Development of Comprehension Monitoring in Beginner Readers

Language and Reading Research Consortium

Gloria Yeomans-Maldonado

The Ohio State University

Accepted for publication in Reading and Writing, 16 June 2017

Author Note

This paper was prepared by a Task Force of the Language and Reading Research Consortium (LARRC) consisting of Laura M. Justice (Convener), Kate Cain, and Gloria Yeomans-Maldonado. LARRC project sites and investigators are as follows:

Ohio State University (Columbus, OH): Laura M. Justice (Site PI), Richard Lomax, Ann O'Connell, Jill Pentimonti[1], Stephen A. Petrill[2], Shayne B. Piasta.

Arizona State University (Tempe, AZ): Shelley Gray (Site PI), Maria Adelaida Restrepo.

Lancaster University (Lancaster, UK): Kate Cain (Site PI).

University of Kansas (Lawrence, KS): Hugh Catts[3] (Site PI), Mindy Bridges, Diane Nielsen.

University of Nebraska-Lincoln (Lincoln, NE): Tiffany Hogan[4] (Site PI), Jim Bovaird, J. Ron Nelson.[5]

1. Jill Pentimonti is now at American Institute for Research.

2. Stephen A. Petrill was a LARRC co-investigator from 2010-2013.

3. Hugh Catts is now at Florida State University.

4. Tiffany Hogan is now at MGH Institute of Health Professions.

5. J. Ron Nelson was a LARRC co-investigator from 2010-2012.

This work was supported by grant # R305F100002 of the Institute of Education Sciences' Reading for Understanding Initiative. We are deeply grateful to the numerous staff, research associates, school administrators, teachers, children, and families who participated. Key personnel at study sites include: Lisa Baldwin-Skinner, Lauren Barnes, Garey Berry, Beau Bevens, Jennifer Bostic, Shara Brinkley, Janet Capps, Beth Chandler, Lori Chleborad, Willa Cree, Dawn Davis, Jaclyn Dynia, Michel Eltschinger, Kelly Farquharson, Tamarine Foreman, Rashaun Geter, Sara Gilliam, Cindy Honnens, Miki Herman, Jaime Kubik, Trudy Kuo, Gustavo Lujan, Junko Maekawa, Carol Mesa, Denise Meyer, Maria Moratto, Kimberly Murphy, Marcie Mutters, Amy Pratt, Trevor Rey, Amber Sherman, Shannon Tierney, Stephanie Williams, and Gloria Yeomans-Maldonado.

The views presented in this work do not represent those of the federal government, nor do they endorse any products or findings presented herein. Correspondence concerning this work should be sent to Laura M. Justice (justice.57@osu.edu), The Ohio State University, 356 Arps Hall, Columbus OH 43210.

Abstract

The current study was designed to understand the development of comprehension monitoring among beginner readers from first to third grade, and to determine the extent to which first graders' comprehension monitoring predicts reading comprehension in grade three. Participants were 113 children (57% female) from four US states who were followed from Grade 1 (M = 7 years, SD = 4 months) to Grade 3 (M = 9 years, SD = 4 months). Measures included decoding, vocabulary, working memory, comprehension monitoring, and reading comprehension. Children's ability to monitor comprehension grew significantly from first to third grade, with a deceleration in growth over time. In addition, comprehension monitoring in Grade 1 made a significant contribution to reading comprehension in Grade 3, even after controlling for decoding, vocabulary, and working memory. Together, these findings supplement our understanding of young readers' development of comprehension monitoring as well as its association with reading comprehension at a later time. Practical implications of the results in the context of providing support for higher-level language skills in beginning reading instruction are discussed.

Keywords: Comprehension, Decoding, Oral language, Vocabulary

Development of Comprehension Monitoring in Beginner Readers

The simple view of reading theorizes that reading is the product of both word recognition and language (or listening) comprehension (Hoover & Gough, 1990). There is substantial support for this viewpoint across a range of alphabetic writing systems, which confirms there to be significant and unique relations between both word recognition and language comprehension and children's reading comprehension (e.g., Authors, 2011a; Authors, 2015c; Protopapas, Simos, Sideridis, & Mouzaki, 2012; Tunmer & Chapman, 2012). Early in reading development, word recognition plays a particularly prominent role in reading skill relative to language comprehension; over time, however, as word reading becomes more fluent, language comprehension becomes the prominent source of variance in reading skill (Authors, 2013; Kershaw & Schatschneider, 2012).

Given the importance of language comprehension to general reading ability as children progress as readers, there is increasing interest in learning more about the component skills that contribute substantially to language comprehension, such as comprehension monitoring. Comprehension monitoring refers to the ability to evaluate the adequacy of one's understanding for speech or written text. It is typically assessed with an error detection task, in which the child is presented with materials that include deliberate anomalies or inconsistencies, such as nonwords, violations with prior knowledge, or internal inconsistencies in which two details in the text contradict. The ability to monitor one's comprehension is measured by the reader's (or listener's) judgment that the material does not make sense and typically the participant is also required to identify the error. Thus, it involves deliberate reflection on one's comprehension and, for this reason, comprehension monitoring is considered a metacognitive skill (Wagoner, 1983).

The ability to monitor one's comprehension, assessed by the ability to detect internal inconsistencies, is observed among good readers (Oakhill, Hartt, & Samols, 2005) and explains variance in reading comprehension development between 7 to 11 years of age (Authors, 2004; Authors, 2012a). Investigations of poor comprehenders, who are struggling readers with average or better word recognition yet poor reading comprehension, support the view that comprehension monitoring is a component language skill of significance to reading: children who are poor comprehenders are significantly poorer at monitoring their comprehension compared to typical comprehenders (Oakhill et al., 2005).

In the present study, we examine grade-related changes in comprehension-monitoring skill as children progress through the early stages of beginning reading from first to third grade. Few studies have examined comprehension monitoring across grades among young readers. Although this work supports the viewpoint that comprehension monitoring is not an all-or-none phenomenon (i.e., a metacognitive insight that one has or does not have), we are less clear about the nature of its development over time and the factors that influence this. Further, there are no studies to date that examine its role as a potential foundational skill in the prediction of future reading comprehension in beginning readers, in addition to word reading. Finally, we examine the relations amongst other correlates of reading skill in beginner readers (i.e., working memory, vocabulary) and comprehension monitoring, and assess whether comprehension monitoring among first graders is a unique predictor of future reading skill in the context of these correlates of reading skill. Our interest in this work is therefore, in part, to determine whether comprehension monitoring, as measured at first grade, is a precursor to skilled reading comprehension at third grade.

The Relation Between Comprehension Monitoring and Text Comprehension

To comprehend a text, one must construct a coherent mental model (or representation) of the information presented in the text. This involves going beyond a representation of individual words or sentences; successful comprehenders integrate their meanings into a coherent whole (Johnson-Laird, 1983; Kintsch & Van Dijk, 1983). Importantly, this mental representation is refined continuously; as the text unfolds, readers and listeners integrate successive ideas and concepts into the existing mental model (Rapp & Kendeou, 2007). Thus, comprehension of a text is dynamic and integrative processing is essential. Theoretically, there is a clear role for comprehension monitoring in the construction of a coherent mental model and, therefore, for reading comprehension in general. Individuals who monitor their comprehension will identify when they do not know the meaning of a key word, when the information in the text does not accord with their background knowledge, and when two pieces of information are hard to integrate. In such instances, an individual with good comprehension monitoring skills is not aware that their comprehension has failed but they are able to take appropriate remedial action, such as re-reading or asking questions, generating inferences, and looking up the meanings of unknown words (Markman, 1981).

In this study we focus on texts that contain two conflicting pieces of information (i.e., internal inconsistencies). This type of error can be detected only if the participant actively engages in the construction of a mental model, integrating the information from each new proposition into the current representation. Thus, it is clear that the evaluation of one's comprehension is intimately connected with the type of integrative processing essential to construct a coherent mental model (Singer, 2013). Individuals who process text word-by-word or sentence-by-sentence, rather than integrating individual propositions and ideas with the

message as a whole, will fail to detect conflicts of internal consistency and will fail to construct a coherent mental model. For such reasons, comprehension monitoring can be conceptualized as a "higher level" language skill, distinct from such lower level language skills as grammar and vocabulary (Authors, 2004), and viewed as such because of its integrative (and thus higher level) role in text comprehension.

Development of Comprehension Monitoring Among Beginning Readers

Comprehension monitoring is a complex cognitive skill, yet prototypical monitoring behavior is observed even as early as the preschool years on structured comprehension tasks. For example, children under 3 years of age show awareness when actors, actions, objects and the temporal sequence of events in familiar stories are altered (Skarakis-Doyle, 2002). Comprehension monitoring improves between 8 and 11 years (Helder, van Leijenhorst, & van den Broek, 2016). Research that has compared performance on different types of errors demonstrates comprehension monitoring even in 5 year-olds with improvements seen up to the age of 11 (Baker, 1984). Young children find nonwords the easiest type of inconsistency to detect and internal inconsistency detection most difficult (Baker, 1984). This may be because nonword and knowledge violations involve checking information in the text against the young comprehender's stored knowledge, whereas detection of an inconsistency requires the comparison of the just-read or heard information with her current mental model. In addition, it may be that younger readers process text word-by-word or sentence-by-sentence, and so attend to the meaning of the individual propositions rather than the comprehensibility of the message-level of the text (Paris & Myers, 1981). However, Baker (1984) notes that this early work may provide an inaccurate estimate of monitoring skill because the children were given feedback after each attempt, which may have served as additional instruction in how to evaluate.

