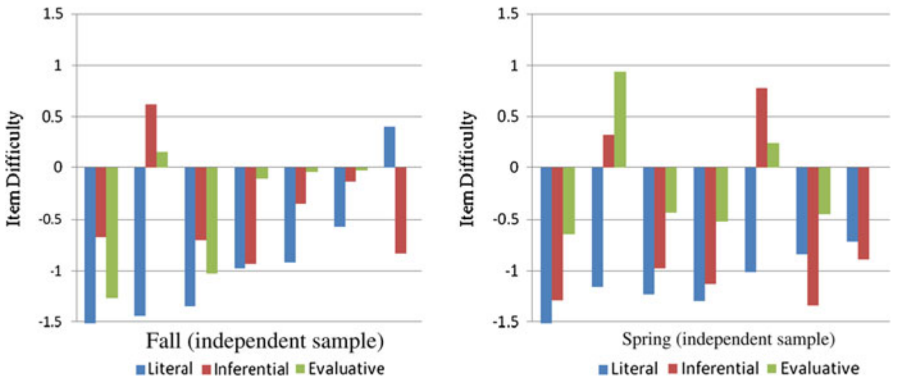

Fig. 1 Comparison of mean Rasch item difficulties for literal, inferential, and evaluative items

1970). This finding is supported by examination of the Rasch item difficulties that overlap in the fall, winter, and spring for samples, meaning that, in some cases, inferential items were easier for students than literal or evaluative items, or that even evaluative items were easier than literal items. Examination of the descriptive statistics reported in Table 1, for example, reveals that for the independent sample in the winter the item difficulties of literal items ranged from −1.34 to −0.45, inferential items ranged from −1.1 to −0.56, and evaluative items ranged from −0.95 to −0.04; some inferential and evaluative items, in other words, had item difficulties comparable to literal items (and vice versa). The relations among the mean item difficulties for these three item types originally proposed are depicted in Fig. 1 to demonstrate the overlap between the three levels of comprehension.

These findings indicate that although literal items, for example, are *on average* easier than inferential and evaluative items, they can also be more challenging than inferential or evaluative items. In addition, the wide range of difficulties observed for each item type shows that items can be easy and challenging *within* each level of comprehension (e.g.., items requiring elaborative inferences may be more challenging than those requiring text-based inferences; McNamara & Magliano, 2009). These findings, then, run counter to the primary criticism of the levels of comprehension theory, which argues that within this framework one level of comprehension is a prerequisite for the subsequent one; that is, literal comprehension must be mastered to acquire inferential comprehension, and inferential comprehension must be mastered to acquire evaluative comprehension (Lapp & Flood, 1983; Vacca et al., 2008). Instead, the overlapping item difficulties support the notion that, although conceptually literal comprehension is required for inferential and evaluative comprehension, some inferential and evaluative questions may be easier for students to answer than literal comprehension items and therefore the relationship among the levels is not hierarchical. As previously argued by Alonzo et al. (2009) these findings provide empirical support for the levels of comprehension theory espoused in reading instruction texts (Carnine et al., 2010; Vacca et al., 2008) via the estimation of the differences in difficulty of the levels of comprehension.

Potential explanations for lack of differences observed in the fall

One finding that warrants further exploration is the lack of statistically significant differences in the mean item difficulties and in the mean ranks of the three groups of items in the fall of fifth grade for both samples. The independent samples *t* tests comparing the mean item difficulties for literal, inferential, and evaluative items were not significant for significant for the independent sample, $t(18) = -1.91$, $p > 0.05$, or the replication sample, $t(18) = -1.95$, $p > 0.05$. Moreover, examination of the mean ranks of the items reveals that they were higher for the fall than the spring while the mean ranks of the inferential and evaluative items were relatively consistent, indicating that the range of scores in the fall was much narrower compared to that observed in the winter or spring.

