The Dissertation Committee for Deborah Kay Reed

certifies that this is the approved version of the following dissertation:

The Contribution of Retell to the

Identification of Struggling Adolescent Readers

Committee:

____
Sharon Vaughn, Supervisor

____
Diane P. Bryant

____
Herbert J. Rieth

____
Greg Roberts

____
Audrey Sorrells

The Contribution of Retell to the Identification of Struggling Adolescent Readers

by

Deborah Kay Reed, B.A.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2010

DEDICATION

This dissertation is dedicated to the many people who helped me develop my skills and grow both academically and professionally; I sincerely appreciated all opportunities to learn from the best. I owe special thanks to my advisor, Sharon Vaughn, for compelling me to make the switch to the Special Education Department and, then, taking me on as a student and fledgling researcher. I am better for the experiences of working with and studying under you, and I will continue to strive for the high standards you set.

I am also grateful for the contributions of my committee members. Greg Roberts believed I was capable of doing a CFA study, and Audrey Sorrells made sure I did not lose touch with reality while trying. Diane Bryant and Herbert Rieth made time in very busy schedules to provide me equally strong guidance and suggestions. Although not on my committee, Yaacov Petscher's patient assistance with learning the analysis was invaluable. I thank you all, as well as the others not named here who kept me focused on the end goal.

Finally, and most importantly, I dedicate this to my dear husband who supported us all by defending our nation in two wars while I worked on my doctorate. Your sacrifices have been far greater than any I have made over the course of my program, yet you supported and encouraged me as though the milestones were just as significant. I carry your heart with me, Patrick. I carry it in my heart. "And whatever is done by only me, is your doing, my darling" (ee cummings).

ACKNOWLEDGEMENTS

The author wishes to gratefully acknowledge the Texas Education Agency for its support of the research reported in this dissertation. In addition, the author acknowledges the National Institute for Child Health and Human Development for its support of the Learning Disability Center at University of Houston and The University of Texas at Austin, which resulted in the extant database used for the current study. The opinions expressed do not necessarily reflect those of the funding institutions, and no official endorsement should be inferred.

The Contribution of Retell to the Identification of Struggling Adolescent Readers

Publication No. ___________________________

Deborah Kay Reed, Ph.D.

The University of Texas at Austin, 2010

Supervisor: Sharon Vaughn

This measurement study examined the construct validity of the retell component of the Texas Middle School Fluency Assessment (Texas Education Agency, University of Houston, & The University of Texas System, 2008a) within a confirmatory factor analysis framework. The role of retell, provided after a one-minute oral reading fluency measure, was investigated by comparing the fit of a three-factor model of reading competence to the data collected on a diverse sample of seventh- and eighth-grade students (N=394). The final model demonstrated adequate to mediocre fit ($\chi^2 = 97.316$ {32}; CFI = 0.958; TLI = 0.941; RMSEA = .081). Results suggest that retell was a significant contributor to comprehension ($\Delta\chi^2$=16.652{1}, p < .001), fluency ($\Delta\chi^2$=10.882{1}, p = .001), and word identification ($\Delta\chi^2$=7.84{1}, p = .005). However, the χ^2 difference was greater for comprehension, as was the factor loading for comprehension (.250, p < .001) compared to fluency (.194, p < .001) and word identification .167, p < .001). Retell did, however, have a large residual variance (.938), suggesting it did not function well as a measure of comprehension in its current state with low inter-rater reliability ($K = .37$).

Narrative retell scores (.352, p< .001) were better predictors of comprehension than expository retell scores (from .2221 to .264, p < .001) or the combination of all three scores ($\Delta\chi^2$=134.261{19}; p < .001), but average retell scores produced a more parsimonious model than narrative retell scores alone (ΔAIC = 58.275; ΔBIC = 58.275). Average retell was only weakly correlated to other measures of comprehension (from r = .155 to r = .257, p < .01). However, the relationship was stronger than the relationship between retell and other measures of fluency (from r = .158 to r = .183, p < .01) or word identification (r = .132, p < .05). In addition, retell did not demonstrate differential item functioning when student characteristics (e.g., primary language, socioeconomic status, ability level) were entered as covariates, even though there were overall latent differences.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Over the past decade, many advances have been made in the identification and progress monitoring of young children with reading difficulties. The use of easily and frequently administered assessments is viewed as critical to planning effective instruction and preventing reading failure (Coyne, Kame-enui, Simmons, & Harn, 2004; Stecker & Fuchs, 2000). Despite the success of systematic intervention plans at the lower elementary level (Kamps & Greenwood, 2005; Simmons et al., 2007), many students demonstrate persistent reading difficulties with a low response to targeted instruction (Torgesen et al., 2001; Vellutino et al., 1996). Purportedly, at least 8 million students in grades 4 through 12 struggle with reading (Williamson, 2006).

The middle school years are often seen as a critical time for equipping students to be successful in post-secondary settings (ACT, 2008; Balfanz & Herzog, 2005; Dynarski et al., 2008). Approximately 27% of eighth-graders scored below the "basic" level on the National Assessment of Educational Progress ([NAEP]; Lee, Grigg, & Donahue, 2007) where "basic" is defined as the ability to identify the main topic of a passage, recognize explicitly stated supporting details, and make simple inferences (US Department of Education, 2006). To prevent continued reading failure and improve the educational attainment of adolescents, educators in the middle grades are attempting to apply the kinds of intervention practices that have been successful in early elementary. However, much less is known about effective, ongoing formative assessments for identifying the specific needs of adolescents with reading difficulties.

A common approach to diagnosing problems and monitoring the progress of students in grades 1 through 5 involves the use of oral reading fluency (ORF) measures. ORF is determined by calculating students' rate and accuracy as they read short passages aloud, usually for one minute. There are both formal ORF instruments, such as the *Dynamic Indicators of Basic Early Literacy Skills* ([DIBELS]; Good & Kaminski, 2002a), and informal or curriculum-based measures ([CBM]; Deno, 1986; Fuchs & Fuchs, 1984), the latter of which rely upon teacher-selected passages written on the student's current grade-level. Research on ORF measures has consistently produced high correlations between elementary students' rate and accuracy and their scores on standardized measures of reading (Burke & Hagan-Burke, 2007; Jenkins & Jewell, 1993; Spear-Swerling, 2006) as well as state criterion-referenced reading assessments (Stage & Jacobsen, 2001; Wiley & Deno, 2005). It is, therefore, theorized that ORF is indicative of students' general reading ability (Burke, Hagan-Burke, Kwok, & Parker, 2009; Fuchs, Fuchs, Hosp, & Jenkins, 2001; Shinn, Good, Knutson, Tilly, & Collins, 1992), and those who do not read quickly and accurately are the children who would profit from instructional intervention (Madelaine & Wheldall, 2005).

Although the data support ORF as a diagnostic and progress monitoring instrument for elementary students, many educators have expressed concern over the emphasis placed upon reading fluency and the use of words read correctly in a minute as a gauge of text comprehension (Applegate, Applegate, & Modla, 2009; Goodman, 2006; Shinn et al., 1992). This is often considered an issue of social validity (Fuchs, Fuchs, & Maxwell, 1988), but there are reasons to question the construct validity of ORF measures for adolescents. Researchers have found that assessments of overall comprehension do not measure equivalent cognitive processes (Cutting & Scarborough, 2006; Keenan, Betjemann, & Olson, 2008; Spooner, Baddeley, & Gathercole, 2004), particularly if they differentially employ narrative and expository texts (Best, Floyd, & McNamara, 2008). Hence, an instrument designed to measure only one type of ability (e.g., word identification or vocabulary knowledge) might fail to identify those students whose reading difficulty rests largely in another domain. To better understand the domains considered to comprise *reading competence* for adolescents (Catts, Adlof, & Weismer, 2006; Fletcher, Lyon, Fuchs, & Barnes, 2007), it is necessary to operationally define *word identification*, *fluency*, and *comprehension*.

Operational Definitions of the Constructs

Word identification. This construct encompasses the word-level skills associated with phonological processing such as letter-sound correspondences, the blending of sounds, knowledge of syllable patterns and morphemic structure. Word identification is demonstrated by the accurate identification of real words and/or the correct pronunciation of nonsense words (patterns of letters used to represent phonetically regular sounds).

Fluency. This construct is concerned not only with the accuracy of identifying printed words, but also with the speed in which those words are read. It rests upon verbal efficiency theory (Perfetti, 1985) that conceptualizes reading as being constrained by limited processing capacity. Fluent reading utilizes fewer cognitive resources for recognizing words or producing letter-sound correspondences because those basic skills are happening with automaticity. Fluency is demonstrated by the number of words read correctly in a limited time and is usually expressed as a reading rate.

Comprehension. This construct is defined simply as making meaning from text. It involves understanding what is literally stated in a passage as well as what must be inferred by making connections between passage content and prior knowledge or experiences. Comprehension can be demonstrated by expressing the main idea or gist of the passage, summarizing the ideas, making predictions about content not yet read, identifying the structure or organizational pattern of the ideas presented, recognizing the author's purpose and tone, recalling word meanings as used in context, and/or by drawing conclusions based on the information (Spearritt, 1972).

Significance of the Problem: Assessing the Reading Competence of Adolescents

Given the distinction among the three domains of ability, it is perhaps no surprise that a synthesis of fluency interventions found that improvements in adolescents' reading rate did not necessarily result in concomitant improvements in comprehension (Wexler, Vaughn, Edmonds, & Reutebuch, 2008). In contrast to the findings from grades 1 through 5, studies conducted with older students indicate a less robust correlation between ORF and reading comprehension (Schatschneider et al., 2004; Wiley & Deno, 2005). In addition, rate and accuracy scores have shown a tendency to asymptote in the middle grades (Fuchs et al., 2001; Stage & Jacobsen, 2001). Possible explanations for this are that the contribution of decoding to reading comprehension diminishes somewhat in adolescence, (Gough, Hoover, & Peterson, 1996; Keenan, Betjamann, & Olson, 2008), and/or older students have more highly developed compensatory strategies that lessen their reliance on word identification skills (Savage, 2006).

Supplementing ORF with a retell prompt might assist in identifying students who are reading dysfluently but with adequate understanding or, conversely, those who are reading fluently but with poor comprehension (Marcotte & Hinze, 2009; Roberts, Good, & Corcoran, 2005). Retell, or free recall, is a frequent component of reading comprehension measures (Fuchs et al., 1988; Nilsson, 2008; Talbott, Lloyd, & Tankersley, 1994). In comprehension research, the skills of retelling, recalling, summarizing, and paraphrasing are considered distinct skills that require differing levels of complex thought and different degrees of telling or transforming knowledge (Kintsch & van Dijk, 1978; Scardimalia & Bereiter, 1987). Within studies examining retell as a measurement tool, however, these skills are treated almost interchangeably (Duffelmeyer & Duffelmeyer, 1987). Depending upon the instrument or study, "retell" and "recall" could be used to elicit main ideas, summaries of the content, or a thorough restatement of the passage. In the most common approach, students are asked to read a passage, either silently or orally, and are then prompted to tell or write about the passage in their own words without referring back to the text.

Retell is an appealing compliment to ORF because it does not add considerable time or expense to the assessment, and it can present a consistent probe of comprehension across passages that is reflective of typical classroom instruction (Roberts et al., 2005). In a study with students ranging in age from 8 to 18, the retell task of an informal assessment was much less sensitive to decoding ability (as measured by word recognition of isolated real words and word attack of nonsense words) than other standardized measures of comprehension (Keenan et al., 2008). In addition, errors consistent with the meaning of the sentence or passage were more strongly related to fourth-graders' recall of important ideas than their reading accuracy scores (Kucer, 2009). Unfortunately, the most commonly available reading assessments with a retell task have not sufficiently established the technical adequacy of the retell component (Reed & Vaughn, manuscript under review).

Statement of Purpose

The purpose of this study is to examine the validity of the retell task included in the Texas Middle School Fluency Assessment ([TMSFA]; Texas Education Agency, University of Houston, & The University of Texas System, 2008a) within a confirmatory factor analysis (CFA) framework (Brown, 2006; Byrne, 1988; Marsh & Bailey, 1991; Thompson, 2004).