Empirically, the contribution of comprehension monitoring to children's reading comprehension has been demonstrated in different research designs. First, a number of studies have examined the concurrent relations between children's comprehension monitoring and their reading or listening comprehension when controlling for other key variables, such as vocabulary and working memory (Authors, 2001, 2004; Kim, 2015, 2016; Strasser & Rio, 2014). Such work shows that children's comprehension monitoring ability is a unique source of variance in reading comprehension in 8 to 10-year-olds (Authors, 2004) and listening comprehension and story book understanding in 6-year-olds (Kim, 2015; Strasser & Rio, 2014), although there is some evidence that the effects are only indirect (Kim, 2016). Second, another set of studies have examined comprehension monitoring among children who have difficulties comprehending what they read. These poor comprehenders have particular difficulties with detecting internal inconsistencies, a sign of poor integrative processing and inadequate mental model construction (Oakhill et al., 2005; van der Schoot, Reijntjes, & van Lieshout, 2012).

Several studies have examined the relation between working memory and comprehension monitoring. Working memory refers to the memory systems used to store and process information simultaneously (Baddeley & Hitch, 1974), which is essential for integrating new information with the existing mental model providing a theoretical base for a relation between working memory and both reading comprehension in general, and comprehension monitoring specifically. Performance on independent measures of working memory predicts variance in reading and listening comprehension in children and adults (Authors, 2004; Daneman & Merikle, 1996) and independent measures of working memory are related to comprehension monitoring in poor reading comprehenders aged 8 years and over (Oakhill et al., 2005; van der Schoot et al., 2012). Of interest, comprehension monitoring explains unique variance in concurrent measures

of reading comprehension at 8, 9, and 11 years over and above the contribution of working memory to reading comprehension (Authors, 2004). This finding indicates that working memory in itself is not sufficient to ensure good reading comprehension, and that higher-level language skills such as comprehension monitoring also play a key role, at least in the concurrent prediction of reading and listening comprehension (see also Kim, 2015; Strasser & Rio, 2014).

A final factor that we consider is vocabulary knowledge. Vocabulary knowledge is a powerful predictor of reading comprehension and comprehension monitoring skill (Authors, 2004). However, it is also related to working memory (Nation, Adams, Bowyer-Crane, & Snowling, 1999) leading some to argue that the relation between working memory and higher-level language skills such as comprehension monitoring or reading comprehension is indirect and mediated by vocabulary knowledge (Nation et al., 1999; van Dyke, Johns, & Kukona, 2014). Recent work on inference making, another higher-level skill that involves integrative processing, suggests that the relation between working memory and inference making is mediated by vocabulary knowledge (Authors, 2015a). Further, work with preschoolers finds that comprehension monitoring partially mediates the relation between working memory and story book comprehension (Strasser & Rio, 2014), although for 6-year-olds there is evidence for only an indirect relation between comprehension monitoring and listening comprehension (Kim, 2016) . Thus, to understand the unique relation between comprehension monitoring and reading comprehension, we need to control for variance associated with both of these correlates, working memory and vocabulary.

Research Aims and Hypotheses

The present study represents a longitudinal investigation of the growth of comprehension monitoring from first to third grade. Research questions examined were twofold. First, we

determined whether and to what extent there are grade-related changes in comprehension monitoring among beginning readers from first to third grade. Although we are aware of no prior studies of the development of comprehension monitoring longitudinally, several prior cross-sectional studies lead us to hypothesize that children's comprehension monitoring would show significant improvements as children moved from the first to third grade. In an early study, for instance, Markman showed that third graders exhibited significantly better comprehension monitoring than first and second graders, with the first and second graders performing similarly (Markman, 1977). That study examined children's ability to detect misleading and incomplete information in the orally presented instructions for performing a magic trick and playing a game, and thus did not examine comprehension monitoring within the context of constructing a mental model of a text. Baker (1984) did use narrative style texts to examine children's comprehension monitoring within an oral context. In that study, 5-, 7-, and 9-year-olds listened to passages with embedded problems (e.g., internal inconsistencies, nonsense words, and/or violations of prior knowledge) and were asked to identify when they noticed a problem. The youngest age groups were particularly poor at detecting the internal inconsistencies on first reading with detection of less than one (out of four) of this error type by the youngest children. Baker showed a systematic increase in comprehension monitoring across the ages, with a very large effect size (~1.9) differentiating the youngest and oldest groupings. The present study sought to determine whether these apparent grade-related changes are observed in a longitudinal sample followed from first to third grade. Due to the known association between comprehension monitoring with vocabulary and working memory skills (Nation et al., 1999; van Dyke et al., 2014; Authors, 2004), we were also interested to determine if these two correlates measured at Grade 1 were associated both with initial comprehension monitoring and growth in this ability.

The second question concerned whether and to what extent first graders' comprehension monitoring serves as a unique predictor of reading comprehension two years later when children are in third grade, accounting for decoding, working memory, and vocabulary in Grade 1. While studies have shown there to be concurrent relations between comprehension monitoring and reading comprehension (e.g., Authors, 2004; Kim, 2014), few studies have examined these relations prospectively and not in this age group (see Authors, 2014, for the prediction of comprehension monitoring at 8 years to reading comprehension at 11). Further, as noted above, the inter-relations among vocabulary, memory, and both reading comprehension and comprehension monitoring ability, require that we control for these language and memory correlates to understand fully the theoretical basis for any relation between comprehension monitoring and reading comprehension.

Methods

Participants

Participants were enrolled in a 5-year multi-site longitudinal study designed to investigate the language bases of reading comprehension from pre-kindergarten (~4 years) to third grade (~9 years). Up to 100 pre-kindergarteners and about 25 kindergarten through third graders at each of four study sites were recruited in year 1 of the study, during the 2010-11 academic year, and followed for up to five years until the child reached third grade. Full details of our enrolment can be found in Authors, 2016.

Across the four study sites, children were recruited in the first year of the study from multiple local school districts which invited their preschool to third-grade teachers to participate. In classrooms in which teachers consented to participate in the study, a recruitment packet was sent home to all children that included a brief questionnaire and a consent form. For the

consented children, study personnel examined the caregiver questionnaire and a teacher-completed screening form to exclude any child identified to have severe or profound disabilities and to be unable to converse in English. This process was conducted at four sites simultaneously until study quotas were achieved. Overall, 420 preschoolers, 124 kindergartners, 125 first graders, 123 second graders, and 123 third graders were recruited into the study at Year 1.

The present analysis includes children who were in Grade 1 (n = 125) in year 1 of the study and who were subsequently followed to Grade 3. The attrition rate from year 1 to year 3 was 9.6% (12 children); five children dropped out of the study between grades one and two and seven dropped between grades two and three, for a final sample of 113 third graders. For these 113 children, 64 were female (57%) and 85% resided in homes in which English was the only language spoken. The children were mostly White (n = 88, 78%), followed by Asian (n = 3, 3%), and African American (n = 3, 3%); in addition, four children were multi-racial (4%; 12% missing data). Median overall annual household income level was in the range of $30K-$60K; specifically, 25 caregivers (22.1%) reported earning $30K or less, 27 (23.9%) reported earning between $30K and $60K, 16 (14.2%) reported earning between $60K and $85K, and 28 caregivers (24.8%) reported a family income of $85K or more (15% missing data).

Attrition analyses conducted on demographic variables (mother's education and family income) as well as on the variables used in the main analyses for this study (see Table 1) showed that the attrited children were more likely to have less-educated mothers than those who were maintained in the study ($\chi^2(1) = 7.40$, $p = .005$), with no difference in annual household income. With respect to the main study variables, the attriters had significantly lower reading comprehension scores than those who remained in the study (M = 8.2, SD = 2.97 vs M = 10.35, SD = 3.16; $t(113) = 2.07$, $p = 0.04$) on one of the three reading comprehension measures (i.e. the

experimental reading comprehension measure), and a significant difference between attriters and non-attriters was apparent for expressive vocabulary (M = 88.33, SD = 13.76 vs M = 97.86, SD = 13.78; $t(123)$= 2.28, p = .02) and receptive vocabulary (M = 115.60, SD = 16.94 vs M = 130.60, SD = 16.37; $t(123)$ = 3.01, p < .01). No significant differences between the two groups were found on comprehension monitoring, working memory, decoding, word classes receptive and word classes expressive, as well as the two other indicators of reading comprehension.

General Procedures

The study procedures of relevance to this study involved assessments of children's skills during each year of the project. Each child completed one battery of assessments implemented across multiple sessions over a 22-week assessment window (January-May). There were approximately 35 direct measures implemented per year, divided into 11 blocks of two to four measures each, ranging in administration time from 15-45 minutes depending on the block. Most measures were administered at schools, although alternative locations were used on occasion (e.g., libraries, children's homes).

To ensure consistent administration of measures, training materials were developed by the study's staff and were available via a centralized project website for ease of accessibility to trainers and assessors irrespective of location. Prior to implementing any measure, assessors completed training modules specific to that measure, to include watching narrative presentations with video exemplars and completing online quizzes. Trainees then completed two mock administrations with a reliable research staff member at their respective site, during which research staff responded to the trainee's administration of the measure using a standardized script and provided feedback regarding administration. Additionally, research staff scored the trainee

for adherence to the administration and scoring protocols using a fidelity checklist specifically designed for that measure. In order for assessors to be able to administer a given measure, a fidelity of 90% or better was required.

With the exception of one measure that was administered to small groups of two to five participants, all children were assessed individually and were given frequent breaks between assessments. Measures requiring complex responses and scoring were audio-recorded and post-scored at a different time. Once measures were administered at each site, they were scanned to a centralized site that completed all required data processing.

Measures

Measures reported here represent five constructs: (1) decoding, (2) vocabulary, (3) working memory, (4) comprehension monitoring, and (5) reading comprehension. All measures were given to the children at all three grades of relevance to this study (first, second, third), although analyses used only one time-point typically.