There may be two potential explanations for these findings. First, examination of the descriptive statistics of mean item difficulties for each item type presented in Table 1 reveals that the average literal comprehension item difficulty in the fall was approximately one-tenth larger in value on the logit scale compared to average literal comprehension item difficulties in the winter and spring. Average difficulties for literal items in the fall, for example, were −0.95 and −0.91 compared to average difficulties ranging from −1.02 to −1.14 in the winter and spring, differences that are small in absolute magnitude but are relatively large on the logit scale. Similarly, examination of the descriptive statistics reveals that inferential items, on average, were more challenging in the fall than in the winter or spring. The average difficulty of inferential items in the fall, for example, ranged from −0.43 to −0.40 (for the independent and replication samples, respectively) compared to ranges of −0.76––0.81 in the winter and −0.64––0.66 in the spring. Just as literal items appeared to be more challenging in the fall compared to winter and spring, so too did inferential items appear to be more challenging for students (a difference of approximately three-tenths on the logit scale). The average item difficulties for literal, inferential, and evaluative comprehension are also closer in value in the fall for both samples compared to the winter and spring, meaning that less variability was observed in student performance on the items, thus making it more difficult for statistically significant differences to be observed between the three item types. Similar trends are observed in item difficulties when examining the mean item difficulties for the literal, inferential, and evaluative item types produced by the independent samples *t* tests. Each of these results provides support for further investigation of possible form effects as these relationships among item types were only observed in the fall of fifth grade.

A second potential explanation may be related less to the structure of the measures than to the instruction at different points during the school year. A review of the literature, for example, supports the possibility of a summer effect, or a decrease in students' scores from the spring of the previous school year that may be a byproduct of the absence of instruction during the summer months (Burkam, Ready, Lee, & LoGerfo, 2004; Lawrence, 2011; McCoach, O'Connell, Reis, & Levitt, 2006; Zvoch, 2009) as an explanation for dramatic differences observed in the fall compared to later in the school year. It is worth noting that this negative effect is particularly evident for inferential and evaluative items whose difficulties are anywhere from 0.38 (inferential) to 0.25 (evaluative) larger in the fall compared to the winter or spring, meaning that the summer effect may have a negative effect on all question types, but particularly on inferential and evaluative questions. From a policy perspective, these data speak to a need for continued support of student's comprehension during the summer, perhaps via summer school programs or structured comprehension activities that students learn throughout the school year to support deeper-thinking about texts that they can then implement independently or with an adult during the summer. In addition, these data may also support the implementation of explicit, teacher-directed comprehension instruction focused specifically on student engagement with the text from the very beginning of the school year in an effort to minimize the potentially long-lasting effects of the summer effect. Although no claims about a true summer effect can be made based on these data because no scores from the spring of students' fourth grade year were included in the analyses, it may be possible that students' performed equally poorly on literal, inferential, and evaluative comprehension in the fall of fifth grade after several months without explicit, teacher-directed instruction. This interpretation is supported by the fact that average item difficulties in the fall are greater (i.e., less negative) than those for all item types in the winter and for literal and inferential items in the spring. This explanation must be interpreted with caution, however, and warrants further exploration.

Limitations

The generalizability of these findings is limited by the fact that all data from this study were collected using one specific multiple-choice reading comprehension measure with one specific population of students (i.e., fifth graders). As discussed earlier, the findings obtained in this study also introduce the possibility of a form effect, meaning that further investigation of data from the fall benchmark probe is warranted to see if similar performance patterns among literal, inferential, and evaluative items are observed with other samples of students. It is also possible that different results may have been obtained if different test formats, such as cloze, maze, or think-aloud procedures had been used to examine students' literal, inferential, and evaluative comprehension abilities; recent research has advocated for the use of multiple test formats that require different levels of interaction to most appropriately assess this cognitively complex construct (Campbell, 2005; Sweet, 2005).