CHAPTER 2

Review of Literature

This chapter provides a critical analysis of the literature on retell measures. A more comprehensive synthesis of retell studies and measures is provided in Appendix A, but the following sections will address the important issues that pertain to this study. The first part of the review focuses on the results of research on retell to provide a framework for understanding: (a) the correlation of retell to other reading assessments, (b) predicting and monitoring student progress in reading comprehension, (c) inter-rater reliability, (d) measurement artifacts, and (e) ability differences among student participants. The second part of the review focuses on the evidence establishing the reliability and validity of commercially or publicly available retell measures.

Extant Research on Retell: Critical Analysis

Correlations of Retell to Other Reading Assessments. Studies examining the correlation of retell scores to other measures of reading have demonstrated a rather consistently moderate correlation between recall and assessments of overall reading including letter-word identification, academic knowledge, vocabulary, reading comprehension, and fluency. The strength of the correlations discussed in this section will be judged conservatively using the following scale of absolute correlation coefficient values (Williams, 1968):

0.00 – 0.30: weak; almost negligible relationship

0.30 – 0.70: moderate correlation; substantial relationship

0.70 – 1.00: high/strong correlation; marked to perfect relationship

The more conservative estimations of the strength of correlation were used here because the study was formative. A more stringent parameter would increase the confidence that the data represents reliable findings.

With a large sample of first-grade participants (Riedel, 2007), oral retell results were more moderately correlated (from *r* = .39 to *r* = .69) to the vocabulary and comprehension subtests of two standardized measures of reading, GRADE and TerraNova. A study of third-graders' comprehension of narrative versus expository text comprehension revealed that free and cued oral recalls of both narrative and expository text were moderately correlated (from *r* = .36 to *r* = .58) with the Woodcock-Johnson academic knowledge test (Best et al., 2008). Narrative free and cued oral recall, as well as expository free oral recall, were also moderately correlated with the Woodcock-Johnson letter-word identification test (from *r* = .48 to *r* = .64).

One exception to the pattern of correlations was found in a study of third- and fifth-graders where oral retell was not significantly correlated with researcher-developed measures of phrasing ability (Rasinski, 1990). Retell was, however, moderately correlated with both miscue and reading rate (from *r* = .38 to *r* = .52). It should be noted that this is the only study for which the retell scoring procedure could not be determined, so the basis of the correlation calculation is unknown. For all other studies reporting correlation data, retells were scored by a numerical count of the words or pre-determined idea units/propositions the student included (see section on inter-rater reliability for more information).

Stronger correlations between retell and fluency were found by Fuchs et al. (1988) with slightly older students. Retell scores of fourth- through eighth-graders were highly correlated to an ORF measure (mean *r* = .75) and moderately to highly correlated with the Stanford Achievement Test (SAT-7) reading comprehension and word study subtests (from *r* = .47 to *r* = .82). This is one of only two studies identified in the extant literature that incorporated both oral and written retells, so it is noteworthy that the researchers found consistently and significantly higher correlations for the written recalls than those for oral recalls. Yet, ORF scores were more highly correlated with the SAT-7 than any of the other measures included in the study. Moreover, ORF had higher correlations with the SAT-7 reading comprehension subtest than the word study subtest.

In another study of upper-middle grades students (Carlisle ,1999), oral recall scores of the sixth- and eighth-grade participants were moderately to highly correlated to scores on researcher-developed sentence verification (from *r* = .50 to *r* = .74) and moderately correlated to science vocabulary (from *r* = .49 to *r* = .51) tests. Results were similar in a study of fifth- and sixth-grade students (Hansen, 1978). Spearman's rank correlation coefficient revealed the proportion of idea units recalled was moderately to highly correlated with performance on open-ended, factual comprehension questions (from ρ = . 46 to ρ = .77).

Finally, Loyd and Steele (1986) found weak to moderate correlations between eleventh- and twelfth-graders' written recall of idea units and SRA reading comprehension and language arts mechanics scores (from *r* = .28 to *r* = .56). Holistic coherence scores on those written retells were, however, all in the slight or weak range (from *r* = .11 to *r* = .39). In sum, across all grade levels and test types in the identified studies providing validity data, retell measures tended to be moderately correlated with both formal and informal assessments of reading ability. These findings included the results of students from a range of different backgrounds and ability levels.

Predicting and monitoring student progress in reading comprehension. Equally few studies have provided data on the predictive validity of retell measures or the adequacy of retell scores for tracking student progress over time. For first graders (Riedel, 2007; Roberts et al., 2005), results indicate that ORF scores are the best predictor of reading performance. Overall, adding oral retell scores only improved the predictive accuracy by 1% or less than ORF alone. For some students, however, retell performance was notably inconsistent with their ORF performance. It is important to note that in neither of these first grade studies was it possible to determine whether narrative or expository passages were used. The measures include both genres, but the particular selections used as stimuli in the research were not specified.

In a study comparing third-graders' oral recall of narrative and expository passages (Best et al., 2008), decoding skill was the strongest predictor of narrative recall, but background academic knowledge was the stronger predictor of expository recall. In addition, Shinn et al. (1992) found the residual variance of written retells for narrative passages to be so high (74%) that "they did not function well as measures of reading constructs for fifth-grade students" (p. 470). Because this factor analysis did not employ expository passages or oral retells, it is not possible to determine if the text genre or format of the retell would have produced a different model of reading. However, there was an apparent developmental difference in the factor structure. A one-factor model of narrative text reading was most parsimonious at grade 3, with ORF demonstrating the highest factor loading (.90). At grade 5, a two-factor model of narrative text reading was most parsimonious, and ORF no longer demonstrated the highest factor loading. In the two-factor model, ORF loaded on decoding, and written retell loaded on reading comprehension.

Only 2 studies were identified as exploring the consistency or stability of students' retells, which would indicate the adequacy of such measures for tracking student progress. Fuchs and Fuchs (1992) found that a written retell measure administered to fourth- and fifth-graders twice weekly over 15 weeks produced instable scores which, when graphed for monitoring purposes, produced small average slopes in relation to the average standard error of estimate. Therefore, the researchers concluded the retells (scored quantitatively) were difficult to use for interpreting students' growth in performance. It is not clear from the article whether students were provided particular instruction related to retell in between testing points. Nonetheless, in a study of fourth-graders, oral retell scores were inconsistent across the multiple baseline probes administered over a 26-week period of multiple strategy instruction related to retell (Mason, Snyder, Sukhram, & Kedem, 2006). The results of these studies reflect a narrow range of grade levels (4 – 5) and a limited number of participants (n = 47). In fact, no identified studies of retell measures for the purposes of predicting or monitoring progress were conducted with students above grade 5.

Inter-rater reliability. The element of the technical adequacy reported most often in the literature is the extent to which different raters reach the same conclusion on evaluating students' retell responses. The overall range of reported inter-rater reliabilities is 72% to 100% agreement. Higher agreements were noted for some written retells (Fuchs & Fuchs, 1992; Fuchs et al., 1988; Loyd & Steele, 1986; Marcotte & Hintze, 2009; Mason et al., 2006; van den Broek et al., 2001) and for scoring procedures that relied upon the number of pre-determined idea units, story structure elements, or propositions recalled in oral retells (Best et al., 2008; Gambrell, Koskinen, & Kapinus, 1991; Gambrell, Pfeiffer, & Wilson, 1985; Horowitz & Samuels, 1985; McGee, 1982; van den Broek et al., 2001; Wright & Newhoff, 2001; Zinar, 1990). Lower inter-rater reliabilities (generally below .90) were noted for scale scores of writing coherence (Loyd & Steele, 1986) or of the match between the composition's organizational structure and that of the text (Richgels, McGee, Lomax, & Sheard, 1987); holistic scores of orally recalled story elements (Gambrell & Jawitz, 1993; Pearman, 2008; Popplewell & Doty, 2001); and holistic scores of overall retell quality (Mason et al., 2006).

The most common method for scoring students' retells involved numerical counts of words, idea units, propositions, or story elements. Although the studies reviewed had some variation in the quantitative procedures, the particular method used does not seem to influence the retell results. Fuchs et al. (1988) found no significant differences among scoring by number of words, percent of content words matching original text, or percent of predetermined idea units. This consistency in results across quantitative scoring procedures is particularly noteworthy because Fuchs and colleagues (1988) employed both written and oral retells after both oral and silent reading. However, only narrative passages were administered, and the students were allowed 10 minutes to respond with repeated prompting if they paused for 30 seconds. This was a longer period that involved more cuing by the examiner than was reported in other studies of oral retell. Nevertheless, the inter-rater reliabilities were consistent with those reported across the studies identified.

What was not addressed in the studies was interpretation of the numerical counts. In some cases, the counts were converted into a proportion of idea units recalled (e.g., Best et al., 2008; Gambrell et al., 1991; Hansen, 1978; McGee, 1982; Richgels et al., 1987; van den Broek et al., 2001; Zinar, 1990). However, little guidance was provided for making conclusions about what a desirable percentage of recalled idea units might be, or what percentage might indicate comprehension difficulty. Hansen (1978) noted that even on-grade-level readers recalled only about one-third of the idea units. In comparing third- and fifth-grade students, McGee (1982) found on-level third-graders recalled, on average, less than 20% of the main ideas and less than 30% of the details. Whereas, average achieving fifth-graders recalled, on average, about 50% of the main ideas but less than 40% of the details. Fifth-grade students identified as below-level readers recalled about 30% of both main ideas and details.

In all studies with quantitative scoring techniques, inter-rater reliability was based on the count itself, not on a translation of the tally or proportion to categories of “better” or “weaker” reading comprehension skill. The extant literature revealed no studies examining teacher or student factors that might influence the scoring and/or interpretation of results. Therefore, it not possible to determine if variables unrelated to retell or comprehension ability accounted for any of the variance among raters.

Measurement artifacts. Several studies of retell have explored issues related to factors of the testing conditions that might influence student performance, such as the influence of text genre. Although children as young as first grade (Moss, 1997) were able to accurately and completely provide main ideas and details in informational trade books, retell information in the proper sequence, and summarize what was most important about what they read, it was reported that students' responses varied widely. When comparing recall of expository texts with that of narratives, Best and colleagues (2008) found that third grade students recalled significantly more pre-determined propositions in narratives (10 – 15 versus 4 – 7 in expository text). With neither genre did students include many inferences (1 – 3%).

Similarly, fifth-graders were more likely to include explicitly-stated causal information from expository texts than when the causal information was implicit (Zinar, 1990). Students in that study who were identified as having comprehension difficulties did not include any causal information in their free recalls, but they included comparable amounts of causal information as their higher ability counterparts when probed. As in the Best et al. (2008) study, having students freely recall information from the passage did not produce as much acquired information as when students were specifically cued to provide information, including inferences, they initially left out of their retell. Hence, the use of specific follow-up prompting influenced student performance in quantitative as well as qualitative ways, particularly for students otherwise considered to have difficulty with reading comprehension (Zinar, 1990).

It is also important to note that the causal relationships targeted in the Zinar (1990) study were reported by Richgels and colleagues (1987) to be least often known by students of all abilities. When probed on their awareness of four expository text structures (collection, comparison-contrast, causation, problem-solution) and recall of texts written in those structures, sixth-graders were most aware of and able to convey information from the comparison-contrast structure. Conversely, students were least aware of or able to produce compositions in the causation structure. The more aware students were of a text's structure, the more likely they were to understand and remember that text as reflected in their written recalls. Furthermore, students demonstrated better-organized recalls in response to passages they read than to structured discussions in which they participated without the aid of a written text or other guide.

The issue of delivery formats for content to be retold was examined more specifically in three studies utilizing only narrative stories. Doty and colleagues (2001) compared second-grade students' retell performance when reading from an electronic medium versus a traditional print book. Research with a small sample of students found no significant differences in students' oral retellings of print versus electronically-based stories. Pearman (2008) found similar results with second-graders. However, when students were separated by reading ability (high-, medium-, low-proficiency), low reading proficiency students' mean retelling scores were significantly higher on electronically-based stories where students could access other supports such as labels, vocabulary definitions, and pronunciations. Changing the delivery format by adding a melody line, so that stories are sung rather than spoken, did not show more promise than the electronic formats. Kinder- and first-grade students demonstrated no significant differences in retell, reading comprehension questions, or mean length of utterance when stories were sung or spoken to them (Kouri & Telander, 2008). Students included a greater number of different words (a higher type-token ratio) when retelling sung stories, but they had greater attention and on-task behaviors when listening to spoken stories.