Decoding. This construct was defined as a latent variable using four different measures. Two were subtests of the Woodcock Reading Mastery Test – Revised/Normative Update (WRMT-R/NU, Woodcock, 1998). The Word Attack subtest measures an individual's ability to apply phonic and structural analysis skills to unfamiliar words. It contains 45 items and administration continues until the 6 highest-numbered items on an easel page are failed or until all items have been administered. Reliability as reported in the manual for Grade 1 was .94. For our sample, reliability in Grade 1 was .92. The Word Identification subtest measures an individual's ability to identify written words in isolation. It contains 106 items; the basal is achieved with the first 6 consecutive correct responses that begin with the first item on an easel page, and the ceiling is defined as the last 6 consecutive incorrect responses that end with the last

item on an easel page. Reliability as reported in the manual for Grade 1 was .98. For our sample, reliability in Grade 1 was .96. In addition, two subtests of the Test of Word Reading Efficiency – Second Edition (TOWRE-2; Torgesen, Wagner, & Rashotte, 1999) were administered. The Sight Word Efficiency subtest is a timed assessment that measures the number of English words, ranging from high to low frequency, children can pronounce in 45 seconds with no errors. The Phonemic Decoding Efficiency subtest assesses the number of pronounceable nonwords, which range in complexity, children can pronounce in 45 seconds with no errors. The average test-retest reliability reported in the manual for the sight word and phonemic decoding was .93 and .94, respectively.

Vocabulary. This construct was defined as a latent variable using three different measures. First, the standardized and norm-referenced *Expressive Vocabulary Test, Second Edition* (EVT-2; Williams, 1997) was used to assess expressive vocabulary. During the assessment, children were shown a picture (e.g., an apple) and asked to provide one word to label it; then, they were asked to provide a single word synonym for the target word. Items were scored as correct only if both the label and a synonym were provided (0=incorrect, 1 = correct). The EVT-2 manual reports internal consistency across Grades 1-3 (alpha ranging from .94 to .97); for the current sample, internal consistency was also adequate for Grade 1 at .94. Second, the fourth edition of the Peabody Picture Vocabulary Test (PPVT-4; Dunn & Dunn, 2007) was used to assess receptive vocabulary. For this measure, assessors read a word and asked the child to point to the picture, out of four, that corresponded to the meaning of the target word. The PPVT-4 manual reports internal consistency ranging from .96-.97; for our sample internal consistency was adequate at .95. Last, the Word Classes receptive and expressive subtests from the Clinical Evaluation of Language Fundamentals – Fourth Edition (CELF-4; Semel, Wiig, &

Secord, 2003) were used to evaluate children's ability to understand relationships between words related by semantic class features (receptive subtest) as well as express the similarities and differences between those relationships (expressive subtest). For the receptive subtest children listened to three or four words and selected the two that were related. For the expressive subtest, children had to describe the relationship between the two words that they had previously selected. The expressive portion of the test was audio recorded and scored offsite by trained personnel in the research lab (ICC = .99). As reported in the manual, internal consistency for the receptive and expressive subtests were good, at .80 and .81, respectively. For the current sample, reliabilities of the receptive and expressive subtests were adequate at .88 and .77, respectively.

Working memory. The auditory memory subtest of the *Woodcock-Johnson III NU Test of Cognitive Abilities* (WJNU; Woodcock, McGrew, & Mather, 2001) was used to measure short-term auditory working memory, and required the children to store and manipulate verbal information. Children listened to a series containing both digits and words (e.g., 3…bread...1…lion). The children were then asked to reorder the series by first naming the words followed by the digits, while maintaining the sequential order in which they were presented (e.g., bread-lion-3-1). The stimuli were presented through a digital recording device (iPod or MP3 player) to ensure that the same timing gaps were used with all participants. There were seven blocks of increasing difficulty of three items each for a total 21 items, and a ceiling rule of three consecutive incorrect items within a block was applied. The reliability coefficient for the auditory working memory subtest reported in the test manual ranged from .89 to .96 across the 4- to 9-years-old age range. Internal consistency for our sample was adequate at .80.

Comprehension monitoring. Children's comprehension monitoring was assessed using an experimental measure based on the work of Authors (2006, 2011). The measure assessed children's ability to detect inconsistencies in orally presented stories. There were 12 test stories that were either entirely consistent (4 stories) or included inconsistent information (8 stories). Only the scores for the inconsistent stories were used for analyses. An example of an inconsistent story is: *"Last night Jill walked home through the park. There was no moonlight so Jill could hardly see her way. Jill often takes this route home. She walked along a narrow path. The moon was so bright that it lit the way. Jill lives on the other side of the park."* For the 8 inconsistent stories presented, children were asked (a) if the story made sense, and if they answered correctly to this question, they were then asked (b) what was wrong with the story (i.e., *"There was no moonlight so Jill could hardly see her way"* and *"The moon was so bright that it lit the way."*) Part (b) was scored as either incorrect (score = 0) or correct (score = 1), and it was this part which was used to compute the total score for this task which ranged from 0 to 8 points.

Although the same test stories were used at all three grades, the version used with second and third graders differed slightly from that administered to first graders in that for 6 of the 8 inconsistent stories there was more additional story material between the inconsistent parts within a story, to make these slightly more challenging. For example, in Grade 1 an excerpt from one of the inconsistent passages read as *"There was no moonlight so Jill could hardly see her way. Jill often takes this route home. She walked along a narrow path. The moon was so bright that it lit the way."* For the Grade 2-3 version of this passage, the story material between the inconsistent parts of the story read as *"Jill often takes this route home, because it is a good shortcut to her house. She walked along a narrow path."* In addition, the amount of additional text between the inconsistencies varied in length. In other words, in the previous example the

added text had nine words (i.e. *because it is a good shortcut to her house)* but the length of the added text varied between stories.

Before administering the experimental items, sufficient practice was given to ensure that children understood the task requirements. Up to five practice stories were administered (practice stories 1, 3 and 4 were inconsistent, and practice stories 2 and 5 were consistent). Children were prompted with the following instructions from assessors: *"I have some short stories. Some of them make sense, but some of them have got silly mistakes in them. I would like you to listen to each story and tell me whether it makes sense or does not make sense. Let's have a practice. Listen carefully to this story and tell me if it makes sense or does not make sense."* The first three practice stories were administered to all children. If children answered these correctly, they proceeded to the experimental items. However, if practice story 3 (inconsistent) was answered incorrectly children were administered practice story 4 (inconsistent). For this group of children, practice story 5 (consistent) was administered only if practice story 2 (consistent) was incorrectly answered. Children who answered practice story 3 (inconsistent) correctly, but not practice story 2 (consistent) were administered story 5 (consistent) but were not assessed on story 4 (inconsistent). In general, assessors provided more feedback during practice stories, and when children were not able to identify the inconsistency they were explicitly told what the inconsistency was. This was different from the experimental items, where no explanation or feedback was provided if the child was not able to identify the inconsistent part of the story. Internal consistency for the present sample ranged from .73 to .84 across grades. Prior analysis of concurrent validity, for a British version of the task, showed correlations with standardized measures of reading comprehension to range from .41 to .50 (Authors, 2011b). For the current sample, concurrent validity at each of the three grades was examined via correlations

with three different measures of reading comprehension (detailed below); for first, second, and third grade, correlations ranged from .37-.41, .32-.44, and .38-.46, respectively. In addition, a recent published study looking at the multi-dimensional nature of language in Grades 1 to 3, provided evidence of construct validity for this particular experimental measure (Authors, 2015c). Specifically, this same experimental comprehension monitoring task significantly loaded into the higher-level language construct with standardized loadings ranging from .59 to .70 across the three grades.

Reading comprehension. This construct was defined as a latent variable using three measures administered to assess reading comprehension. First, the Passage Comprehension subtest of the *Woodcock Reading Mastery Test – Revised/Normative Updated* (WRMT; Woodcock, 1998) was used to assess children's ability to read and comprehend a short passage and identify a key missing word. Before administering the test, examiners practiced with the students by pointing to a given sentence in the test book and saying "This says '*The cat is playing with a…'*" and then prompting the child to fill in the sentence with the word that belonged in the blank space. During actual administration, children read the one- or two-sentence passage to themselves. The manual reported split-half reliability of .94 and .92 for Grade 1 and 3, respectively. For the current sample, internal consistency was .89.

Second, the *Gates-MacGinitie Reading Test* was administered to assess how well children can read and understand passages. There were three levels to the test, corresponding to grade levels 1, 2, and 3. For the current analysis, Level 3 (i.e., Grade 3) of Form S of the Gates-MacGinitie Reading Tests, 4th edition (MacGinitie, MacGinitie, Maria, & Dreyer, 2000) was administered. This consisted of 11 short stories followed by multiple choice questions, for a

total of 48 items. Reliability (KR-20) as reported in the manual was .92. For the current sample, reliability was .92.

Third, an experimental measure of reading comprehension was administered to assess reading comprehension ability. The Reading Comprehension Measure (RCM) included five narrative passages and questions derived from the *Qualitative Reading Inventory-5* (QRI; Leslie & Caldwell, 2011), as well as six new passages and questions that were developed specifically for this study (1 narrative passage and 5 expository passages). The RCM involved presenting passages to children (both narrative and expository) followed by questions designed to assess comprehension of both implicit and explicit content of the passages. The passages differed for each grade. The Grade 3 measure, which is relevant to the present study, included a total of 4 stories (2 narrative and 2 expository), and 28 questions (possible range of 0-28). For the current sample, reliability was .80.

Results

Preliminary Data Analysis and Considerations

Children in the present study were those in the longitudinal study enrolled in first grade with follow-through to third grade. Children were nested in classrooms at each year of the study, with increasing dispersion of children into different classrooms with each successive grade. At the beginning of the study (Grade 1), the 113 children were nested in 42 classrooms and the mean number of study children per classroom was 2.0 (SD = 2.1, range from 1 to 9). At Grade 2, children were nested in 60 classrooms with a mean number of study children per classroom of 1.9 (SD = 1.04, range from 1 to 5), and at Grade 3, children were nested in 68 classrooms with a mean of 1.7 study children per classroom (SD = 1.31, range from 1 to 8). For the purposes of this study and due to the small number of children nested within classrooms, we did not model

the nesting at the classroom level. Thus, to examine growth in comprehension monitoring, a latent growth curve model with three measurement occasions (Grades 1-3) was analyzed.