A further limitation may involve the definitions of literal, inferential, and evaluative comprehension used during the construction of these measures. Although a review of the research reveals a general consensus on the definitions and importance of literal and inferential comprehension to overall understanding, there is less agreement among researchers and educators about the idea of evaluative comprehension and whether it differs at all from higher-level inferencing. Lastly, the use of readability formulae in general and only one readability formula in particular for the leveling of the comprehension measures may also be problematic as different readability formulae rely on different indices of text complexity to estimate text difficulty (Bailin & Grafstein, 2001; Bruce & Rubin, 1988). A review of recent research (Ardoin, Williams, Christ, Klubnik, & Wellborn, 2010; Christ & Ardoin, 2009; Francis et al., 2008) suggests that other methods, such as the use of Lexiles in conjunction with an equating procedure, may be more appropriate for ensuring that the passages used to measure reading comprehension are of comparable difficulty.

Conclusions

The multiple-choice reading comprehension measures within the easyCBM assessment system are designed as general outcome measures, intended to guide instructional decision making by providing teachers with periodic snapshots of student performance that facilitate the monitoring of student progress (Deno, 1985). As noted earlier, most early literacy research, including the development of CBMs, has focused on brief, individually administered measures of skills that can be firmly mastered, such as phonological awareness, alphabetic understanding, and oral reading fluency, which tend to lose their usefulness as an instructional planning tool for most students beyond third grade (Yovanoff, Duesbery, Alonzo, & Tindal, 2005). Most CBMs of reading comprehension developed thus far have traditionally relied on maze or cloze procedures whereby reading comprehension is assessed via students' ability to correctly supply or identify words missing from a passage (Good & Kaminski, 2011; Shinn & Shinn, 2002). In contrast, the measure employed in this study was a longer, group-administered, multiple-choice measure that was constructed using complex statistical analyses that can provide additional insights about student performance beyond an overall total score. One potential direction for further research using these same analytic techniques (i.e., the bifactor model) would be to test whether the domain-specific factors predict external variables, such as student performance on a reading comprehension subtest of a large-scale assessment, over and above the general reading comprehension factor (Chen et al., 2006); given the intended purposes of these multiple choice reading comprehension measures, one of which is to determine whether a student is on track to meet later reading goals, this possibility seems relevant.

Because CBMs are intended to inform instructional decision-making, one significant advantage of the reading comprehension measure employed in this study is the ability to obtain information about students' abilities to respond correctly to literal, inferential, and evaluative comprehension questions. Just as careful examination of student performance on a measure of alphabetic understanding, such as whether students are reading pseudo-words sound-by-sound or as whole words can help teachers target instruction (e.g., providing scaffolded blending instruction for sound-by-sound readers versus increased blending practice to build fluency for whole word readers; Basaraba, Travers, & Chaparro, 2011; Harn, Stoolmiller, & Chard, 2008), so too can the results obtained from the easyCBM reading comprehension assessment be used to target instruction. If examination of item-level performance reveals, for example, that students' have mastered literal comprehension but are struggling with inferential and evaluative questions teachers can plan and design lessons that target these higher-level comprehension skills.

Although not intended for diagnostic purposes, teachers can nonetheless use this level of detailed information to inform, guide, and plan reading comprehension instruction.

Acknowledgments This study was partially funded by a grant from the Office of Special Education Programs: Model Demonstration Centers on Progress Monitoring (Award H326M050003). The findings and conclusions, however, are those of the authors only and do not necessarily represent the perspectives of the U.S. Department of Education or the Office of Special Education Programs.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: Massachusetts Institute of Technology.

Allington, R. L., Chodos, L., Domaracki, J., & Trueax, S. (1977). Passage dependency: Four diagnostic oral reading tests. *The Reading Teacher, 30*, 369–375.

Alonzo, J., Basaraba, D., Tindal, G., & Carriveau, R. (2009). They read, but how well do they understand? An empirical look at the nuances of measuring reading comprehension. *Assessment for Effective Intervention, 35*, 34–44.