Identified studies that explored the influence of instruction in or practice with retelling had somewhat conflicting results. Second-grade students classified as high-, medium-, and low-proficiency readers all demonstrated no significant difference between mean scores on a first- versus second- administration of an oral retell measure (Pearman, 2008). However, second-grade students, who were accustomed to providing retells when conferencing with their teachers about the stories they are reading, performed significantly better on a retell assessment than students who did not practice retelling as part of their literacy instruction (Popplewell & Doty, 2001). Fourth-grade students provided multiple strategy instruction in elements of oral and written retelling demonstrated some improvement in the number of main ideas included (Mason et al., 2006). Although, the improvement was not evident in all of the 9 participants, and those students who did show progress were still inconsistent in the number of main ideas they included. Similarly, fourth-grade students provided opportunities to practice identifying the important ideas and supporting details in passages performed significantly better on written and oral retell tasks than students who practiced illustrating the important ideas

(Gambrell et al., 1985). Moreover, the students who practiced retelling had significantly higher free recall scores 2 days after the treatment as compared to the immediate free recall scores of the students who were in the comparison group and practiced illustrating.

Besides the age difference of the participants in the grade 2 and grade 4 studies, there was also a difference in the genre of text. The second-grade participants (Pearman, 2008) were reading narrative passages; whereas, the fourth-graders were reading expository (Zinar, 1990) or informational narrative (Gambrell et al., 1985) passages. In a separate study of grade 4 students (Gambrell et al., 1991), practice effects were also evident in students' oral retells of narrative stories, as well as their ability to answer cued-recall questions. Therefore, the data seem to indicate that the inconsistency in results might be attributable to developmental differences more so than text type. Unfortunately, this cannot be concluded with confidence because none of the available studies examined practice effects at different grade levels with both narrative and expository passages.

Developmental trends were also noted in a study of the effects of causal relation questions on students' written recall performance (van den Broek et al., 2001). When comparing the performance of fourth-graders, seventh-graders, tenth-graders, and undergraduate college students, younger students tended to recall less information overall than did older students. In addition, the school-age students generally recalled significantly less information when provided questions during and after reading, with the youngest students showing the most severe impairment in recall with questions used during reading. In contrast, the college students benefited from the inclusion of causal relation questions and recalled significantly more information when provided the questions during reading. Students of all ages included in their recall of what they read significantly more story propositions that were also needed to answer the questions. The researchers concluded that memory of and attention to information was universally heightened by the nature of the questions asked during or after reading. Students in grade 10 and students in college recalled similar amounts of information not specifically probed in the causal relation questions as did students who were not provided any questions. Students in grades 4 and 7 recalled significantly less information not specifically probed in the questions than students in the comparison. Hence, it seems students' sensitivity to potential measurement artifacts varies with age or developmental level. It cannot be determined from available data whether students' cultural-linguistic backgrounds are related to any variations in retell performance.

Ability differences among student participants. Ability has been addressed as an interaction variable in several studies of retell measures, 3 of which reportedly included high percentages of culturally and economically diverse students. The youngest participants ([grade 2]; Pearman, 2008) were categorized as having high-, medium-, and low-reading proficiency and were assessed with a retell protocol after reading traditional print and electronically-based stories. Although there were no differences in retell performance on the two text formats between students classified as high- and medium-proficiency, students with low-reading proficiency performed significantly better on the retell measure when reading electronically-based stories with hyper-textual supports in the form of labels, vocabulary definitions, and pronunciations of words or segments of text.

When only reading traditional print narratives, fourth-graders classified as proficient- and less-proficient readers made similar improvements in their abilities to answer cued-recall questions and to recall text-based propositions, themes, and plot episodes after four testing sessions (Gambrell et al., 1991). However, only the proficient-readers included significantly more appropriate elaborations with practice.

A comparison of the retell performance of students in grades 5-6 with and without LD (Hansen, 1978) found students with LD included significantly fewer idea units. Both groups accurately retold just over one-third of the total propositions when reading instructional-level material, had similar amounts of "other" information, and included few inaccuracies (mostly isolated, specific details). Students without LD had more partially-correct propositions and recalled significantly more super-ordinate propositions. However, both groups included similar amounts of subordinate details.

Similarly, Zinar (1990) found that fifth-graders with higher comprehension ability freely recalled significantly more pre-determined propositions than students identified as having low comprehension. In addition, high comprehenders were more likely to include explicitly-stated causal information; whereas, low comprehenders did not include any causal relationships unless probed. Then, low comprehenders included similar amounts of causal information and similar amount of pre-determined propositions as the high comprehenders. Low comprehenders seemed to understand the expository passages just as well as the students considered to have better reading ability, but the former students did not offer as much information unless specifically probed. They did not offer any more non-target information than the high comprehenders, rather the low comprehenders just did not say as much.

This consideration of target/significant and non-target/less significant information from the passage was explored further in Carlisle's (1999) study, which scored students' retells not only by the number of words and idea units included, but also by the importance or centrality of the ideas. Even after controlling for students' scores on researcher-developed sentence verification and science vocabulary tests, sixth- and eighth- grade students with learning disabilities (LD) still performed more poorly on recall than their peers without LD. Both ability groups included similar numbers of ideas and total words. However, the students without LD had better constructed and elaborated oral recalls of the expository passage. Among the better readers, a significantly greater proportion of their overall scores were attributable to main ideas, as opposed to the subordinate details. The follow-up prompting in this study was not specific to the missing information as was the case in the Zinar (1990) study with fifth-graders, so it is not possible to determine if these results confirm or contrast with the earlier study.

These results are consistent with a comparison of fifth-grade on-level, fifth-grade below-level, and third-grade on-level readers when providing retells for an expository passage written on the third-grade level (McGee, 1982). Although there were no significant differences among the groups on the number of subordinate ideas recalled, the better fifth-grade readers included a greater proportion and more total ideas than their peers reading below grade-level. Below-level fifth-graders recalled a greater proportion and more total ideas than third-grade on-level readers. As in the Zinar (1990) study, McGee (1982) found that students' sensitivity to the organizational structure of information in the text was related to their retell performance. Fifth-grade better readers were more likely to match the organization of their response to the structure of the passage read and include more super-ordinate ideas. Fifth-grade below-level readers demonstrated only a partial match to the structure of the text and included similar amounts of super- and sub-ordinate ideas in their recalls. Third-grade on-level readers, however, responded in list-like fashion with no match to the text's structure and included a greater proportion of subordinate ideas. McGee speculated that the differences in performance could be related to the degree of difficulty the expository text presented to students. Fifth-grade better readers not only found the text (written on a third-grade level) easier, but were also more likely to have the requisite background knowledge and experience with expository text.

Similarly, Horowitz and Samuels (1985) examined the recalls of sixth-grade students classified as "poor" and "better" readers when listening to and reading expository passages. Retells were scored with respect to the number of idea units and the rank of those ideas in the text hierarchy. The results did not differentiate between lower- and higher-order information, and follow-up prompting was not specific to missing information. Overall, poor readers performed better when listening to text, and better readers demonstrated significantly higher recall than their lower ability counterparts when reading text. When retell results were disaggregated by the level of text difficulty, both better and poor readers performed better when listening to easier texts. However, the two ability groups had no significant within group difference between listening and reading recall with more difficult texts.

In contrast, Wright and Newhoff (2001) did not report significant differences among the retell performance of students in grades 3-7 with and without language-learning disabilities (LLD) when reading or listening to narrative stories with a difficulty level that does not exceed the students' oral vocabulary or identified reading level. However, students with and without LLD did perform significantly better on inferential comprehension questions when the stories were read to them. In comparing the retell performance of students with LLD, those without LLD matched by chronological age, and those without LLD matched by language ability, the chronological-age-matched group produced more sentences, more verbatim information, and retold significantly more story grammar parts than the other two groups. There were no significant differences between the retell performance of students in the LLD and language-ability-matched groups. The researchers noted that age-matched students generally provided a longer retell, thus giving themselves more opportunity to include story components. As there was no follow-up prompting described for the retell portion, it is possible that students in the other groups might have provided more story components had they been specifically prompted as in the Zinar (1990) study.

Across the identified studies, students who are considered to be struggling with reading performed more poorly than average achieving or better readers when the retell protocol was administered in a more traditional format (i.e., with print-based passages read independently by the student and assessed with a generic recall prompt). Because the former students have previously exhibited difficulties, it is, perhaps, not surprising that they would perform better on a retell comprehension measure when they receive some assistance with reading the passages – either through electronic hypertext or from the teacher reading the passage aloud. The more compelling data suggest that these younger and middle grades students may not retell as much as they actually do comprehend unless they are specifically cued to provide missing information. However, they still do not provide the degree of elaboration or strength of retell construction exhibited by better readers.

Existing Retell Measures

Existing assessments that include a retell measure were identified in an ancestral search of articles on reading comprehension assessments. In addition, the databases of test publishers (e.g., ProEd, Pearson, McGraw Hill, Kendall Hunt) were manually searched for Informal Reading Inventories (IRIs), which the extant literature indicated were the most common type of comprehension assessment to include a retell component. The 12 instruments included in this review are designed for students in kindergarten through twelfth-grade, include a stated protocol for administering an oral or written retell, and are commercially or publicly available in all states. Assessments tied to commercial reading programs (e.g., *Houghton Mifflin Leveled Reading Passages Assessment* in the Houghton Mifflin reading series) were excluded unless the measure had been used in a study of retell. Instruments tied to commercial reading program were otherwise excluded because those examined tended to be reliable and valid only within the context of that program. The goal of this review was to describe the measures that could indicate students' reading ability irrespective of the instructional program in use.

Norming sample characteristics. Although 8 retell measures reported at least some information on the norming samples of students, only 1 had a large and diverse sample that represented the full span of grade levels for which the assessment was intended (Applegate, Quinn, & Applegate, 2008). A second measure reported a more limited sample of students identified in grade groupings (i.e., elementary, middle school, secondary, adult) for the reliability study, but did not utilize all grade levels for the validity study and did not report student ethnicities (Bader & Pearce, 2009). A third measure reported employing a diverse sample representative of all grades, but did not make it clear whether that sample was administered the optional retell subtest (Karlsen & Gardner, 1996). Similarly, a fourth measure had a large and diverse sample of all grade levels excluding the youngest (preK) and oldest (grade 9) for which the instrument is intended; however, the retell measure was not separated from the overall analysis of the assessment in the reliability study and reported no validity study (Cooter, Flynt, & Cooter, 2007).

The remaining 4 measures included only a single grade (Good & Kaminski, 2002b; Johns, 2008) or a small span of grades out of all those for which the assessment is intended (Beaver, 2003; Leslie & Caldwell, 2006). Among those 4 measures, one only conducted a reliability study (Johns, 2008) and another only reported the norming sample for the criterion validity study (Leslie & Caldwell, 2006). Bilingual students were reported in one measure's reliability study sample, but not the validity study sample

(Beaver, 2003; Beaver, 2006). Overall, few existing retell measures reported information about the norming sample demographics suitable for determining the generalizability of results across students of different ages and backgrounds.

Established reliability of existing retell measures. Authors and publishers of existing retell measures were more likely to report the inter-rater reliability of the instruments than any other type of established reliability (e.g., alternate form or test-retest reliability). Half of the instruments (n = 6) provide information on the agreement of different scorers. As was evident in the research on retells, higher inter-rater reliabilities were reported in 3 of the instruments that score retells on the number of pre-determined idea units a student includes in the recall ([.90 - .98+]; Applegate et al., 2008; Bader & Pearce, 2009; Leslie & Caldwell, 2006).

Only two measures that score retells holistically or with a more subjective scale provided inter-rater reliabilities (Beaver, 2003; Beaver, 2006; Johns, 2008). These were lower (.74-.81) as is consistent with what was reported in the research studies. A third measure utilizing holistic scores reported “some variation” in scoring but “great consistency” determining the overall reading level of students; however, the authors did not quantify the percent agreements among scorers to define their descriptors (Woods & Moe, 2007).