Table 1 reports descriptive statistics of the measures that were used for analyses as well as children's age in months at each of the study time-points; raw scores are reported for all measures and standard scores are reported when available. Notice that the only measure in Table 1 that is being used across all three grades is comprehension monitoring. Table 2 presents the Pearson correlations amongst all measures included in the present study. Except for the correlation between word attack (WRMT) in Grade 1 and comprehension monitoring in Grade 2, receptive vocabulary in Grade 1 (PPVT) and the receptive subtest of word classes receptive in Grade 1, and word classes receptive in Grade 1 with comprehension monitoring in Grades 1 and 3, which were small and non-significant, all other correlations were significant at $p < .05$.

As part of the preliminary data analysis, the three latent constructs used across aims 1 and 2 (i.e. decoding, vocabulary, and reading comprehension) were individually examined to ensure that the loadings and model fit indices were acceptable. To interpret the results of these confirmatory factor analysis (CFAs), we report three model fit indices as suggested by Hancock and Mueller (2006): an absolute (Standardized Root Mean Square Residual, or SRMR; values below .08 are considered acceptable), a parsimonious (Root Mean Square Error of Approximation, or RMSEA values below .05 are considered acceptable), and an incremental index (Comparative Fit Index, or CFI; values above .95 are considered acceptable). Note that as suggested by Kenny, Kaniskan, and McCoach (2015), models with low degrees of freedom (as is the case for these constructs) can have artificially large values for RMSEA. For the decoding construct, results suggested good fit for CFI = .97, and SRMR = .03, but inadequate fit based on RMSEA = 0.21, 90% CI(0.11,0.33). Standardized loadings for the indicators were all significant

and greater than .70. For the vocabulary construct, results suggested good fit for CFI = 1.00, and SRMR = 0.015. Although the point estimate of the RMSEA was excellent (RMSEA = 0.00), the 90% confidence interval included a range from excellent to poor RMSEA values, i.e. 90% CI (0.00, 0.15). For the vocabulary construct, the residual variances of the CELF Word Classes Expressive and Receptive indicators were allowed to correlate. Further, the residual variance of the Expressive Vocabulary Test (EVT) was set to zero due to an unacceptable solution. Standardized loadings for the vocabulary construct were all significant, and the magnitude of the loadings for the PPVT, CELF Word Classes Expressive, and CELF Word Classes Receptive were .78, .45, and .29, respectively. Last, the reading comprehension standardized loadings were all significant and greater than .70. Since the reading comprehension construct had only three indicators, this model was just identified and evaluating its model fit does not apply because by default such solutions always have perfect fit (Brown, 2006, p. 66).

Grade-Related Changes in Comprehension Monitoring

The first aim of this study was to examine whether and to what extent there were grade related changes in comprehension monitoring among beginning readers from first to third grade. Further, we were also interested to determine if vocabulary and working memory skills, measured at first grade, were associated with initial comprehension monitoring and growth in this ability. We first describe the results of grade-related change in comprehension monitoring followed by the results of the associations with vocabulary and working memory.

The data presented in Table 1 show that children's mean scores on the comprehension monitoring measure increased from Grade 1 (M = 4.56, SD = 2.11) to Grade 2 (M = 6.03, SD = 1.68) and Grade 3 (M = 6.37, SD = 1.43). Specifically, the means of comprehension monitoring across time suggested that there was an increase in this skill from Grade 1 to Grade 3, with the

average change between Grade 1 and Grade 2 ($t(108) = 6.92$, $p < .0001$, $d = .67$) being significantly larger than that from Grade 2 to Grade 3 ($t(106) = 1.62$, $p = 0.108$, $d = .17$). To further understand the growth in comprehension monitoring across time as well as to investigate the variability of comprehension monitoring at Grade 1 (i.e. intercept), a series of latent growth models were run in Mplus 7 (Muthén & Muthén, 2012). Prior to fitting growth models, spaghetti plots looking at all three time points of a random group of participants were studied (see Figure 1). In addition, individual plots for all participants were also examined to visualize trends and growth patterns based on all available data. These plots suggested a pattern of non-linear growth in comprehension monitoring. Further, the pattern in the observed means of comprehension monitoring across time, as well as the attenuation of correlations in comprehension monitoring between Grade 1 and Grade 2 ($r = .33$) compared to Grade 2 and Grade 3 ($r = .20$) provided further support of a non-linear trend.

When fitting growth curves, the number of parameters is equal to the number of time points minus 1, if all parameters are to be estimated. In the case of three time points, which is what we had, one could fit a linear model if all eight parameters (i.e. intercept, linear slope, variance of intercept, variance of slope, covariance of intercept and slope, and residual variances of all the three time points) were to be estimated. This growth curve will be identified and will have one degree of freedom. Alternatively, one could also fit a non-linear growth curve with three time points but at least one variance/covariance component would need to be fixed for the model to be estimated and identified. There are some recent examples in the literature in which quadratic and piece-wise models have been fit with three time points by constraining the variance or covariance of one of the growth parameter components (Codding, Mercer, Connell, Fiorello, & Kleinert, 2016; Kamata, Nese, Patarapichayatham, & Lai, 2013).

Table 3 presents the results of the latent growth curve models for comprehension monitoring, in which time was coded using grade (i.e. Grade 1 as 0, Grade 2 as 1, Grade 3 as 2) where Grade 1 represented the intercept or starting point. For model 1, an unconditional growth model with linear time suggested a positive and significant effect in growth of comprehension monitoring; for every additional grade, comprehension monitoring increased by 0.81 points. A model estimating the variance in the linear slope (not shown) failed to converge because the variance term of the linear slope was negative. Constraining this negative and non-significant linear variance term to zero eliminated this problem and produced plausible parameter estimates. After fixing the variance in the linear slope to zero, a model with a quadratic slope was estimated. This non-linear model is consistent with patterns in plots and the pattern of the overall means for comprehension monitoring across grades (Table 1), which suggested a deceleration in the positive growing trend. When modeling quadratic growth, we first ran a model that estimated the variance in the quadratic term (not shown). Convergence problems analogous to those reported for the variance of the linear slope parameter (i.e. negative variance term) were encountered and thus the variance of the quadratic slope parameter was constrained to zero. The results of the quadratic model presented in model 2 (Table 3) provide evidence of a deceleration in the growth of comprehension monitoring. To assess the improvement in model fit between the linear model (model 1) and the quadratic model (model 2), a chi-square deviance test was used. Based on this chi-square test, $\chi^2(1) = 9.61$, $p < .001$, we viewed the model with quadratic time as superior to the model with only the linear trend.

Associations of Vocabulary and Working Memory with Grade-Related Changes in Comprehension Monitoring

As a continuation of the first aim, we were also interested in exploring whether vocabulary and working memory, measured at first grade, were associated with both initial comprehension monitoring and growth in this ability. From the results of the latent growth models examining grade-related changes in comprehension monitoring we learned that there was overall growth in comprehension monitoring that was faster between first and second grade than between second and third grade. However, as indicated by the lack of reliable variance in the linear- and quadratic- growth parameters, there were no individual differences in the growth of comprehension monitoring, suggesting that all children grew at the same rate over time. Since there was not individual variability in the growth rate of comprehension monitoring, this implied that we could not examine the association between vocabulary and working memory with this growth. Since the only variability in the comprehension monitoring growth models came from the individual variation in the intercept or starting point (i.e. Grade 1), associations of vocabulary and working memory in first grade could only be studied as predictors of the intercept's variance. We describe these associations next.

Model 3 presents results of the association of vocabulary and working memory (both measured at Grade 1) with the intercept (i.e. Grade 1) of the growth trajectory of comprehension monitoring. Results of this analysis appear as model 3 in Table 3, and show that vocabulary had a significant and positive concurrent relationship with comprehension monitoring ($b = 0.05$, $p < .001$), although this was not true for working memory ($b = 0.02$, $p = .18$). To assess the unique contribution of Grade 1 children's vocabulary skills on the variance around the intercept (i.e., Grade 1) of comprehension monitoring, we looked at the reduction in variance around the

intercept. To do so, model 3 was re-run without vocabulary to obtain the intercept variance estimate prior to adding vocabulary as a predictor (intercept variance = .72). Then this variance in the intercept was compared to that of model 3 which allowed us to assess the unique contribution of vocabulary as opposed to vocabulary and working memory. We calculated the proportion of variance explained due to vocabulary by comparing the variance in the intercept of model 3 in Table 3 (intercept variance = .33) and the intercept variance when vocabulary was not included in this model (intercept variance = .72). This suggested that the proportion of variance accounted for by the addition of vocabulary was 54.16% (i.e. (.72-.33)/.72).

Note that model fit indices for models 1 and 2 as reported in table 3 were weak by commonly used benchmarks, but individual components of the growth trajectories (intercept, slope, variability in the intercept) were of more substantive interest (Bollen & Curran, 2006). In addition, model fit improved once covariates were included in the model to explain the variability in the intercept. In summary, models 1 and 2 suggest that comprehension monitoring grows significantly from first to third grade, but that there is deceleration over time. Model 3 suggests that vocabulary at Grade 1 is responsible of explaining some of the observed variability in the intercept or starting point (i.e. Grade 1) of comprehension monitoring.