Alonzo, J., Liu, K., & Tindal, G. (2007). *Examining the technical adequacy of reading comprehension measures in a progress monitoring assessment system*. Technical report #41. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2008). *Examining the technical adequacy of fifth-grade reading comprehension measures in a progress monitoring assessment system*. Technical report no. 0808. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., & Tindal, G. (2009). *Alternate form and test–retest reliability of easyCBM reading measures*. Technical report no. 0906. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). *easyCBM online progress monitoring assessment system*. Eugene, OR: Center for Educational Assessment Accountability. Available at http://easycbm.com.

Anderson, R. C. (2004). Role of the reader's schema in comprehension, learning, and memory. In R. B. Ruddell & N. J. Unrau (Eds.), *Theoretical models and process of reading* (5th ed., pp. 594–619). Newark, DE: International Reading Association.

Anderson, D., Alonzo, J., & Tindal, G. (2011a). *easyCBM reading criterion related validity evidence: Oregon state test 2009–2010*. Technical report no. 1103. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2011b). *easyCBM reading criterion related validity evidence: Washington state test 2009–2010*. Technical report no. 1101. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Applegate, M. D., Quinn, K. B., & Applegate, A. J. (2002). Levels of thinking required by comprehension questions in informal inventories. *The Reading Teacher, 56*, 174–180.

Ardoin, S. P., Williams, J. C., Christ, T. J., Klubnik, C., & Wellborn, C. (2010). Examining readability estimates' predictions of students' oral reading rate: Spache, Lexile, and Forcast. *School Psychology Review, 39*, 277–285.

Bailin, A., & Grafstein, A. (2001). The linguistic assumption underlying readability formulae: A critique. *Language & Communication, 21*, 285–301.

Baker, S. K., Smolkowski, K., Katz, R., Fien, H., Seeley, J. R., Kame'enui, E. J., et al. (2008). Reading fluency as a predictor of reading proficiency in low-performing, high-poverty schools. *School Psychology Review, 37*, 18–37.

Basaraba, D., Travers, P., & Chaparro, E. (February, 2011). *Application of Ehri's theory: Instructional implications of students' decoding skills*. Paper presented at the National Association of School Psychology annual conference, San Diego, CA.

Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology, 29*, 137–164.

Boulineau, T., Fore, C., Hagan-Burke, S., & Burke, M. D. (2004). Use of story-mapping to increase the story-grammar text comprehension of elementary students with learning disabilities. *Learning Disability Quarterly, 27*, 105–121.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology, 44*, 62–91.

Bruce, B., & Rubin, A. (1988). Readability formulas: Matching tool and task. In A. Davison & G. M. Green (Eds.), *Linguistic complexity and text comprehension: Readability issues reconsidered* (pp. 5–22). Hillsdale, NJ: Lawrence Erlbaum.

Burkam, D. T., Ready, D. D., Lee, V. E., & LoGerfo, L. F. (2004). Social-class differences in summer learning between kindergarten and first grade: Model specification and estimation. *Sociology of Education, 77*, 1–31.

Bus, A. G., & van Ijzendoorn, M. H. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology, 91*, 403–414.

Cain, K., Oakhill, J., Barnes, M. A., & Bryant, P. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & Cognition, 29*, 850–859.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96*, 31–42.

Campbell, J. R. (2005). Single instrument, multiple measures: Considering the use of multiple item formats to assess reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 347–368). Mahwah, NJ: Lawrence Erlbaum.

Carnine, D. W., Silbert, J., Kame'enui, E. J., & Tarver, S. G. (2010). *Direct instruction reading* (5th ed.). Boston, MA: Merrill.

Chard, D. J., Simmons, D. C., & Kame'enui, E. J. (1998). Word recognition: Research bases. In D. C. Simmons & E. J. Kame'enui (Eds.), *What reading research tells us about children with diverse learning needs: Bases and basics* (pp. 141–168). Mahwah, NJ: Lawrence Erlbaum.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*, 189–225.

Chikalanga, I. (1992). A suggested taxonomy of inferences for the reading teacher. *Reading in a Foreign Language, 8*, 697–709.