The second most common type of reliability reported among the existing measures was passage equivalency or alternate form reliability. Five measures provided data that ranged from a low of .57 (Good & Kaminski, 2002b) to a high of .90 (Leslie & Caldwell, 2006). Most of these measures (n = 4) include both narrative and expository passages, so the wide range in coefficients is not necessarily attributable to having different text genres in the assessment. However, the low passage equivalencies found in some instruments suggest that a possible measurement artifact exists in these assessments.

It could not be determined with confidence whether or not measurement artifacts existed due to a lack of corroborating evidence, such as on the instruments' test-retest reliability. Only 2 of the 12 instruments reported this data, and neither reported alternate form reliability (Beaver, 2003; Beaver, 2006; Cooter et al., 2007). Test-retest reliability ranged widely from .67 to .93 in the measure incorporating both narrative and expository text (Cooter et al., 2007) and were in the .90 range for the measure that primarily utilizes narrative stories (Beaver, 2003; Beaver, 2006).

Several measures reported other reliability information; although, some of the information was similar to that considered validity data by other test developers. For example, Johns (2008) reported moderate correlations between his instrument and two other commercially prepared reading inventories (from r = .64 to r = .73). Similarly, Leslie and Caldwell (2006) reported low to moderate correlations (from r = .34 to r =.60) between retell scores and comprehension question scores on the passages in their measure. The correlations had high variability, particularly at lower grade levels.

The remaining reliability data included an estimated reliability of 3 passages for the retell fluency measure (.80) based on the Spearman-Brown prophecy formula (Good & Kaminski, 2002b); the percent agreement (66%) on reading instructional level between the reading inventory and a clinician-constructed inventory (Johns, 2008); and the internal consistency of the overall reading comprehension portion of the instrument which included the retell protocol as an optional component ([from r = .79 to r = .97]; Karlsen & Gardner, 1996). Only one measure provided information to establish the reliability of the pre-determined idea units used to score students' retells (Leslie & Caldwell, 2006). The propositions deemed important were recalled by 20% of the students and/or were identified by 50% of the teachers in the field test. However, the norming sample was not described. Two measures reported no reliability data (Roe & Burns, 2007; Silvaroli & Wheelock, 2004). These same instruments provided no validity data either.

Established validity of existing retell measures.. Five of the 12 instruments reported no information on validity; however, 2 of those measures included correlation data in sections of the technical manuals labeled as "reliability" that was similar to what other measures reported in sections labeled "validity" (Johns, 2008; Leslie & Caldwell, 2006). These two measures provided correlation coefficients between the retell scores and other instruments or other test components as described in the previous section.

Four measures provided correlations among test components as validity data. Although the results were somewhat consistent in indicating moderate correlations, some measures lacked specific information or a broader sample that would increase the confidence in and generalizability of the data. A moderate correlation (r = .51) was reported between the retell score on the *Critical Reading Inventory* (Applegate et al., 2008) and the total comprehension score on narrative passages, but a less robust correlation (r = .43) was reported for informational passages. Leslie and Caldwell (2006) reported the retell component of the *Qualitative Reading Inventory* (*QRI-4*) was correlated with prior knowledge scores from kindergarten through upper middle school, but no coefficients were provided. In addition, the overall reading comprehension score was correlated with word identification and rate at preK, second-, third-, and fourth-grades, but no information on the complete norming sample and no coefficients were provided. With a limited sample of first-graders, the average retell fluency score on the *Vital Indicators of Progress* ([*VIP*]; Good & Kaminski, 2002b) was moderately correlated (*r* = .61) with the oral reading fluency average. Finally, the continuity of the *Stanford Reading Diagnostic Test* (*SDRT*) across grade levels was established with moderate to strong correlations between corresponding subtests (from *r* = .59 to *r* = .87), but the optional retell subtest was not disaggregated in the data.

Test developers often provided information on only one type of validity (e.g., concurrent, predictive, construct, or criterion validity), and rarely did two measures include data on the same type. The developers of the *SDRT* sought to establish the instrument's construct validity (how accurately the test measures the construct of reading and academic performance) by correlating results to scores on a standardized measure, the *Otis-Lennon School Ability Test*. In contrast, researchers of the *VIP* correlated results to scores on a standardized measure of general reading achievement, the *Broad Reading Cluster*, in order to establish the *VIP*'s predictive validity (how accurately the test represents students' future reading ability or performance). Despite the different purposes, results in neither validity study were highly encouraging. Correlations between the *SDRT* and the *Otis-Lennon* for a large sample of students in grades 2 through 12 were reported from a moderate .43 to a strong .95, a wide range without disaggregated data on the optional retell subtest. The correlation of the *VIP* with a limited sample of first-graders was a moderate .51, but the retell measure only explained an additional 1% of the variance in the *Broad Reading Cluster* results compared to the variance accounted for by ORF scores alone (Roberts et al., 2005).

The assessment labeled as "parallel" to the *VIP*, the *Dynamic Indicators of Basic Early Literacy* ([*DIBELS*]; Good & Kaminiski, 2002a), provided data on criterion-related validity. Consistent with the *VIP* data, the correlation between the *DIBELS* retell component and the *Oregon State Assessment Test* was a moderate .50. However, the test publishers did not directly report the norming sample or the percent variance explained by the *DIBELS* retell. In addition to predictive validity, information was provided on the measure's concurrent validity (how accurately the test represents the student's current level of reading ability or performance). The correlation between DIBELS and ORF scores was, again, reported as moderate (r = .59), with no information on the norming sample.

The developers of both the *Developmental Reading Assessment* ([*DRA*]; Beaver, 2003; Beaver, 2006) and the *QRI-4* provided results on the correlation of their measures to the *Iowa Test of Basic Skills* (*ITBS*). Data for the *QRI-4* were used to establish the instrument's criterion validity; whereas, the developer of the *DRA* did not specify what type of validity the data were to establish. As with the intra-correlations of test components reported earlier, results were similar but lacked specific information on the norming samples or were based on samples that did not reflect the full spectrum of grade levels for which the assessments are intended. The *DRA* was moderately correlated (from r = .68 to r = .83) with *ITBS* grade-equivalent scores and national curve equivalents as well as Lexile measures. However, only students in grades 1, 2, and 3 participated in the validation studies. Interestingly, the developers of the QRI-4 did not administer the ITBS to students in grades 1 through 3 but, instead, administered the *California Achievement Test* for these lower grade levels.

Correlations between the *QRI-4* and the *ITBS* (for grades 3-8) or the *California Achievement Test* (for grades 1-3) were reported in a wide range, with some non-significant findings and inconsistent results on narrative versus expository passages in the *QRI-4*. For narrative text, correlations ranged from a weak and non-significant .27 at grade 6 to a strong .85 at grade 1. For expository text, correlations ranged from a weak and non-significant .28 at grade 7 to a moderate .55 at grade 9. The norming sample was reported as including students in grades 1 through 8, so it is unclear how the results for the grade 9 students were obtained. The *QRI-4* is intended for use through high school. Test developers also reported a moderate correlation (r = .75) between the *QRI-4* and the *Woodcock Reading Mastery* passage comprehension subtest, but did not specify the type of validation study conducted or the norming sample on which the results were based.

The developer of the *DRA* took a unique approach to establishing the content validity (how well the test taps reading behaviors and skills that it is supposed to measure) of the instrument. Reportedly, 89% of the teachers at the test development site (n = 84) agreed that the measure was helpful in evaluating students' reading progress, and 82% agreed that the *DRA* was helpful in determining instructional goals. The only other instrument reporting similar data was the *BADER Reading and Language Inventory* (Bader & Pearce, 2009). Without specifying the type of validity they were attempting to establish, the test developers reported a high correlation between *BADER* scores to school reading specialists' judgments of students' reading level (r = .93) and to classroom teachers' judgments of students' reading levels (r = .89). These results were obtained with scores from limited samples of students in restricted grade levels.

Study Framework

Results from the review of research indicate that retell was moderately correlated with other measures of reading and had more variability at younger grade levels. Of note was the finding that no studies of retell as a progress monitoring tool were identified with students above grade 5 where retell performance shows more sensitivity to practice and less sensitivity to decoding ability. The review of existing retell instruments revealed very little data substantiate the reliability and validity of existing retell measures. Therefore, this dissertation study seeks to examine the contribution of retell to a theoretical model of reading for middle school students.

As a measurement study, rather than an intervention study, the framework derives from theories of how the construct(s) of reading are defined. It examines how performance on measured reading skills contributes to latent variables or theoretically defined components of reading. Extant research suggests various conceptions of *reading competence* as a single construct or as a composite of 2 to 4 distinct constructs (i.e., decoding, fluency, vocabulary, and comprehension). The number of component skills seems to depend on the age of the individual(s) and the operational definitions of the constructs.

The next section reviews the research basis for defining a model of reading competence in adolescence. The number of latent variables identified in data obtained from students at different grade-levels or ages are reported. In addition, the correlation among skills measured as related to a model of reading competence is provided. The section concludes by positioning the current study within the existing framework.

Component skills. Factor analyses conducted on the scores of younger students in the middle of first-grade (Burke & Hagan-Burke, 2007) and in third-grade (Shinn et al., 1992), indicate that measures of phoneme segmentation, word reading in isolation, nonsense word reading, oral reading fluency, retell, and comprehension all load onto a single factor. For these students, ORF performance had the highest factor loadings. In predicting the reading development of kindergarten students, phonological awareness alone was more closely associated with passage comprehension ability through second-grade (Kirby, Parrila, & Pfeiffer, 2003). Although the effect of low phonological awareness continued to be evident through grade 5, naming speed (as measured by rapid color and picture naming) became the powerful predictor of reading comprehension.

By grade 5, the results of studies suggest that two distinct constructs can be identified. A two-factor model that differentiated decoding (defined by measures of word reading, nonsense word reading, and oral reading fluency) from reading comprehension (defined by measures of multiple-choice questions and retell) was most parsimonious for fifth-grade reading competence (Shinn et al., 1992). Similarly, research conducted with fourth- and fifth-grade students distinguished those who suffered from comprehension deficits alone, word-level deficits alone (word reading, nonsense word reading, spelling, phonological awareness, and naming speed), and those with a combination of comprehension and word-level deficits (Leach, Scarborough, & Rescorla, 2003). These results are consistent with those found at grade 8 (Catts et al., 2006). However, the way in which *decoding* or *word-level* skill is defined could result in the identification of a third construct of reading competence.

When decoding accuracy (phonological processing as measured by accurate word and nonsense word reading) is considered separately from naming speed or text reading rate, researchers have categorized students from grade 5 through adulthood based on deficits in one more of the following domains: decoding, fluency, and comprehension (Buly & Valencia, 2003; Hock et al., 2009; Jackson, 2005; Valencia & Buly, 2004; Vukovic, Wilson, & Nash, 2004). Among adolescents, difficulties in comprehension and fluency account for greater percentages of students who struggle with reading than difficulties in word identification (Hock et al., 2009; Texas Education Agency, University of Houston, & The University of Texas System, 2008b; Valencia & Buly, 2004). In reviewing the research on the cognitive correlates of fluency, Fletcher and colleagues (2007) found support for the independence of naming speed/fluency and phonological awareness/decoding.

Although more recent research indicates acceptance of a three-factor model of reading, particularly for older students, there is little evidence that vocabulary knowledge exists as a fourth construct. Principal components analysis conducted with eighth- and ninth-grade participants identified decoding, fluency, vocabulary, and comprehension as distinct categories (Hock et al., 2009). However, the high correlations between vocabulary and comprehension make it difficult to consider the skills distinct (Carlo et al., 2004; Snow, 2002). Rather, the relationship is more accurately described as bi-directional (Wagner, Muse, & Tannenbaum, 2007). Where students with low comprehension can be differentiated from students who are low in word identification, vocabulary knowledge tends to be consistent with comprehension performance (Leach et al., 2003; Valencia & Buly, 2004). That is, students in the studies who demonstrated higher vocabulary knowledge also were likely to demonstrate stronger comprehension performance, and vice versa. Even Hock and colleagues (2009) identified few (approximately 4% of the sample) students scoring above the 40th percentile on standardized measures of reading comprehension who demonstrated low vocabulary skill.