Beginning Readers' Comprehension Monitoring and Future Reading Comprehension

Our second aim was to determine whether and to what extent first graders' comprehension monitoring serves as a unique predictor of reading comprehension two years later when children are in third grade, accounting for decoding, working memory, and vocabulary at the first time-point. This aim was informed by the results in aim 1, where we found that the intercept in comprehension monitoring (i.e. Grade 1) had significant variability. Thus, the intercept of the growth trajectory in comprehension monitoring was used as the predictor of

Grade 3 reading comprehension. In summary, a structural equation model (SEM) was used to predict children's reading comprehension at grade three; three predictors all measured at Grade 1 (a latent construct of decoding, a latent construct of vocabulary, an observed measure of working memory), and the intercept of the growth in comprehension monitoring as specified in Table 3 model 2, were included in the final model. Decoding, vocabulary, and working memory were used as covariates given their positive, predictive correlations with reading comprehension (Authors, 2004).

To determine the extent to which the intercept from the comprehension monitoring growth model served as a unique predictor of reading comprehension, the model fitting strategy was conducted in three steps. In step 1, decoding was included in the model. In step 2, both vocabulary and working memory were added to the model fitted in step 1. In step 3, the intercept of comprehension monitoring was added to the model fitted in step 2. At each step, R^2 was noted to be able to determine the unique contribution of comprehension monitoring, our main predictor of interest. The models are presented in Table 4, and a graphical representation of the final model is included in Figure 2.

Model 1 in Table 4 presents the result of including Grade 1 decoding as a predictor of reading comprehension in Grade 3; not surprisingly, decoding in first grade was a significant and positive predictor of reading comprehension ($\beta = .75$, $p < .0001$), explaining about 56% of the variance in third-grade reading comprehension. In terms of model fit, it had a good fit based on SRMR (.04), and good fit based on CFI (.95). For RMSEA, model fit was not acceptable (.15, 90% CI[.10, .20]); this held true across the three models.

For model 2, vocabulary and working memory were added to model 1. Decoding remained a significant and positive predictor of reading comprehension ($\beta = .44$, $p < .0001$), as

did vocabulary (β = .35, p < .0001) and working memory (β = .16, p = .03). Adding vocabulary and working memory accounted for about 10% of additional variance in Grade 3 reading comprehension. In terms of model fit, it had a good fit based on both SRMR and CFI.

For model 3, the intercept of the comprehension monitoring growth model was added. Even after controlling for decoding, vocabulary, and memory, first-graders' comprehension monitoring substantially and positively predicted third-grade reading comprehension (β = .36, p < .001). Additionally, comprehension monitoring was uniquely responsible for explaining an additional 8% of the variance in reading comprehension in Grade 3. Substantively, the pattern of results for the decoding and working memory predictor included in step 1 and 2 remained the same but the vocabulary and working memory constructs were no longer significant in the final model: decoding (β = .43, p < .0001), vocabulary (β = .15, p = .131), and memory (β = .13, p = .054). For model fit, it had an acceptable SRMR and close to adequate fit based on CFI.

Discussion

In this study, we investigated comprehension monitoring for beginning readers as they progressed from first to third grade. Our first aim was to determine whether there are grade-related changes in comprehension monitoring within this period and the extent of those changes. Additionally, we were interested in determining if first-grade vocabulary and working memory skills were related to this growth. We found that children's ability to monitor their comprehension grew significantly from first to third grade, but that there was a deceleration in growth over time. Between first and second grade, children's comprehension monitoring grew more so than occurred between second and third grade. Further, we found no reliable individual differences in this growth trajectory suggesting that all children grew at the same rate over time. Since the only reliable variability in the growth trajectory of comprehension monitoring was

found in the intercept, we could assess only the association of first-grade vocabulary and working memory skills with the intercept of comprehension monitoring, which was centered at Grade 1. For this later set of analysis, we found that Grade1 vocabulary, but not working memory, was a significant predictor of the observed variation in initial comprehension monitoring skills. Our second aim was to determine whether comprehension monitoring (defined as the intercept centered at Grade 1 from the comprehension monitoring growth model) was a unique predictor of reading comprehension two years later (in third grade), after taking into account the contributions made by variables that are strongly associated with reading comprehension. Importantly, even after controlling for the contributions of word decoding, working memory, and vocabulary, comprehension monitoring in Grade 1 predicted unique variance in future reading comprehension. The results of this study, in the aggregate, extend our understanding of young readers' development of comprehension monitoring, the theoretical relations between comprehension monitoring and reading comprehension, and yield practical implications for the elementary-grade classroom, which we discuss below.

The first major finding is that children's comprehension monitoring is actively developing between first and third grade, with the greatest volume of growth occurring between first and second grade. This time period corresponds to one of significant growth in word-reading abilities, yet the study results show that component skills related significantly to reading comprehension are also in an active state of development. The deceleration in comprehension monitoring from second to third grade, following a period of significant growth, is not entirely unexpected. Work by Skibbe and her colleagues, which studied reading development from preschool through second grade, showed rapid growth in decoding and comprehension skills from kindergarten to first grade followed by slowing growth in second grade (Skibbe, Grimm,

Bowles, & Morrison, 2012). This pattern is similar to what we observed in this study with respect to comprehension monitoring. Pragmatically, such work suggests that first grade is a highly prominent context in which children show rapid, substantial growth in an array of reading skills, including component skills that contribute to reading comprehension. As the first year of formal, full-time, academically focused schooling for many children, reading development appears to approximate a 'growth spurt' followed by deceleration. Notably, our study did not find evidence of reliable individual variability in the growth of comprehension monitoring. In other words, our findings suggest that all children grew at the same rate and that the only reliable variability in the growth of comprehension monitoring was at the intercept (i.e. Grade 1). The lack of individual variability in the growth of comprehension monitoring is discussed in more detail in the limitations.

Theoretically, the finding of an overall growth in comprehension monitoring that was faster between first and second grade than between second and third grade, highlights the need to consider the construct of comprehension monitoring as a developmental construct. One interpretation of this pattern of growth is that the metacognitive awareness required to successfully monitor comprehension develops substantially between grades 1 and 2, such that additional gains in performance determined by language skills or cognitive resources will be slight, once such awareness is achieved. Other metacognitive skills related to literacy have been described as an all-or-none phenomenon, representing an insight, rather than a skill that gradually develops over time, for example children's ability to recognize the symbolic relations between written language and spoken language (Bialystok & Luk, 2007). However, our data do not fully support that interpretation for two reasons. First, we find growth between Grade 1 through 3, and second scores in Grades 2 and 3 do not indicate ceiling performance by a majority

of children. Instead, this finding suggests that after awareness of sense monitoring is achieved, performance may still be influenced by other factors that limit children's ability to monitor their comprehension. For example, across grades 1 through 3, the ease with which children retrieve word meanings and construct sentence meanings improves; children who are more fluent processors of language will have greater attentional and processing resources to devote to comprehension monitoring. Thus, developing language and cognitive skills may explain additional gains in comprehension monitoring.

Methodologically, it is also possible that the deceleration observed in the growth trajectory of comprehension monitoring is an artifact of our measure of comprehension monitoring. For instance, it may be that our comprehension monitoring task was too easy for many children in Grade 3, and there were some children who hit the ceiling of the measure at each time-point. At the same time, it is also the case that by Grade 3, all of the children had prior experience with this paradigm in the earlier grades and practice may have played a role in 'overestimating' performance. However, it should be pointed out that the results observed here for growth in comprehension monitoring, with deceleration between second and third grade following a period of substantial skill incline, are similar to those reported by Skibbe and colleagues, as referenced earlier.

With respect to improving our understanding of comprehension monitoring in beginning readers, we found that vocabulary skill but not working memory was related to children's ability to monitor their own comprehension. It is important to note that vocabulary (measured at Grade 1) was associated with variability in the intercept of the trajectory of comprehension monitoring, not with its growth rate. Since there was no reliable variation in the growth trajectory of comprehension monitoring, vocabulary nor working memory could be used to predict this

growth rate since it was fixed. With regards to the association of vocabulary skill but not working memory with the intercept of comprehension monitoring, we find convergence and also a notable difference with a recent study by Kim (2015) of Grade 1 children in South Korea. In Kim's study vocabulary was found to be a strong predictor of concurrent comprehension monitoring skill in Grade 1. Similarly, our study found that children's vocabulary in Grade 1 accounted for significant variability in the comprehension monitoring skills of first graders. Together, these findings confirm the influence of vocabulary's status as a lower-level or foundational language skill that supports more complex types of language processing. Other recent work also has demonstrated the influence of vocabulary on another higher-level language skill, namely inference making (Authors, 2015a, 2015b; Lepola, Lynch, Laakkonen, Silven, & Niemi, 2012). The size of the correlation between the two variables, together with other work demonstrating a separation between lower-level language skills (including vocabulary) and higher-level language skills (including comprehension monitoring) in grades 1 through 3 (Authors, 2015c), supports the viewpoint that these language skills are related but cannot be assumed to serve as proxies for each other. Critically, our study adds to Kim's (2015) finding with a different population, demonstrating that the relation between the lower-level skill of vocabulary and the higher-level skill of comprehension monitoring generalizes across language and school systems.

An interesting difference in findings between the present study and that of Kim's (2015 and 2016) studies was the contribution of memory: this was a significant predictor in Kim's work, but did not explain unique variance in comprehension monitoring in ours. We propose that this difference arises largely for methodological reasons. To assess working memory, Kim (2015, 2016) used a listening span task, in which participants first judge the truth of a sentence

and then remember a target word (in this instance, the first word) from that sentence for later recall. It is not surprising that a memory task that taps sentence comprehension and monitoring for sense should predict performance on a passage-level comprehension-monitoring task. In contrast, our measure of working memory sought to minimize the language comprehension component of the task; children were required to re-order unrelated digits and words. Also, it did not require metacognitive sense judgments. Our study therefore suggests that working memory *per se* is not a unique predictor of concurrent comprehension monitoring, and that some of the reported relations between working memory and language comprehension tasks may arise because of shared processing requirements and metacognitive demands. Other work also has demonstrated a significant correlation between comprehension monitoring and a listening span working memory task, but a much weaker relation with a number based working memory task (Authors, 2004).