Christ, T. J., & Ardoin, S. P. (2009). Curriculum-based measurement of oral reading: Passage equivalence and probe-set development. *Journal of School Psychology, 47*, 55–75.

Coleman, C., Lindstrom, J., Nelson, J., Lindstrom, W., & Gregg, K. N. (2010). Passageless comprehension on the *Nelson-Denny Reading Test*: Well above chance for university students. *Journal of Learning Disabilities, 43*, 244–249.

Davey, B. (1988). Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers. *Journal of Experimental Education, 56*, 67–75.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219–232.

Dimino, J., Gersten, R., Carnine, D., & Blake, G. (1990). Story grammar: An approach for promoting at-risk secondary students' comprehension of literature. *The Elementary School Journal, 91*, 19–32.

Dole, J. A., Duffy, G. G., Roehler, L. R., & Pearson, P. D. (1991). Moving from the old to the new: Research on reading comprehension instruction. *Review of Educational Research, 61*, 239–264.

Duran, N. D., McCarthy, P. M., Graesser, A. C., & McNamara, D. S. (2007). Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods, 39*, 212–223.

Durkin, D. (1989). *Teaching them to read* (5th ed.). Boston, MA: Allyn & Bacon.

Fien, H., Baker, S. K., Smolkowski, K., Mercier Smith, J. L., Kame'enui, E. J., & Thomas Beck, C. (2008). Using nonsense word fluency to predict reading proficiency in kindergarten through second grade for English learners and native English speakers. *School Psychology Review, 37*, 391–408.

Fitzgerald, T. P., & Fitzgerald, E. F. (1978). A cross-cultural study of three measures of comprehension at the primary and intermediate levels. *Educational Research Quarterly, 3*, 84–89.

Fox, E. (2002). The role of reader characteristics in processing and learning from informational text. *Review of Educational Research, 79*, 197–261.

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*, 315–342.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's mini talks. *Language Testing, 16*(2), 2–32.

Gernsbacher, M. A., Robertson, R. R., Palladino, P., & Werner, N. K. (2004). Managing mental representations during narrative comprehension. *Discourse Processes, 5*, 53–72.

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. K. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of the research. *Review of Educational Research, 71*, 279–320.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., et al. (2006). Full information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*, 4–19.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bifactor analysis. *Psychometrika, 57*, 423–436.

Good, R. H., Baker, S. K., & Peyton, J. A. (2009). Making sense of nonsense word fluency: Determining adequate progress in early first-grade reading. *Reading and Writing Quarterly, 25*, 33–56. doi: 10.1080/1058/3560802491224.

Good, R. H., & Kaminski, R. A. (2011). *DIBELS next assessment manual*. Eugene, OR: Dynamic Measurement Group.

Good, R., Simmons, D., & Kame'enui, E. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Reading Studies, 5*, 257–288.

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

Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology, 93*, 103–128.

Harn, B. A., Stoolmiller, M., & Chard, D. J. (2008). Measuring the dimensions of alphabetic principle on the reading development of first graders: The role of automaticity and unitization. *Journal of Learning Disabilities, 41*, 143–157.

Herber, H. L. (1970). *Teaching reading in the content areas*. Englewood Cliffs, NJ: Prentice Hall.

Kendall, J. R., Mason, J. M., & Hunter, W. (1980). Which comprehension? Artifacts in the measurement of reading comprehension. *The Journal of Educational Research, 73*, 233–236.

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

Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 71–92). New York, NY: Routledge.

Kintsch, W., & Rawson, K. A. (2005). Comprehension. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 209–226). Malden, MA: Blackwell.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293–323.

Lapp, D., & Flood, J. (1983). *Teaching reading to every child*. New York: MacMillan.

Lawrence, J. F. (2011). English vocabulary trajectories of students whose parents speak a language other than English: Steep trajectories and sharp summer setback. *Reading and Writing: An Interdisciplinary Journal*. doi:10.1007/s11145-011-9305-z.