A path analysis of five predictor variables found that vocabulary made a larger contribution to the reading comprehension of ninth-grade students than background knowledge, inference ability, strategy use, or a word reading accuracy and fluency composite (Cromley & Azevedo, 2007). In addition, vocabulary had a small effect on comprehension mediated by inference ability and was significantly correlated to both word reading and background knowledge, the latter of which made the second largest contribution to comprehension among the five predictor variables. The direct and inferential mediation model ([DIME]; Cromley & Azevedo, 2007) adds indirect pathways to the structural equation models popularized by Kintsch (1988) and Perfetti (1985) that also rely on the five predictor variables of vocabulary, background knowledge, inference ability, word reading, and strategy use. Cromley and Azevedo concluded that the role of inference ability in mediating the effects of vocabulary, background knowledge, and strategy use on comprehension distinguish literal comprehension performance, modeled by the direct pathways from the other four predictors, from inferential comprehension.

This dissertation study will expand on the work of Shinn and colleagues (1992) conducted with third- and fifth-grade students by modeling the latent constructs of *reading competence* for students in grades 7 and 8. Data will be analyzed to determine if findings that distinguish comprehension from word-level deficits (Catts et al., 2006) as well as decoding accuracy/word identification from reading rate (Fletcher et al., 2007) can be confirmed. This study is different from previous research that has sought to categorize middle school students who struggle with reading by particular skill deficits (Buly & Valencia, 2003; Hock et al., 2009; Valencia & Buly, 2004) because the final model will be based on data obtained from students at a range of ability levels, including those considered typically achieving in reading. Although indirect pathways to comprehension will not be examined, results will contribute to the field by providing empirical data on the relationship of retell as a previously unexamined variable in the construct of reading comprehension among adolescents (Cromley & Azevedo, 2007; Kintsch, 1988; Perfetti, 1985).

CHAPTER 3

Methodology

Overview of Research Design

The reliability and validity of the retell component of the TMSFA (Texas Education Agency, University of Houston, & The University of Texas System, 2008a) was examined. In measurement research, *validity* was traditionally divided into four different categories: predictive, concurrent, content, and construct (Cronbach & Meehl, 1955). More recently, however, construct validity has been considered to encompass the other forms of validity as a unified or overarching quality within which particular relationships among the test being developed and other established assessments are explored (Brown, 2006; Trochim & Donnelly, 2006). The goal of construct validity is to experimentally demonstrate that the new instrument measures the construct it intends to measure. The construct is some attribute or ability that has been established in theory and observed in practice. In this study, the construct of interest is reading comprehension in a 3-factor model of reading competence, which also includes the constructs of word identification and fluency.

With *a priori* constructs of reading, confirmatory factor analysis (CFA) is the most appropriate method of evaluating the construct validity of retell to determine whether it measures observable skills that predict reading comprehension ability (Brown, 2006; Byrne, 1988; Marsh & Bailey, 1991; Thompson, 2004). As described by Shinn et al. (1992):

> Confirmatory factor analysis tests whether the theoretically derived model is one of the models that would fit the data adequately. Thus, instead of relying on the subjective judgment that the theoretical model is adequately reflected by the empirical model as in exploratory factor analysis, the researcher can test explicitly the hypothesis that the theoretical model adequately fits the data. (p. 466)

Research Questions

Given that the research establishing a three-factor model included retell as an assessment for the construct of comprehension among adolescents (Burke & Hagan-Burke, 2007; Jackson, 2005; Shinn et al., 1992), it could be expected that the retell component of the TMSFA would measure reading comprehension ability. Similarly to the procedure used in two of the aforementioned studies (Burke & Burke, 2007; Shinn et al., 1992), the TMSFA retell protocol is administered after a student reads a passage under timed conditions. Based on the premise that immediate recall is one common element of reading comprehension measures, retell is intended to provide unique information on how well the student understood the passage at a literal level (Jackson, 2005; RAND Reading Study Group, 2002).

However, no commercially or publicly available retell instruments have established the construct validity (Reed & Vaughn, manuscript under review). Among 12 existing instruments, only one ([SDRT]; Karlsen & Gardner, 1996) specifically mentioned construct validity in the technical manual, but the correlation coefficients for the optional retell subtest were not disaggregated from that of the primary components of the reading comprehension assessment. Three other retell developers reported the correlation of their instruments to a state reading assessment (Good & Kaminski, 2002a) or to a standardized measure of reading achievement (Beaver, 2003; Beaver, 2006; Leslie & Caldwell, 2006). Yet, no technical manual for any identified instruments was found to report results of factor analyses conducted with retell data.

Therefore, the primary research questions addressed about the TMSFA retell were:

1. What is the factor structure of reading competence expressed in the data obtained from a large, heterogenous sample of students in grades 7 and 8?
2. Is model fit improved by including only narrative retell, only expository retell, narrative and expository retells entered individually, or the average retell performance on narrative and expository passages combined?
3. How and to what extent does retell contribute to comprehension? Does retell contribute to fluency and word identification?
4. What are the patterns of associations (correlation, regression) between the TMSFA retell instrument and other standardized measures of reading?
5. Is retell influenced by differences in primary language, ability level, or socioeconomic status over and above the effects of reading comprehension?

Research Setting and Participants

This study relies upon an extant database compiled by researchers at The University of Texas at Austin and the University of Houston under a grant from the National Institute of Child Health and Human Development (NICHD) and the Texas Education Agency (TEA). Participants were from 7 middle schools in Texas. In all, 394 students were tested: 149 from school A, 12 from school B, 30 from school C, 37 from school D, 61 from school E, 47 from school F, and 58 from school G. Of the 394 students, 260 were enrolled in grade 7, 134 were enrolled in grade 8, 184 were female, and 211 were male. The sample was culturally and ethnically diverse with approximately 37% African-American students, 1% Asian, 47% Hispanic, 14% Caucasian, and 63% classified as economically disadvantaged (based on free/reduced lunch status). Students represented a range of ability levels with 13% receiving special education services, 16% classified as limited English proficient or enrolled in English as a second language (ESL) classes, and 23% classified as having reading difficulties (based on scale scores on the state criterion-referenced reading assessment).

After removing outliers, the final sample consisted of 311 students, evenly divided between males and females. The racial/ethnic make-up did not change. The percentage of students in special education (12% of the sample) and the percentage of students classified as having reading difficulties (22% of the sample) were only slightly smaller than in the original sample. There were, however, a greater number of students classified as limited English proficient or enrolled in ESL classes (24% of the sample), and a greater number were classified as economically disadvantaged (71%).

Measures. All students were administered 11 reading assessments conceptualized as measuring word identification, fluency, and/or comprehension. With the addition of an intelligence test and the TMSFA retell (the instrument under study), the total number of measures included in this study was 13. Data from these assessments were gathered at post-test (May 2008) at the seven school sites. Each instrument is fully described in the following sections and examples of the TMSFA components are provided in Appendix B.

Word identification. Students were administered five measures of word identification and word attack. Three of these were subtests of the *Woodcock-Johnson III Tests of Achievement* ([WJ-III]; Woodcock, McGrew, & Mather, 2001): Word Attack, Letter-Word Identification, and Spelling. Word Attack is assessed by having students read aloud phonetically regular nonsense words. The median coefficient alpha reported for this subtest is .87, and the median test-retest reliability coefficient was .83 with a 1-year interval between test administrations (McGrew & Woodcock, 2001). Letter-Word Identification is assessed by having students name letters and read aloud lists of real words. The median coefficient alpha reported for this subtest is .94, and the median test-retest reliability alpha was .95, again with the 1-year interval between administrations (McGrew & Woodcock, 2001). Taken together, these two individually-administered subtests comprise the Basic Reading Skills Cluster of the WJ-III, which was moderately to highly correlated with the *Kaufman Test of Educational Achievement Reading Decoding Scale* (r = .66; Kaufman & Kaufman, 2004) and the *Wechsler Individual Achievement Test Basic Reading Scale* (r = .82; The Psychological Corporation, 1992).

Although spelling is the encoding of sounds rather than decoding, spelling ability is related to reading ability and reflects a student's understanding of word structure (Blachman, Tangel, Ball, Black, & McGraw, 1999; Cassar, Treiman, Moats, Pollo, & Kessler, 2005; Ehri, 2000). Therefore, the spelling subtest of the WJ-III was included as a measure of the decoding construct. It requires students to encode letters and words as they are dictated orally. In a modification from the typical individual administration, this data was gathered through group administration with a set list of items. The median coefficient alpha for this subtest was reported as .90 (McGrew & Woodcock, 2001).

Two other individually-administered tests were included for the decoding construct: the *Test of Word Reading Efficiency* ([TOWRE]; Torgesen, Wagner, & Rashotte, 1999) and the TMSFA Word Reading Fluency subtest. Both the Sight Word Efficiency (SWE) and the Phonemic Decoding Efficiency (PDE) subtests of the TOWRE were administered. The SWE assesses how many real words students can accurately identify in a 45-second time limit. As with the WJ-III Word Attack subtest, the PDE assesses how many phonetically regular nonsense words a student can identify with the time limit. The mean alternate forms reliability coefficients for the TOWRE all exceeded .90, and the test-retest reliability alpha ranged from .83 to .96 (Torgesen, Wagner, & Rashotte, 1999). Because the two subtests are highly correlated, the combined TOWRE Summary score was used for analysis.

The Word Reading Fluency subtest of the TMSFA assesses the number of real words a student can read accurately in 60 seconds. Students are presented 3 lists in succession, each of increasing difficulty as defined by the length and frequency of the words (Zeno, 1995). Substitutions, mispronunciations, alterations, reversals, skips, and 3-second hesitations are all counted as errors. The mean intercorrelation of performances on the three word lists ranged from .89 to .98 with a sample of students in grades 6 through 8 (Texas Education Agency, University of Houston, The University of Texas System, 2008b). The criterion validity of the Word Reading Fluency subtest (r = .36) was established with the Texas Assessment of Academic Skills (TAKS) reading test (Texas Education Agency, 2004).

Fluency. Students were administered four measures of reading fluency. One, the TOWRE (Torgesen, Wagner, & Rashotte, 1999) was also included as a measure of decoding because it assesses a student's ability to identify words. However, its timed nature results in a score reflective of reading rate, so it is also included as a measure of fluency. As previously mentioned, the mean alternate forms reliability coefficients for the TOWRE all exceeded .90, and the test-retest reliability alpha ranged from .83 to .96 (Torgesen, Wagner, & Rashotte, 1999).

Similarly, the Word Reading Fluency subtest of the TMSFA is included as a measure of both decoding and fluency because it assesses the number of words in isolation that students can read correctly in one minute. The other individually-administered subtest of the TMSFA, Passage Reading Fluency, utilizes connected text to assess the number of words read correctly. Students are presented three passages in succession, each of increasing difficulty or Lexile levels (The Lexile Framework, 2007). The three passages at each testing point represent a combination of narrative and expository text. Substitutions, mispronunciations, alterations, reversals, and skips are all counted as errors. If a student hesitates for 3 seconds, the examiner is to provide the word but mark it as an error. All passages were equated and the mean intercorrelation of the performances on five passages across testing points ranged from .86 to .98 with a sample of students in grades 6 through 8 (Texas Education Agency, University of Houston, The University of Texas System, 2008b). The criterion validity of the Passage Reading Fluency subtest (*r* =

.50) was established with the Texas Assessment of Academic Skills (TAKS) reading test (Texas Education Agency, 2004).

The *Test of Sentence Reading Efficiency* ([TOSRE]; Wagner, in press) is a group-administered measure that assesses students' ability to determine whether a statement is truthful or logically correct. For example, the sentence: "A fish lives on land," should be marked "false." Scores are based on the number of sentences marked correctly in 3 minutes, minus the number of sentences marked incorrectly. The mean intercorrelation of performances across five time points ranges from .79 to .96 with a sample of students in grades 6 through 8 (Wagner, in press).