The most important contribution of the present study is the demonstration that comprehension monitoring defined as the intercept from the growth in comprehension monitoring predicts reading comprehension in Grade 3, over and above the contributions of children's decoding, vocabulary, and working memory. Of note, a sizeable, 74%, proportion of variance in third-graders' reading comprehension was explained by these four variables. Theoretically, this finding adds to the growing body of evidence showing that higher-level language skills are critical foundations for reading and listening comprehension, over and above lower-level skills such as vocabulary (see also Lepola et al., 2012). Clearly, words and sentences are the building blocks of meaning, and it therefore follows that weaknesses at this basic level may limit the ability to engage in the integrative and evaluative processing involved in comprehension monitoring, which supports the construction of the mental model of a text's

meaning. However, our findings do not lend support to the viewpoint that variation in foundational language skills or memory is the critical determinant of comprehension outcomes (Hulme & Snowling, 2011; Perfetti, Stafura, & Adlof, 2013). Instead, our finding that comprehension monitoring skills in Grade 1 predict reading comprehension at Grade 3, adds to a growing body of evidence that identifies an independent contribution of higher-level language skills, such as comprehension monitoring (Authors, 2012a) and also inference making, which we did not evaluate here (Lepola et al., 2012; Authors, 2015b). Thus, we conclude that basic language skills or memory alone are insufficient to support the construction of the mental model.

Practically, the finding that comprehension monitoring at Grade 1 predicts reading comprehension in Grade 3 confirms the need to include support for higher-level language skills in beginning reading instruction, to include explicitly helping young children to learn how to monitor their own comprehension of text. Recent research suggests that young children can improve their comprehension monitoring in the context of explicit instruction. Specifically, a study of the impacts of an experimental curriculum designed to improve children's lower- and higher-level language skills, including comprehension monitoring, showed that preschool-aged children exposed to the curriculum demonstrated improvements on a comprehension-monitoring task compared to controls (Authors, 2015d). The approach used was derived from strategy instruction used with struggling adolescent readers, who were taught how to identify when text they are reading is 'clicking' versus 'clunking' (i.e., making sense or not making sense) (Klingner & Vaughn, 1999). The experimental curriculum mentioned above similarly involves teachers explicitly teaching young children how to analyze when texts they are listening to make sense or do make sense. Given the importance of comprehension monitoring to reading comprehension, it will be necessary to identify effective strategies to improve young children's

ability to monitor their comprehension and, in turn, determine whether improvements in comprehension monitoring have positive, causally interpretable effects on reading comprehension. Given the evidence of lack of variation in the rate of growth of comprehension monitoring, this study suggests that the best timing to boost comprehension monitoring in children will be early on during formal education (i.e. Grade 1).

Limitations, Implications, and Conclusions

A strength of our study is the inclusion of children from several geographical locations across the USA, the longitudinal design, the multiple indicators used to define vocabulary, decoding, and reading comprehension, and the convergence of key findings with Kim (2015). However, several key limitations warrant note. The first key limitation was the use of single indicators for comprehension monitoring, and the possibility that ceiling effects limited our ability to detect further growth in comprehension monitoring. Measuring comprehension monitoring using a single measure prevented us from empirically testing that what was measured was the intended construct of comprehension monitoring. Specifically, the fact that vocabulary accounted for about 54% of the variance in the intercept of comprehension monitoring could be seen as a lack of support for thinking of comprehension monitoring as measuring something other than vocabulary skills. Although we expect these skills to be correlated, as evidenced in this study and as shown in Kim's work (2015, 2016), future research should be devoted to the understanding of the dimensionality of comprehension monitoring and the extent to which it represents something other than vocabulary skills. Future work should also consider the use of additional measures of comprehension monitoring including nonsense words (or infrequent unfamiliar words for ecological validity) and violations of prior knowledge, and also the memory processing requirements of different types of comprehension monitoring tasks. However, we

note that the inconsistency detection paradigm is a widely used measure of comprehension monitoring in many other studies (e.g., Kim, 2015, 2016, Oakhill et al., 2005; Authors 2012). The second limitation is that the comprehension monitoring measure that we used is an experimental measure, and the validity of our materials needs to be confirmed with a different sample. Related to this, future studies should also look into the developmental trajectory of comprehension monitoring in beginning readers and assess whether reliable individual growth is observed. The third limitation is that our measure of comprehension monitoring did not assess whether children detected the errors during presentation of the story, or later when prompted by the sense question. In older readers, the study of moment-by-moment processing of text with eye tracking has been successful (Connor, Radach, Vorstius, Day, McLean, & Morrison, 2015). Paradigms sensitive to younger children's comprehension monitoring, such as listening time tasks (Fecica & O'Neill, 2010) are needed to determine the locus of error identification. Last, we cannot assess the causal relations between early comprehension monitoring and future reading comprehension. Training studies that seek to improve comprehension monitoring and subsequent effects on reading comprehension will be important for understanding whether comprehension monitoring is a correlate of reading comprehension or whether it directly supports improved comprehension of text.

In closing there are two practical implications that stem from this work that we choose to highlight. The first is that successful reading comprehension is determined by multiple oral language skills and cognitive resources: it requires a foundation of decoding, lower-level and higher-level oral language skills, in addition to working memory. That viewpoint is shared with a recent analysis of the National Early Literacy Panel (Authors, 2012b), and suggests that higher-level language skills, such as comprehension monitoring, should be included in the beginning-

reading curriculum. The second is that we should consider training higher-level oral language skills, such as comprehension monitoring, to minimize the risk of later reading comprehension failure. Comprehension monitoring can successfully be trained in the early years, as we noted previously (Connor et al., 2014). A priority for future work is to determine whether such training benefits distal measures of reading comprehension.

References

Authors, (2001)

Authors, (2004)

Authors, (2006)

Authors, (2011a)

Authors, (2011b)

Authors, (2012a)

Authors, (2012b)

Authors, (2013)

Authors, (2014)

Authors, (2015a)

Authors, (2015b)

Authors, (2015c)

Authors, (2015d)

Authors, (2016)

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *The psychology of learning and motivation, 8*, 47-89.

Baker, L. (1984). Children's effective use of multiple standards for evaluating their comprehension. *Journal of Educational Psychology, 76*(4), 588.

Bialystok, E., & Luk, G. (2007). The universality of symbolic representation for reading in Asian and alphabetic languages. *Bilingualism: Language and Cognition, 10*(02), 121-129.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective* (Vol. 467): John Wiley & Sons.

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Codding, R. S., Mercer, S., Connell, J., Fiorello, C., & Kleinert, W. (2016). Mapping the Relationships Among Basic Facts, Concepts and Application, and Common Core Curriculum-Based Mathematics Measures. *School Psychology Review School Psychology Review, 45*(1), 19-38.

Connor, C. M., Phillips, B. M., Kaschak, M., Apel, K., Kim, Y. S., Al Otaiba, S., . . . Lonigan, C. J. (2014). Comprehension tools for teachers: Reading for understanding from prekindergarten through fourth grade. *Educational Psychology Review, 26*, 379-401.

Connor, C. M., Radach, R., Vorstius, C., Day, S. L., McLean, L., & Morrison, F. J. (2015). Individual differences in fifth graders' literacy and academic language predict comprehension monitoring development: An eye-movement study. *Scientific Studies of Reading, 19*, 1140134.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3*(4), 422-433.

Dunn, D. M., & Dunn, L. M. (2007). *PPVT-IV*: *Peabody picture vocabulary test*. Pearson.

Fecica, A. M., & O'Neill, D. K. (2010). A step at a time: Preliterate children's simulation of narrative movement during story comprehension. *Cognition, 116*, 368-381. doi: 10.1016/j.cognition.2010.05.014

Hancock, G. R., & Mueller, R. O. (2006). *Structural equation modeling: A second course*. Greenwich, Conn: IAP.

Helder, A., Van Leijenhorst, L., & van den Broek, P. (2016). Coherence monitoring by good and poor comprehenders in elementary school: Comparing offline and online measures. *learning and individual differences, 48*, 17-23.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and writing, 2*(2), 127-160.

Hulme, C., & Snowling, M. J. (2011). Children's reading comprehension difficulties: nature, causes, and treatments. *Curent Directions in Psychological Science, 20*, 139-142.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*: Harvard University Press.

Kamata, A., Nese, J. F. T., Patarapichayatham, C., & Lai, C.-F. (2013). Modeling Nonlinear Growth with Three Data Points: Illustration with Benchmarking Data. *Assessment for Effective Intervention, 38*(2), 105-116.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The Performance of RMSEA in Models With Small Degrees of Freedom. Sociological Methods & Research, 44(3), 486-507. doi:10.1177/0049124114543236

Kershaw, S., & Schatschneider, C. (2012). A latent variable approach to the simple view of reading. *Reading and writing, 25*(2), 433-464.

Kim, Y. S. (2015). Language and cognitive predictors of text comprehension: Evidence from multivariate analysis. *Child development, 86*(1), 128-144.

Kim, Y. S. (2016). Direct and mediated effects of language and cognitive skills on comprehension of oral narrative texts (listening comprehension) for children. *Journal of Experimental Child Psychology, 141*, 101-120.

Kintsch, W., & Van Dijk, T. A. (1983). Strategies of discourse comprehension.

Klingner, J. K., & Vaughn, S. (1999). Promoting reading comprehension, content learning, and English acquisition though Collaborative Strategic Reading (CSR). *The Reading Teacher*, 738-747.

Lepola, J., Lynch, J. S., Laakkonen, E., Silven, M., & Niemi, P. (2012). The role of inference making and other language skills in the development of narrative listening comprehension in 4–6-year-old children. *Reading Research Quarterly, 47*, 259-282. doi: 10.1002/RRQ.020

Leslie, L. & Caldwell, J. (2011). Qualitative reading inventory-5. MN: Pearson Assessment.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). Gates-MacGinitie Reading Tests (4th ed.). Itasca, IL: Riverside Publishing.