Leu, D. J., & Kinzer, C. K. (1999). *Effective literacy instruction (K-8)* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Masson, M. E., & Miller, J. A. (1983). Working memory and individual differences in comprehension and memory of text. *Journal of Educational Psychology, 75*, 314–318.

McCoach, D. B., O'Connell, A. A., Reis, S. M., & Levitt, H. A. (2006). Growing readers: A hierarchical linear model on children's reading growth during the first 2 years of school. *Journal of Educational Psychology, 98*, 14–28.

McCormick, S. (1992). Disabled readers' erroneous responses to inferential comprehension questions: Description and analyses. *Reading Research Quarterly, 27*(1), 54–77.

McNamara, D. S. (1997). Comprehension skill: A knowledge-based account. In *Proceedings of the 19th Annual Meeting of the Cognitive Science Society* (pp. 508–513). Hillsdale, NJ: Erlbaum.

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of reading comprehension. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 51, pp. 297–384). New York, NY: Elsevier Science.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus: Statistical Analysis with latent variables (version 5.21)*. Los Angeles, CA: Muthén & Muthén.

Nation, K. (2005). Children's reading comprehension difficulties. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 248–265). Oxford, MA: Blackwell.

National Institute of Child Health and Human Development (NICHD). (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of scientific research literature on reading and its implications for instruction: Reports of the subgroups*. NIH publication no. 00-4754. Washington, DC: U.S. Government Printing Office. Also available on-line: http://www.nichd.nih.gov/publications/nrp/report.htm.

National Research Council. (1998). Organizational strategies for kindergarten and the primary grades. In C. E. Snow, M. S. Burns, & P. Griffin (Eds.), *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

No Child Left Behind Act of 2001: Reading first, 20 U.S.C. § 6319 (2008).

Oakhill, J., & Cain, K. (2007). Issues of causality in children's reading comprehension. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 47–72). Mahwah, NJ: Lawrence Erlbaum.

Oakhill, J., Hartt, J., & Samols, D. (2005). Levels of comprehension monitoring and working memory in good and poor comprehenders. *Reading and Writing, 18*, 657–686.

Otero, J., Léon, J. A., & Graesser, A. C. (Eds.). (2002). *The psychology of science text comprehension*. Mahwah, NJ: Lawrence Erlbaum.

Pearson, P. D., & Fielding, L. (1991). Comprehension instruction. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 815–860). White Plains, NY: Longman.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices—Past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13–70). Mahwah, NJ: Lawrence Erlbaum.

Pearson, P. D., & Johnson, D. D. (1978). *Teaching reading comprehension*. New York, NY: Holt, Rinehart, & Winston.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.

Perfetti, C. A. (1999). Comprehending written language: A blueprint of the reader. In C. M. Brown & P. Hagoart (Eds.), *The neurocognition of language* (pp. 167–208). New York, NY: Oxford University Press.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Malden, MA: Blackwell.

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest, 2*, 31–74.

Rupley, W. H., & Blair, T. R. (1983). *Reading diagnosis and remediation: Classroom and clinic* (2nd ed.). Boston: Houghton Mifflin.

Rupley, W. H., & Willson, V. L. (1996). Content, domain, and word knowledge: Relationship to comprehension of narrative and expository text. *Reading and Writing: An Interdisciplinary Journal, 8*, 419–432.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*(4), 441–474.

Saéz, L., Park, B., Nese, J. F., Jamgochian, E., Lai, C-F., Anderson, D., et al. (2011). *Technical adequacy of the easyCBM reading measures (grades 3–7), 2009–2010 version*. Technical report #1005. Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Sarroub, L., & Pearson, P. D. (1998). Two steps forward, three steps back: The stormy history of reading comprehension assessment. *The Clearinghouse, 72*, 97–105.

Shinn, M. R., & Shinn, M. M (2002). AIMSweb training workbook: Administration and scoring of reading MAZE for use in general outcomes measurement. Retrieved August 25, 2011 from http://www.aimsweb.com/uploads/pdfs/scoring_maze.pdf.