Comprehension. Students were administered four measures of comprehension. The *AIMSweb Reading Maze* (Harcourt Assessment, 2007; Shinn & Shinn, 2002) is a group-administered measure utilizing short passages (150-400 words in length) with every seventh word after the first sentence deleted. In the word's place are three words inside parentheses. Scores are based on the number of words within parentheses selected by students to correctly complete the cloze for the passage. The intercorrelation of performances across testing points ranges from .69 to .91 with a mean of .81, and the reliability of estimated growth was .66 (Shinn, Deno, & Espin, 2000).

Two subtests of the *Group Reading Assessment and Diagnostic Evaluation* ([GRADE]; Williams, 2001) were administered, as the name implies, to groups of students. The Passage Comprehension subtest requires students to read a short passage (one or more paragraphs) silently and then respond to multiple-choice questions focused on questioning, predicting, clarifying, and summarizing. The Listening Comprehension subtest requires students to listen to a sentence read orally by the examiner and then decide which of four pictures best matches the sentence. Items are intended to target comprehension of vocabulary, grammar, idioms, inference, and non-literal expressions. Reliability coefficients for alternate form and test-retest were in the .90 range (Williams, 2001).

The WJ-III Passage Comprehension test is individually-administered to students by having them read aloud a sentence or short paragraph in which words have been removed. This subtest assesses students' ability to use their vocabulary knowledge and make inferences from context in order to correctly supply the missing word. The median coefficient alpha for this subtest was reported as .88 (McGrew & Woodcock, 2001).

The *Texas Assessment of Academic Skills* ([TAKS]; Texas Education Agency & Pearson Educational Measurement, 2007) is the criterion-referenced assessment used as the accountability test in Texas. Tests are unique to the grade level and are designed to measure student learning of the Texas Essential Knowledge and Skills. Internal consistency reliabilities are reportedly in the high .80s to low .90s range. Scale scores are equated using the Rash model, and the resulting classification accuracy ranges between 81.7% and 95.4% for the TAKS reading tests. Scale scores at the Met Standard performance level predicted ACT English scale scores of 18 and SAT English scale scores of 460 (Texas Education Agency & Pearson Educational Measurement, 2007).

Other measures. In addition to the twelve instruments selected to measure the *a priori* constructs, the Kaufman Brief Intelligence Test – 2 ([K-BIT-2]; Kaufman & Kaufman, 2004) was used to assist in determining whether ability level was a covariate of retell performance. The K-BIT is individually administered and includes items assessing verbal as well as nonverbal intelligence. For the Verbal Scale, the Verbal Knowledge subtest was used. This assesses expressive vocabulary, but does not require reading or spelling. The examiner reads aloud a question, and the student selects from among six illustrations the one that best corresponds to the question. For the Nonverbal Scale, the Matrices subtest was used. This assesses reasoning ability through the use of relationships and analogies. The items contain pictures or abstract words from which students select the one that corresponds to a series of other diagrams or completes a 2 x 2 analogy. Internal consistency values reportedly range between .87 and .95 for all subtests and the composite, and the test-retest reliabilities reportedly range between .80 and .95 (Kaufman & Kaufman, 2004). For the norming sample of students in grades 6 through 8, correlations with other assessments of intelligence ranged between .75 and .90 (Kaufman & Kaufman, 2004).

The final measure included in this study was the TMSFA retell. After the one-minute reading of each passage in the Passage Reading Fluency subtest, the examiner conceals the text and delivers the prompt: "Tell me in your own words what this passage is mostly about." If the student provides only the title or a single word, the examiner prompts again with "Tell me more." This additional prompt is offered only one time. The examiner transcribes the student response as accurately as possible on the record sheet and scores the response using a rubric. Scores from 0 – no response to 3 – strong comprehension are awarded based on accuracy, completeness, and coherency (rubric and exemplar responses are provided in Appendix C).

Procedures

Test administration. The assessments were administered at the school sites by research assistants who attended at least 6 hours of training prior to testing for the first time and a 3-hour “booster” training prior to testing in subsequent waves. All assistants had to achieve 100% accuracy in the administration and scoring procedures, which could take 2 to 4 hours longer than the standard training time. Although numerous testing waves were conducted over the 3-year period of the study from which the data were derived, this study relied only on the year 2 posttest administered to intervention, comparison, and typically achieving students. This specific data set was selected for several reasons: (a) The retell component was not included in year 1 of the study while the Passage and Word Reading Fluency subtests of the TMSFA were being developed and validated; (b) not all students participated in the pre-test or progress monitoring waves, thereby limiting the ability to look for potential covariates; and (c) the sample size from year 3 would have been too small for the type of analysis planned for the validation of the retell component (see *Design and Data Analysis* section for more information).

For group-administered assessments, the research assistants would bring together 10-100 students in a room (e.g., library, cafeteria, vacant room). Students were seated in rows, facing forward, and provided with individual stimuli and pencils. One research assistant would read the directions from the assessment manual to the full group while 2 to 12 additional assistants (depending on group size) would monitor students throughout the room. All assistants remained in the room during the test administration to ensure adherence to the procedures outlined in the assessment manual.

For individually-administered assessments, including the retell, research assistants would pull students from a classroom one-at-a-time and take them to the testing room. The assistant would sit directly across from the student and follow the administration procedures in the assessment manual and/or pre-printed on the examiner document. Stimuli were placed in front of students, typically inside plastic sheet protectors and held in binders. After testing a student, the research assistant tallied and recorded data on the examiner document(s).

Handling of data. After each testing session, research assistants checked the student answer documents from group-administered assessments and the examiner documents from individually administered assessments for completion. Packets with missing data were flagged and make-up testing was conducted with students when necessary and possible. Due to absences and school schedule restrictions, some students did not take all assessments included in a testing wave. The handling of missing data in the analysis will be addressed in the section on design and analysis.

Students recorded their answers from group-administered assessments on teleforms, computer-readable documents that allowed for electronic scoring. Research assistants only checked these documents to ensure students had completely filled-in the bubbles. No hand scoring was conducted for these measures.

Individually-administered assessments that required the counting of words/items missed, calculating of rate or accuracy, and the bubbling of correct responses were double-checked for accuracy by an assistant other than the one who administered the assessment(s). Tallies of missed items, the number of words read correctly per minute, and tallies of correct responses were recorded on teleforms included with the examiner documents. The assistant who “double-scored” for accuracy would check the number recorded on the teleform against the tester’s notations of errors on the examiner documents. When a discrepancy in the count or an error in the calculation was found, the second scorer would draw an “X” over the top of the original scorer’s number on the teleform and, then, record the corrected count or calculation.

When all student packets and data from each school site were accounted for, they were delivered with a manifest to researchers at the University of Houston. Those researchers scanned all teleforms and uploaded the information into an electronic database.

Inter-rater reliability of retell scoring. Transcribed student retells were scored once by the original examiner and scored a second time by the researcher. Both scorers were trained in the use of the rubric. Observed inter-rater agreement, calculated by dividing the number of agreements by the sum of the agreements plus disagreements, was 0.66. This is consistent with the findings from the review of research, indicating that holistic scores of overall quality (e.g., Mason et al., 2006) have weaker inter-rater reliability than quantitative counts of included idea units (e.g., Best et al., 2008; Gambrell et al., 1991; Gambrell, Pfeiffer, & Wilson, 1985; Horowitz & Samuels, 1985; McGee, 1982; van den Broek et al., 2001; Wright & Newhoff, 2001; Zinar, 1990). The estimate of inter-rater reliability for the TMSFA retell scores was then adjusted for the possibility of chance with the kappa statistic (Cohen, 1960):

$$K = \frac{\text{Observed agreement} - \text{Chance agreement}}{1 - \text{Chance agreement}}$$

The resulting kappa (*K* = .47) was interpreted as a moderate agreement (Landis & Koch, 1977). It is important to note that inter-rater reliability for the individual passages was the same. That is, the raw percent agreement was 66% for passage 1 scores, passage 2 scores, and passage 3 scores. Likewise, the kappa statistic was .47 for passage 1 scores, passage 2 scores, and passage 3 scores. This stability in observed and chance agreement across passages implies that, although individual scorers often disagreed on the quality of a response, each rater evaluated the scores in a consistent manner. In other words, the inter-rater reliability was only moderate, but the intra-rater reliability was likely quite substantial.

After averaging the three retells, however, inter-rater agreement decreased to .63. The resulting kappa (*K* = .37) was interpreted as a fair agreement (Landis & Koch, 1977). This decline in observed and chance agreement when using the averages was likely attributable to the small 0-3 range in possible scores. Not surprisingly, the greatest number of disagreements was between scores of 1 and 2, or scores of 2 and 3. With a maximum possible sum of 9 (score of 3 x 3 passages = 9), a discrepancy of only one point on a single passage in the set of 3 would change the average by approximately 0.33. This was often a difference, for example, between a 2.67 (rounded to a 3) and 2.33 (rounded to a 2). With two-thirds fewer scores in calculations using the average versus the retells from each of the three passages, the small discrepancies have a higher magnitude of effect on the percent agreement and kappa statistic.

Design and Analysis

Confirmatory factor analysis (CFA) provided the overall framework for analysis. An integrated model building approach was used to address the research aims, with each analysis providing a foundation for subsequent models. CFA belongs to the class of structural equation models. Accordingly, it provides error-adjusted measures of latent constructs based on the covariance structures of observed variables, yielding more precise estimates of relevant factors than the observed values on which the analysis is based. The preferred method for handling missing data in structural models is the direct maximum likelihood (ML) estimator, which is more efficient and unbiased than ad hoc methods (Allison, 2003; Enders & Bandalos, 2001; Schafer & Graham, 2002). Unlike list-wise deletion, ML uses *all* available data within each given case. Deleting cases reduces the sample size, thus inflating standard errors, decreasing statistical power, and lowering the precision of the parameter estimates (Brown, 2006; Buhi, Goodson, & Neilands, 2008; Enders & Bandalos, 2001). Unlike imputation of missing values, direct ML uses *only* available data rather than replacing missing items with plausible values. Predicting scores by regressing the variable with missing data on other variables in the data set for cases with complete data can result in an underestimation of variances and standard errors, as well as an overestimation of correlations (Brown, 2006; Schafer & Graham, 2002).

More conventional missing data techniques, such as ML and multiple imputation (MI) regard missing data as random variables. Although MI corrects for the decrease in variance created by single imputation (Buhi et al., 2008) and exhibits statistical properties similar to ML (Schafer & Graham, 2002), it does not have a single systematic approach.

Because MI is implemented in different ways based on its particular applications, it produces different results each time it is used (Allison, 2003). Hence, where software programs are available to support the model and analysis, ML is often preferred (Buhi et al., 2008). For this study, SPSS version 17.0 (SPSS Inc., 2009) was used to manage the data and calculate descriptive statistics. M*plus* Version 5.21 (Muthen & Muthen, 2009) was used to estimate confirmatory models.

Structural equation models, including CFA, also provide indices of model fit as a means of evaluating the degree to which the available data conform to the specified model. The comparative fit index (CFI) is frequently reported in CFA research, but sample sizes of greater than 100 can inflate CFI (Brown, 2006). For this reason, other indices of comparative fit and parsimony correction were included in the evaluation of model fit. These indices included the Tucker-Lewis index ([TLI]; Tucker & Lewis, 1973) and root mean square error of approximation ([RMSEA]; Steiger & Lind, 1980).

Finally, because direct ML analyzes covariance structures representing different levels of aggregation (e.g., individual, group, etc.), it is more appropriate than traditional approaches when data are clustered, whether by design (i.e., stratified sampling strategy) or circumstance (e.g., students in schools). The extant database used for this study can be considered clustered by circumstance due to the nesting of students within seven different middle schools. In summary, ML is the preferred technique for handling missing scores from any of the identified measures for the cases included in this study because it represents a more efficient and parsimonious use of data, increases power, and yields more reliable estimates of population parameters.