Markman, E. M. (1977). Realizing that you don't understand: A preliminary investigation. *Child development*, 986-992.

Markman, E. M. (1981). Comprehension monitoring. *Children's oral communication skills*, 61-84.

Muthén, L. K., & Muthén, B. O. (2012). Mplus statistical modeling software: Release 7.0. *Los Angeles, CA: Muthén & Muthén*.

Nation, K., Adams, J. W., Bowyer-Crane, C. A., & Snowling, M. J. (1999). Working memory deficits in poor comprehenders reflect underlying language impairments. *Journal of experimental child psychology, 73*(2), 139-158.

Oakhill, J., Hartt, J., & Samols, D. (2005). Levels of comprehension monitoring and working memory in good and poor comprehenders. *Reading and writing, 18*(7-9), 657-686.

Paris, S. G., & Myers, M. (1981). Comprehension monitoring, memory, and study strategies of good and poor readers. *Journal of Literacy Research, 13*(1), 5-22.

Perfetti, C. A., Stafura, J. Z., & Adlof, S. M. (2013). Reading comprehension and reading comprehension problems: a word-to-text integration perspective. In B. Miller, L. E. Cutting & P. McCardle (Eds.), *Unravelling reading comprehension: behavioral, neurobiological, and genetic components* (pp. 22-32). Baltimore: Paul Brookes Publishing Co.

Protopapas, A., Simos, P. G., Sideridis, G. D., & Mouzaki, A. (2012). The components of the simple view of reading: A confirmatory factor analysis. *Reading Psychology, 33*(3), 217-240.

Rapp, D. N., & Kendeou, P. (2007). Revising what readers know: Updating text representations during narrative comprehension. *Memory & Cognition, 35*(8), 2019-2032.

Semel, E., Wiig, E., & Secord, W. (2003). Clinical evaluation of language fundamentals-IV. Marickville: Harcourt Assessment.

Singer, M. (2013). Validation in reading comprehension. *Current Directions in Psychological Science, 22*(5), 361-366.

Skarakis-Doyle, E. (2002). Young children's detection of violations in familiar stories and emerging comprehension monitoring. *Discourse Processes, 33*(2), 175-197.

Skibbe, L. E., Grimm, K. J., Bowles, R. P., & Morrison, F. J. (2012). Literacy growth in the academic year versus summer from preschool through second grade: Differential effects of schooling across four skills. *Scientific Studies of Reading, 16*(2), 141-165.

Strasser, K., & Río, F. D. (2014). The role of comprehension monitoring, theory of mind, and vocabulary depth in predicting story comprehension and recall of kindergarten children. *Reading Research Quarterly, 49*, 169-187.

Torgesen, J. K., Wagner, R., & Rashotte, C. (1999). TOWRE–2 test of word reading efficiency. *Austin, TX: Pro-Ed*.

Tunmer, W. E., & Chapman, J. W. (2012). The simple view of reading redux vocabulary knowledge and the independent components hypothesis. *Journal of learning disabilities, 45*(5), 453-466.

van der Schoot, M., Reijntjes, A., & van Lieshout, E. C. (2012). How do children deal with inconsistencies in text? An eye fixation and self-paced reading study in good and poor reading comprehenders. *Reading and writing, 25*(7), 1665-1690.

Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition, 131*, 373-403.

Wagoner, S. A. (1983). Comprehension monitoring: What it is and what we know about it. *Reading Research Quarterly, XVIII, 328-346.*

Williams, K. T. (1997). Expressive Vocabulary Test Second Edition (EVT™ 2). *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*, 864-872.

Woodcock, R. W. (1998). Woodcock Reading Mastery Tests – Revised/Normative Update. Circle Pines, MN: American Guidance Service/Pearson Assessments.

Woodcock, R.W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III Test of Cognitive Abilities. Itasca, IL: Riverside Publishing.

Table 1
Descriptive Statistics of Child Assessments

	N	*M*	*SD*	*Range*
Age (G1)	113	84.81	4.19	73-93
Age (G2)	110	96.28	4.06	86-105
Age (G3)	110	108.16	4.00	98-118
Decoding				
WRMT-Word Attack raw (G1)	112	20.89	7.94	6-39
WRMT-Word Attack standard (G1)	112	117.36	8.67	97-139
WRMT-Word Identification raw (G1)	113	49.19	12.71	19-83
WRMT-Word Identification standard (G1)	113	118.98	11.59	93-146
TOWRE-Sight Word raw (G1)	113	45.14	14.74	12-73
TOWRE-Sight Word standard (G1)	113	108.30	15.30	71-142
TOWRE-Phonemic Decoding raw (G1)	113	20.06	10.69	2-48
TOWRE-Phonemic Decoding standard (G1)	113	103.67	14.61	69-145
Vocabulary				
EVT-2 raw (G1)	113	97.86	13.78	58-143
EVT-2 standard (G1)	113	108.53	12.28	76-147
PPVT-4 raw (G1)	113	130.59	16.37	87-195
PPVT-4 standard (G1)	112	111.87	12.64	82-160
CELF-4 Word Classes Receptive raw (G1)	110	19.03	1.84	10-21
CELF-4 Word Classes Expressive raw (G1)	110	15.14	2.47	7-20
WJ-Auditory Memory raw (G1)	111	14.68	5.04	4-29
WJ-Auditory Memory standard (G1)	111	113.22	13.81	81-152
Comprehension Monitoring				
Comprehension monitoring (G1)	112	4.56	2.11	0-8
Comprehension monitoring (G2)	110	6.03	1.68	1-8
Comprehension monitoring (G3)	110	6.37	1.43	2-8
Reading Comprehension				
WRMT-Passage Comprehension raw (G3)	108	37.95	5.50	23-53
WRMT-Passage Comprehension standard (G3)	108	110.71	9.33	88-132
Gates-MacGinitie raw (G3)	109	35.22	8.42	13-47
Gates-MacGinitie standard (G3)	109	5.93	8.41	2-9
Reading Comprehension Measure (G3)	109	20.45	4.04	10-26

Note: WRMT = Woodcock Reading Mastery Test; EVT-2 = Expressive Vocabulary Test; PPVT-4 = Peabody Picture Vocabulary Test; G1 = Grade 1; G2 = Grade 2; Grade 3 = Grade 3.

Table 2

Pearson correlations among study variables (based on raw scores)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. WRMT-WA (G1)	1.00														
2. WRMT-WI (G1)	0.83	1.00													
3. TOWRE-SW (G1)	0.69	0.89	1.00												
4. TOWRE-PD (G1)	0.74	0.80	0.77	1.00											
5. EVT-2 (G1)	0.42	0.66	0.61	0.47	1.00										
6. PPVT-4 (G1)	0.31	0.50	0.41	0.31	0.78	1.00									
7. WC Rec (G1)	0.30	0.28	0.29	0.21	0.29	**0.18**	1.00								
8. WC Exp (G1)	0.40	0.44	0.40	0.34	0.45	0.33	0.68	1.00							
9. WJ-AM (G1)	0.41	0.45	0.43	0.38	0.33	0.22	0.29	0.38	1.00						
10. CM (G1)	0.24	0.38	0.37	0.25	0.50	0.48	**0.18**	0.29	0.22	1.00					
11. CM (G2)	**0.14**	0.27	0.26	0.21	0.36	0.25	0.30	0.24	0.20	0.33	1.00				
12. CM (G3)	0.21	0.32	0.31	0.28	0.44	0.32	**0.10**	0.28	0.20	0.42	0.20	1.00			
13. WRMT-PC (G3)	0.57	0.66	0.57	0.53	0.60	0.50	0.29	0.39	0.34	0.35	0.31	0.38	1.00		
14. Gates (G3)	0.49	0.66	0.64	0.54	0.58	0.45	0.29	0.43	0.39	0.43	0.39	0.46	0.69	1.00	
15. RCM (G3)	0.34	0.42	0.41	0.45	0.45	0.44	0.28	0.38	0.36	0.49	0.37	0.45	0.55	0.70	1.00

Note: WRMT = Woodcock Reading Mastery Test; WA = Word Attack; WI = Word Id; SW = Sight Word; PD = Phonemic Decoding; EVT-2 = Expressive Vocabulary Test; PPVT-4 = Peabody Picture Vocabulary Test; WC Rec = Word Classes Receptive; WC Exp = Word Classes Expressive; WJ-AM = Woodcock Johnson Auditory Memory; CM = Comprehension Monitoring; PC = Passage Comprehension; RCM = Reading Comprehension Measure; G1 = Grade 1; G2 = Grade 2; Grade 3 = Grade 3. All correlations, except for the bolded one, were significant at $\alpha = .05$.

Table 3
Latent growth models for comprehension monitoring for Grade 1 to Grade 3 (unstandardized coefficients reported; standard errors in parenthesis)

	Model 1 (linear)	Model 2 (quadratic)	Model 3 (quadratic and predictors)
Means			
Intercept	4.83*** (0.19)	4.56*** (0.20)	4.22*** (0.35)
Grade (linear)	0.81*** (0.09)	2.04*** (0.38)	2.05*** (0.38)
Grade (quadratic)	---	-0.57** (0.18)	-0.58** (0.18)
[a]Vocabulary (G1)	---	---	0.05*** (0.01)
[a]WJ-Auditory Memory (G1)	---	---	0.02 (0.02)
Variance Components			
Variation in intercept	0.83*** (0.20)	0.86*** (0.20)	0.33* (0.15)
[b]Variation in linear slope	---	---	---
[b]Variation in quadratic slope	---	---	---
Residual variance for G1	3.23*** (0.52)	3.11*** (0.47)	2.91*** (0.40)
Residual variance for G2	2.36*** (0.37)	2.14*** (0.37)	2.19*** (0.34)
Residual variance for G3	1.25*** (0.25)	1.25*** (0.24)	1.30*** (0.23)
Model Fit			
RMSEA	.20	.15	.081
90% CI for RMSEA	[.12,.30]	[.04, .28]	[.03,.13]
CFI	.55	.83	.95
SRMR	.19	.10	.09

****p<.0001, **p<.01,*p<.05;* [a]As predictors of the intercept. [b]Variation in linear slope or quadratic slope was not significant and had an inadmissible solution and thus was set to zero.
Vocabulary = latent construct defined using the Expressive Vocabulary Test (EVT-2), the Peabody Picture Vocabulary Test (PPVT-4), the CELF-4 Word Classes (receptive and expressive subtests); WJ = Woodcock Johnson (observed indicator); RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; SRMR = Standardized Root Mean Square Residual.