Short, E. J., & Ryan, E. B. (1984). Metacognitive differences between skilled and less skilled readers: Remediating deficits through story grammar and attribution training. *Journal of Educational Psychology, 76*, 225–235.

Simms, L. J., Grös, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and Anxiety, 25*, E34–E46.

Smith, S. B., Simmons, D. C., & Kame'enui, E. J. (1998). Phonological awareness: Research bases. In D. C. Simmons & E. J. Kame'enui (Eds.), *What reading research tells us about children with diverse learning needs: Bases and basics* (pp. 61–128). Mahwah, NJ: Lawrence Erlbaum.

Snider, V. E. (Fall 1988). The role of prior knowledge in reading comprehension: A test with LD adolescents. *Direct Instruction News*, 6–11.

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

Stothard, S. E., & Hulme, C. (1996). A comparison of reading comprehension and decoding difficulties in children. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 93–112). Hillsdale, NJ: Lawrence Erlbaum.

Swanson, H. L., & O'Connor, R. E. (2009). The role of working memory and fluency training on reading comprehension in children who are dysfluent readers. *Journal of Learning Disabilities, 42*, 548–575.

Sweet, A. P. (2005). Assessment of reading comprehension: The RAND reading study group vision. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment.* Mahwah, NJ: Lawrence Erlbaum.

Tal, N. F., Siegel, L. S., & Maraun, M. (1994). The role of question type and reading ability in reading comprehension. *Reading and Writing: An Interdisciplinary Journal, 6*, 387–402.

Taylor, B. M., & Beach, R. W. (1984). The effects of text structure on middle-grade students' comprehension and production of expository text. *Reading Research Quarterly, 19*, 134–146.

Taylor, S. E., Frackenpohl, H., & White, C. E. (1989). *EDL core vocabularies in reading, mathematics, science, and social studies*. Orlando, FL: Steck-Vaughn.

Tuinman, J. J. (1973). Determining the passage dependency of comprehension questions in 5 major tests. *Reading Research Quarterly, 9*, 207–223.

Tun, P. A. (1989). Age differences in processing expository and narrative text. *Journal of Gerontology, 44*, 9–15.

Vacca, J., Vacca, R. T., & Gove, M. K. (1987). *Reading and learning to read.* Boston, MA: Little, Brown, & Co.

Vacca, J. L., Vacca, R. T., Gove, M. K., Burkey, L. C., Lenhart, L. A., & McKeon, C. A. (2009). *Reading and learning to read* (7th ed.). Boston, MA: Pearson.

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, P., Lorch, R. P., Linderholm, T., & Gustafson, M. (2001a). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition, 29*, 1081–1087.

van den Broek, P., Tzeng, Y., Risden, K., Trabasso, T., & Basche, P. (2001b). Inferential questioning: Effects of comprehension of narrative texts as a function of grade and timing. *Journal of Educational Psychology, 93*, 521–529.

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

Weaver, C. A., & Kintsch, W. (1991). Expository text. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 230–244). White Plains, NY: Longman.

Wood, R., Wilson, D. T., Gibbons, R., Schilling, S., Muraki, E., & Bock, R. D. (2003). *TESTFACT 4 for windows: Test scoring, item statistics, and item factor analysis (computer software).* Lincolnwood, IL: Scientific Software International, Inc.

Yovanoff, P., Duesbery, L., Alonzo, J., & Tindal, G. (2005). Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. *Educational Measurement: Issues and Practice, 24*, 4–12.

Zvoch, K. (2009). A longitudinal examination of the academic year and summer learning rates of full- and half-day kindergarteners. *Journal of Education for Students Placed at Risk, 14*, 311–333.

Zwaan, R. A., & Brown, C. M. (1996). The influence of language proficiency and comprehension skill on situation-model construction. *Discourse Processes, 21*, 289–327.

Zwaan, R. A., Radvansky, G. A., Hilliard, A. E., & Curiel, J. M. (1998). Constructing multidimensional situation models during reading. *Scientific Studies of Reading, 2*, 199–220.