Question 1: What is the factor structure of reading competence expressed in the data obtained from a large, heterogenous sample of students in grades 7 and 8? The baseline factor model is depicted in Figure 3.1. The model was specified in the Bentler-Weeks method (Bentler & Weeks, 1980) in which all variables are assigned as either independent or dependent variables. The circles represent the latent constructs of word identification, fluency, and comprehension and are marked as independent variables by the arrows pointing away from the circles. The rectangles represent the measured variables and are marked as dependent by the arrows pointing toward the rectangles. Correlations among the constructs are depicted by the two-headed arrows. This is an unconditional model because it does not include covariates or specify model constraints.

The first step in the analysis, and the purpose of research question 1, was to evaluate the degree to which this model fits the data. Traditional fit indices were used to evaluate this model, with relative fit indices (CFI, TLI) of at least .95 and RMSEA of .05 or less used as standards (Bovaird, 2007). Model modification indices were used according to best practice to adjust the model to improve fit. The final model provided the basis for the remaining analyses.

Question 2: Is model fit improved by including only narrative retell, only expository retell, narrative and expository retells entered individually, or the average retell performance on narrative and expository passages combined? Retells by passage type (narrative or expository) were entered as covariates in the final CFA model (see Question 1) to estimate effects on existing model parameters, including factor means and factor loadings. The first comparisons were among scores on individual passages (narrative and expository) and the combination of all three scores. For these nested models, the retell scores were evaluated with the χ^2 difference test for significance. The score(s) that significantly improved the fit of the model were then compared with the average retell score across all three passages. For the non-nested comparisons, scores were evaluated on by the Akaike information criterion ([AIC]; Akaike, 1987) and Bayesian information criterion (BIC). The retell score(s) with the lowest AIC and BIC were used for all subsequent analysis.

Question 3: How and to what extent does retell contribute to comprehension? Does retell contribute to fluency and word identification? The most parsimonious TMSFA retell score(s) identified in the previous question were included in the observed covariance matrices used for fitting models. Once the "best fitting" model was identified, the contribution of retell to the estimation of comprehension, fluency, word identification was evident based on the magnitude and statistical significance of the path coefficients from these latent factors to the retell variable (i.e., path coefficients that differ significantly from 0). To more formally evaluate the contribution of retell to the three latent factors, a series of nested model comparisons was conducted. Difference testing of nested models involved constraining the parameter of interest (the above-mentioned path coefficients to retell) as equal to 0 and comparing the fit of the constrained and the fully specified models.

Standards of measurement invariance differ by area of study and by circumstances of practice. For purposes of this study, a relatively less restrictive standard was used: statistical equivalence on factor loadings (i.e., non-significant difference in χ^2

estimates; [$\Delta\chi^2$]) when factor means are constrained at 0 was sufficient evidence of invariance (Schmitt & Kuljanin, 2008). If the fit of the constrained and full models was not significantly different, the coefficient in question was considered less useful in modeling reading competence. Based on the CFA conducted by Shinn and colleagues (1992) with fifth-grade students, it was anticipated that no differences would be found by constraining the word identification and/or decoding path coefficients to retell, but that the reading comprehension path coefficient to retell would be significantly different when relaxed.

Question 4: What are the patterns of associations between the TMSFA retell instrument and other standardized measures of reading? This phase of the analysis included a calculation of the correlations of the factors to the measures and correlations among the factors described earlier in this chapter (also identified in the rectangles in Figure 1). The expectation was that the TMSFA retell would be moderately correlated to the four other measures of reading comprehension and weakly correlated to the seven measures of word identification and fluency.

Question 5: Is retell influenced by differences in primary language, ability level, or socioeconomic status over and above the effects of reading comprehension? To evaluate group differences in retell ability, cases in the dataset were coded by inclusion in groups: socioeconomic status (defined by participation in free/reduced-price lunch program), bilingual, English language learner (ELL), limited English proficient (LEP), and ability level (defined by participation in general or special education as well as by performance on the K-BIT). Given the smaller sample size, multi-group modeling with nested comparisons based on group (Bovaird, 2007; Mehta & Neale, 2005) was not used. Rather, a multiple indicators, multiple causes or MIMIC model (Joreskog & Goldberger, 1975) was conducted by adding the aforementioned groups as covariates to the CFA. MIMIC models with categorical indicators have demonstrated equivalence to differential item functioning (DIF) analysis and have the advantage of modeling a direct effect of the covariate on the latent factor (MacIntosh & Hashim, 2003; Muthen, Kao, & Burstein, 1991). In this study, DIF would be indicated if the factor means were significantly different at the different levels of the covariates

CHAPTER 4

Results

This study was conducted to examine the validity of the retell task included in the Texas Middle School Fluency Assessment ([TMSFA]; Texas Education Agency, University of Houston, & The University of Texas System, 2008a). An extant database gathered from a diverse sample of 394 seventh- and eighth-grade students was used for the analysis. Of the 13 measures administered, 5 were considered indicative of the latent construct “word identification,” 4 were considered indicative of the latent construct “fluency,” and 4 were considered indicative of the latent construct “comprehension” (Figure 1). The three constructs are said to comprise overall “reading competence” among students of this age group (Buly & Valencia, 2003; Hock et al., 2009; Jackson, 2005; Valencia & Buly, 2004; Vukovic, Wilson, & Nash, 2004). In addition to the 11 assessments included in the baseline model of reading competence, an average of three retell scores was tested as a predictor for comprehension, and the K-BIT was used as a categorical indicator of students’ ability levels.

Primary Questions

The purpose of this study was to examine the validity of the retell task included in the TMSFA (Texas Education Agency, University of Houston, & The University of Texas System, 2008a) within a confirmatory factor analysis (CFA) framework (Brown, 2006; Byrne, 1988; Marsh & Bailey, 1991; Thompson, 2004). The research questions addressed were:

1. What is the factor structure of reading competence expressed in the data obtained from a large, heterogenous sample of students in grades 7 and 8?
2. Is model fit improved by including only narrative retell, only expository retell, narrative and expository retells entered individually, or the average retell performance on narrative and expository passages combined?
3. How and to what extent does retell contribute to comprehension? Does retell contribute to fluency and word identification?
4. What are the patterns of associations (correlation, regression) between the TMSFA retell instrument and other standardized measures of reading?
5. Is retell influenced by differences in primary language, ability level, or socioeconomic status over and above the effects of reading comprehension?

Preparation of the Dataset

In preparing to model the contribution of retell to students' reading competence, the extant database was assessed for normality using tests of skewness and kurtosis (Table 4.1). Several variables were found to have values outside the desired -1 to +1 range. With an adequate sample size, normality is still assumed if the skewness values do not exceed the -2 to +2 range, and the kurtosis values do not exceed the -3 to +3 range (Garson, n.d.). However, two variables, TAKS (taks_ss0708) and the Woodcock Johnson Passage Comprehension subtest (WJ_PassComp), still exceeded acceptable limits.

Table 4.1

Descriptive Statistics: Original Database

	N	Skewness		Kurtosis	
	Statistic	Statistic	Std. Error	Statistic	Std. Error
KBITcomp	370	-.153	.112	.219	.223
TAKS	394	-4.227	.118	22.202	.236
AveRetell	378	-.103	.125	-.315	.250
WJ_LetterWord	385	-.448	.124	1.474	.248
WJ_WordAttack	385	.239	.124	.395	.248
WJ_PassComp	383	-.677	.125	3.401	.249
TOWRE_SightWord	384	.036	.125	.434	.248
TOWRE_PhonDecod	384	.382	.125	.207	.248
TOWRE_Summ	383	-.015	.125	-.045	.249
TMSFA_AveWordES	388	-.116	.124	-.148	.247
TMSFA_AvePassES	386	-.179	.124	.387	.248
AIMSmaze	377	.751	.126	1.313	.251
GRADEcomp	381	.335	.125	1.826	.249
WJ_Spell	371	-.904	.127	1.165	.253
TOSRE_sum	376	-.256	.126	-.036	.251
Valid N (listwise)	320				

A visual inspection of the Q-Q plots indicated there were outliers that might be affecting the distribution of the scores. Therefore, Mahalanobis distances [χ^2 (14, N=394) = 36.123, p < .001] were evaluated for the variables of interest, and 83 cases exceeding the critical value were removed from the dataset. This reduced the sample size to 311, which was still sufficient for the CFA because it met or exceeded the 3 cases: parameter ratio. As Table 4.2 reveals, resulting values were within acceptable ranges. This table also includes the means and standard deviations of each measure.

Table 4.2

Descriptive Statistics: Outliers Removed

	N	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
KBITcomp	287	97.66	13.675	-.371	.144	.607	.287
TAKS	311	2189.79	168.036	.200	.138	.312	.276
AveRetell	311	1.85	.616	-.097	.138	-.237	.276
WJ_LetterWord	311	99.12	11.645	.033	.138	-.125	.276
WJ_WordAttack	311	99.14	10.990	.579	.138	.255	.276
WJ_PassComp	311	93.43	10.866	-.253	.138	1.453	.276
TOWRE_SightWord	311	97.52	11.157	.473	.138	-.045	.276
TOWRE_PhonDecod	311	100.07	15.382	.428	.138	.266	.276
TOWRE_Summ	311	98.27	14.363	.275	.138	-.272	.276
TMSFA_AveWordES	311	74.78	18.541	.049	.138	-.141	.276
TMSFA_AvePassES	311	145.75	31.804	.141	.138	.023	.276
AIMSmaze	311	190.84	58.356	.633	.138	1.559	.276
GRADEcomp	311	90.62	11.212	.458	.138	1.922	.276
WJ_Spell	311	96.18	14.860	-.714	.138	1.101	.276
TOSRE_sum	311	91.13	14.123	-.150	.138	.062	.276
Valid N (listwise)	287						

Before analyzing the baseline model (Figure 3.1), the measures were assessed for multicollinearity to confirm the correct measures or components were being entered into the model. Specifically, the TOWRE subtests were assumed to be highly correlated such that the TOWRE summary score would be preferred over entering the Sight Word Efficiency and Phonemic Decoding Efficiency subtest scores separately. Therefore, the tolerance and variance inflation factor (VIF) values were examined to determine whether measures were dependent upon each other. Tolerance values of .01 or less or VIF values greater than 10 are considered suggestive of multicollinearity (Stevens, 2002).

Table 4.3

Collinearity Statistics

Model		Tolerance	VIF
1	(Constant)		
	WJ_LetterWord	.291	3.431
	WJ_WordAttack	.360	2.780
	WJ_PassComp	.391	2.558
	TOWRE_SightWordEff	.024	41.035
	TOWRE_PhonemeDecodEff	.018	54.235
	TOWRE_Summ	.006	164.012
	TMSFA_AveWordES	.169	5.923
	TMSFA_AvePassES	.197	5.085
	AIMSmaze	.709	1.410
	GRADEcomp	.539	1.854
	WJ_Spell	.507	1.974
	TOSRE_sum	.438	2.285
	KBITcomp	.460	2.172
	TAKS	.474	2.110

The data on Table 4.3 reveal that the TOWRE subtests and summary score all had exceptionally high VIF values and questionable tolerance values. Therefore, the correlations among these three scores were analyzed. As anticipated, the TOWRE Sight Word Efficiency (*r* = .916) and Phonemic Decoding Efficiency (*r* = .943) subtest were both highly correlated to the TOWRE summary score, so the decision to enter only the summary score into the baseline model was confirmed.

Analysis of the Baseline Model: The Factor Structure of Reading Competence

The initial analysis concerned the fit of the baseline model depicted in Figure 3.1. This did not converge in 50,000 iterations, so the model was revised (see Figure 4.1) to remove the cross-loadings of the TOWRE summary and the TMSFA Word Reading Fluency subtest. The former was included as a dependent variable for the fluency construct only, and the latter was included as a measure of the word identification construct only. The revised model did not converge either. Therefore, the dependent variables were redefined conceptually to identify the most theoretically supported measures for each of the three constructs.

Both the TMSFA Word Reading Fluency (*r* = .868) and Passage Reading Fluency (*r* = .813) subtests were strongly correlated to the TOWRE summary. The TMSFA Word Reading Fluency subtest is more similar to the TOWRE in that it assesses words in isolation; whereas, the TMSFA Passage Reading Fluency subtest assesses words correct per minute with connected text. Consequently, the TMSFA Word Reading Fluency subtest was removed from the model because it did not contribute unique information above what was contributed by the TOWRE summary. The TMSFA Passage Reading Fluency subtest was retained as a dependent variable for fluency.