Table 4

Prediction of Grade-Three Reading Comprehension (standardized coefficients reported; standard errors in parenthesis)

	Model 1	Model 2	Model 3
Decoding (G1)	.75*** (.05)	.44**** (.11)	.43*** (.09)
Vocabulary (G1)	---	.35** (.11)	.15 (.10)
WJ-Auditory Memory (G1)	---	.16* (.07)	.13$^{¥}$ (.07)
Intercept from CM Growth Model	---	---	.36*** (.07)
Model Fit			
RMSEA	0.15	0.10	.09
90% CI for RMSEA	[.10,.20]	[.07, .13]	[.07, .12]
CFI	0.95	0.95	0.95
SRMR	0.04	0.07	0.07
R^2	0.56	0.66	0.74

$^{***}p < .0001$, $^{**}p < .01$, $^{*}p < .05$, $^{¥}p < .10$

Decoding = latent construct defined using WRMT – Word Attack, WRMT – Word ID, TOWRE-sight word, TOWRE-phonemic decoding; Vocabulary = latent construct defined using the Expressive Vocabulary Test (EVT-2), the Peabody Picture Vocabulary Test (PPVT-4), the CELF-4 Word Classes (receptive and expressive subtests); WJ = Woodcock Johnson (observed indicator); WRMT = Woodcock Reading Mastery Test; CM = Comprehension Monitoring; RMSEA: Root Mean Square Error of Approximation; CFI: Comparative Fit Index; SRMR: Standardized Root Mean Square Residual; CI = Confidence Interval.

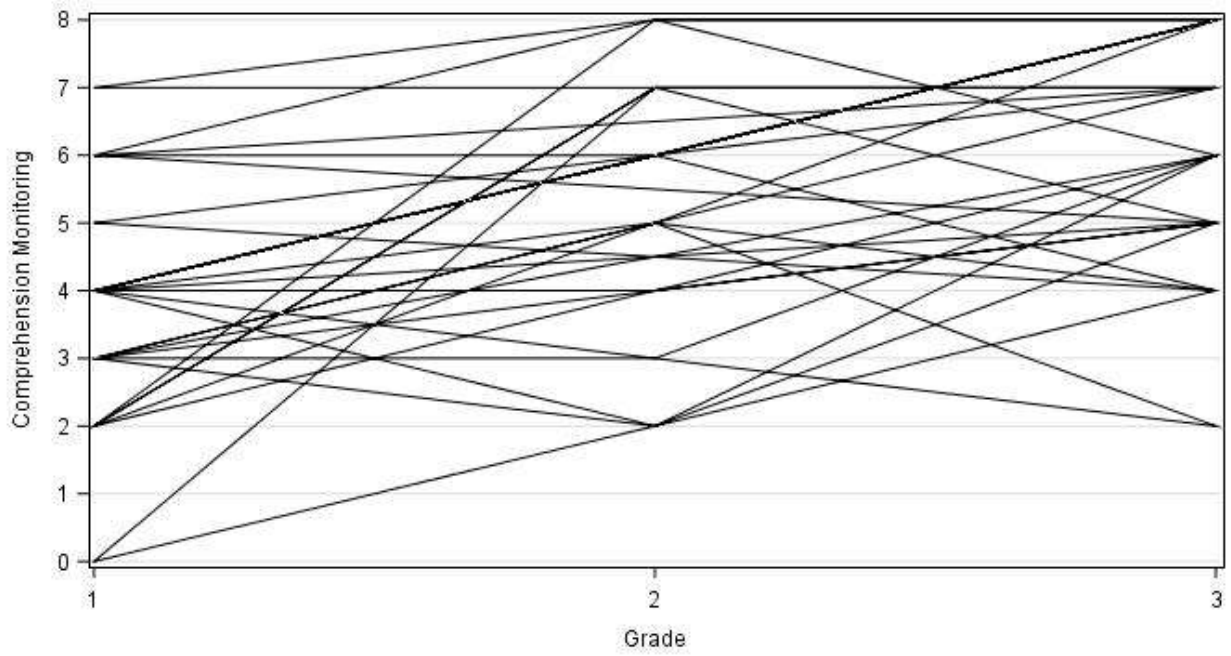

Figure 1. Spaghetti plot of comprehension monitoring for 15 randomly selected children.

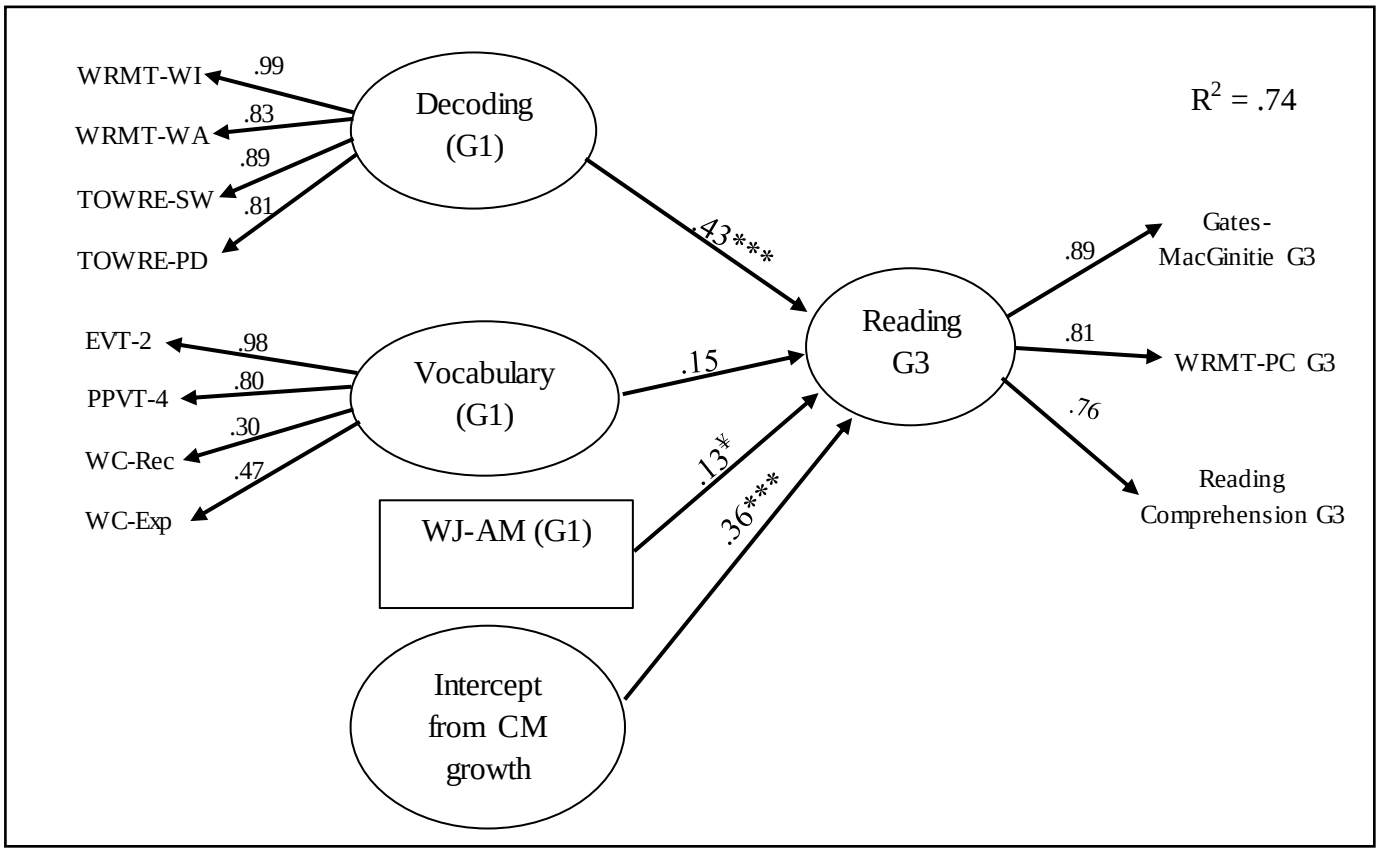

Figure 2. Results from the Structural Equation Model (standardized coefficients). ***p < .0001, **p < .01, *p< .05, ¥ p< .10. WRMT-WI = Word Id; WRMT-WA = Word Attack; TOWRE-SW = sight word; TOWRE-PD = phonemic decoding; EVT-2 = Expressive Vocabulary Test (EVT-2); PPVT-4 = Peabody Picture Vocabulary Test; WC-Rec = Word Classes Receptive; WC-Exp = Word Classes Expressive; WJ-AM = Woodcock Johnson Auditory Memory; WRMT = Woodcock Reading Mastery Test; CM = Comprehension Monitoring. Correlations among latent predictors are not shown in diagram but were included in the model. Correlations between decoding and (a) vocabulary, (b) WJ-AM, and (c) intercept from CM growth model were .68, .45, and .44, respectively. Correlations between vocabulary and (a) WJ-AM and (b) intercept from CM growth model were .34 and .60, respectively.