This left the minimum required three measures for word identification: WJ-III Word Attack, WJ-III Letter Word Identification, and WJ-III Spelling. To determine the third measure for fluency, the correlations among the AIMSweb Reading Maze and other measures were examined. The AIMSweb Reading Maze was intended as a measure of comprehension, but it was only weakly correlated (from *r* = .195 to *r* = .281) to the other dependent variables for this latent construct. In comparison, moderate correlations were evident with the dependent variables for the fluency construct (from $r = .436$ to $r = .517$). Therefore, the AIMSweb Reading Maze was moved within the model to be a measure of fluency.

The TOSRE, however, was removed from the model. It demonstrated moderate correlations with the measures of word identification (from $r = .441$ to $r = .551$), fluency (from $r = .436$ to $r = .646$), and comprehension (from $r = .481$ to $r = .567$). Conceptually, then, it could not be clearly distinguished as a dependent variable for any one latent construct. Moreover, the TOSRE differs from the other measures of fluency in that it is based on sentences correct per minute, rather than words correct per minute. After removing the TOSRE, three measures of fluency remained: TOWRE summary, TMSFA Word Reading Fluency, and AIMSweb Reading Maze.

The three measures of comprehension in the model were the GRADE, WJ-III Passage Comprehension, and the TAKS. The average of the retell scores from the TMSFA was included as a dependent variable of comprehension as well, but because its contribution was still being tested, it was not considered one of the measures needed to meet the minimum specifications for CFA. This conceptually redefined model (see Figure 4.2) converged and demonstrated adequate fit ($\chi^2 = 97.316$ {32}; CFI = 0.958; TLI = 0.941; RMSEA = .081). Although the TLI value is slightly less than the desired .95, Hu and Bentler (1999) suggest a value "close to" .95 is acceptable because the recommended cut point can fluctuate by modeling conditions. A value below .90 would suggest rejecting the model (Bentler, 1990), which was not the case here.

Similarly, an RMSEA of .05 or less is preferable, but Browne and Cudeck (1993) lend support to considering an upper limit of .08. This is confirmed by others who believe that RMSEAs between .08 and .10 are still indicative of "mediocre" fit with the model not rejected until the value exceeds .10 (MacCallum, Browne, & Sugawara, 1996). The 90% confidence interval for the model tested here (Figure 4.2) was from .063 to .100. In addition, the standardized root mean square residual (SRMR) was .039, which is close to the desired SRMR of 0.0 (Brown, 2006). Taking all indices of fit into consideration, the conceptually redefined 3-factor model of reading competence was accepted

Model Fit by Passage Type

The next phase of analysis examined whether the fit could be improved by using one or more retell scores from individual passages rather than the average of the three retell scores. The three retell scores were derived from two expository passages and one narrative passage (passages are provided and labeled in Appendix B). When entered individually into the model, the retell score from the narrative passage was the best predictor with a moderate but significant factor loading on comprehension of .352 ($p < .001$) compared to the weak but significant factor loadings of the expository passages (.264 and .221; $p < .001$). The AIC and BIC values were both lower for the model with the narrative retell score alone than for the model that included all three retell scores. Therefore, this nested model comparison was evaluated with the χ^2 difference test, which was significant ($\Delta\chi^2$=134.261{19}; $p < .001$).

However, entering the average across the three retell scores produced a more parsimonious model than entering the retell score from the narrative passage alone. Not only were the AIC and BIC values lower for the average retell score (ΔAIC = 58.275; ΔBIC = 58.275), but the RMSEA and SRMR values were also slightly lower (Table 4.4). The relative fit indices (CF I, TLI) further confirm that entering the narrative retell score alone decreased model fit. Therefore, the average of the three retell scores was used for all subsequent analysis.

Table 4.4

Tests of Model Fit

	CFI	TLI	RMSEA	SRMR	AIC	BIC
Narrative Retell	.956	.938	.084	.042	24412.788	24536.201
Average Retell	.958	.941	.081	.039	24354.513	24477.926

Contribution of Retell to the Latent Constructs

With the retell score that produces the “best fitting” model identified, the contribution of retell to each of the three latent factors was evaluated through χ^2 difference testing of nested models. The difference between the constrained versus the fully specified path coefficients to retell was significant for comprehension ($\Delta\chi^2$=16.652{1}, p < .001), fluency ($\Delta\chi^2$=10.882{1}, p = .001), and word identification ($\Delta\chi^2$=7.84{1}, p = .005). However, the χ^2 difference and the factor loading (Table 4.5) were greater for comprehension, so the model depicted in Figure 4.2 was not revised. The average retell score remained as an indicator of comprehension only, suggesting it is less indicative of students’ word identification and fluency ability.

Table 4.5

Factor Loadings of Retell on the Latent Constructs

	Estimate	Standard Error (S.E.)
Word Identification by average Retell	0.167*	0.058
Fluency by average Retell	0.194*	0.057
Comprehension by average Retell	0.250*	0.059

*p < .001

The Patterns of Associations among the Measures and Constructs

The correlations among the measures in the final model are provided in Table 4.6. The average retell score was weakly but significantly (p < .01) correlated with the measures of fluency (from *r* = .158 to *r* = .183) and comprehension (from *r* = .155 to *r* = .257). The strongest correlations were with the WJ-III Passage Comprehension (*r* = .208) and TAKS(*r* = .257), the two measures with the highest factor loadings on the comprehension construct. The weakest correlations were between average retell and the measures of word identification where only the correlation coefficient for the WJ-III Letter Word Identification was significant (*r* = .132, p < .05). Consistent with retell's factor loading on comprehension, the average retell score was more related to measures of comprehension than to measures of word identification or fluency. Retell bore the weakest relationship to other measures of word identification, which is in contrast to the moderate and significant relationships between the TMSFA passage reading fluency subtest and the measures of word identification (from *r* = .550 to *r* = .595, p < .01). The

TMSFA passage reading fluency subtest had the weakest relationship to other measures of comprehension (from *r* = .430 to *r* = .498, p < .01). Although still stronger than the relationship of retell to the GRADE, WJ-III passage comprehension, and TAKS, the results suggest that the ORF portion of the TMSFA is more associated with word identification skills than comprehension. The retell component of the same measure, on the other hand, is more associated with comprehension skills than word identification.

Nearly all other measures included in the model demonstrated moderate to strong relationships (p < .01) with each other. The exceptions to this were between the GRADE composite and the WJ-III Word Attack subtest (*r* = .264) and AIMS reading maze (*r* = .195), and between the TAKS and AIMS reading maze (*r* = .281). Recall that the weak relationships between AIMS reading maze and the other measures of comprehension was the reason AIMS reading maze was moved within the model to be a measure of fluency. It is interesting to note that while retell had a consistently weak relationship to the other measures in the model but the strongest relationship to the measures of comprehension, AIMS maze had moderate relationships with measures of word identification (from *r* = .324 to *r* = .358, p < .01) and fluency (from *r* = .462 to *r* = .517, p < .01) but among the weakest relationships to the measures of comprehension (from *r* = .195 to *r* = .329, p < .01). As with the TMSFA passage reading fluency subtest, this seems to suggest that measures assessing words correct per minute are less sensitive to comprehension ability among seventh- and eighth-grade students.

Table 4.6

Correlations Among the Reading Variables

Variable	1	2	3	4	5	6	7	8	9	10
1. WJ_WordAttack		.742**	.555**	.630**	.550**	.328**	.264**	.479**	.322**	.037
2. WJ_LetterWord	.742**		.609**	.644**	.595**	.358**	.408**	.578**	.509**	.132*
3. WJ_Spell	.555**	.609**		.579**	.572**	.324**	.339**	.508**	.387**	.075
4. TOWRE_Summ	.630**	.644**	.579**		.813**	.462**	.336**	.453**	.383**	.158**
5. TMSFA_AvePassES	.550**	.595**	.572**	.813**		.517**	.430**	.498**	.471**	.180**
6. AIMSmaze	.328**	.358**	.324**	.462**	.517**		.195**	.329**	.281**	.183**
7. GRADEcomp	.264**	.408**	.339**	.336**	.430**	.195**		.564**	.555**	.155**
8. WJ_PassComp	.479**	.578**	.508**	.453**	.498**	.329**	.564**		.581**	.206**
9. TAKS	.322**	.509**	.387**	.383**	.471**	.281**	.555**	.581**		.257**
10. AveRetell	.037	.132*	.075	.158**	.180**	.183**	.155**	.206**	.257**	

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

The relationships of the factors to the measures are provided in Table 4.7 and Figure 4.3. Retell was a weak but significant predictor of comprehension with a very high residual variance (depicted in the model in a small circle with an arrow pointing toward the measure). A large residual variance indicates the variable does not function well as a measure of the construct. It is possible the weak inter-rater reliability for the current scoring mechanism (discussed in Chapter 3) is contributing to the poor functioning of the average retell scores in the model. Nevertheless, retell as included in this model was less indicative of comprehension ability than the other, more formal measures (GRADE, WJ-III passage comprehension, and TAKS).

Table 4.7

Relationships of the Factors to the Measures

	Estimate	S.E.	Residual Variance
Word Identification by			
WJ_LetterWord	.891*	.019	.205
WJ_WordAttack	.808*	.024	.347
WJ_Spell	.711*	.033	.495
Fluency by			
TOWRE_Summ	.903*	.018	.185
TMSFA_AvePassES	.901*	.019	.187
AIMSmaze	.541*	.043	.707
Comprehension by			
GRADEcomp	.685*	.037	.531
WJ-PassComp	.825*	.030	.320
TAKS	.737*	.035	.457
AveRetell	.250*	.059	.938

*p < .001

The relationships among the latent constructs are provided in Table 4.8 and Figure 4. As expected from the development of the baseline model, the three constructs were all significantly correlated. The strongest correlations were with word identification. Given that all the measures except TAKS were timed, one hypothesis explaining the correlations among the factors is that the timed tests place pressure on the speed with which words can be read accurately or processed (Catts, Gillispie, Leonard, Kail, & Miller, 2002; Cutting & Scarborough, 2006; Jackson, 2005).

Table 4.8

Relationships Among the Latent Constructs

	Estimate	S.E.
Fluency with Word Identification	.799[*]	.030
Comprehension with Word Identification	.722[*]	.040
Comprehension with Fluency	.640[*]	.045

[*] $p < .001$

Influence of Socioeconomic Status, Primary Language, and Ability Level

In the final phase of the analysis, students' socioeconomic status, language proficiency, and ability level were entered into the model as covariates. Each covariate was treated as a dichotomous variable. For example, students who were receiving free or reduced-priced lunch were coded as "1," and those not receiving free or reduced-priced lunch were coded as "0" on the variable "SES." The K-BIT composite scores were converted into a categorical indicator for ability level. Students whose standard score was greater than or equal to 100 were coded as "1" for "above average," and those whose score was less than 100 were coded as "0" for "below average."

The MIMIC testing followed a two-step approach. The first testing was for overall latent differences on the covariates. As Table 4.9 indicates, there were significant small to moderate group differences on the three latent factors defining reading competence (word identification, fluency, and comprehension). This was particularly true with respect to comprehension performance where all groups except limited English proficient demonstrated DIF.

Table 4.9

Influence of Student Characteristics on Factors

	Estimate	S.E.
Word Identification on		
SES	-0.109	0.057
BILINGUAL	-0.221**	0.070
ELL	-0.098	0.072
LEP	0.174***	0.078
SPECIAL EDUCATION	-0.330*	0.055
KBIT	0.312*	0.055
Fluency on		
SES	0.002	0.059
BILINGUAL	-0.204**	0.072
ELL	-0.080	0.074
LEP	0.017	0.081
SPECIAL EDUCATION	-0.307***	0.054
KBIT	0.221***	0.058
Comprehension on		
SES	-0.154**	0.053
BILINGUAL	-0.315***	0.065
ELL	-0.180**	0.067
LEP	0.066	0.073
SPECIAL EDUCATION	-0.219***	0.050
KBIT	0.469***	0.049

*p < .001; **p < .01; ***p < .05

This initial step of the MIMIC testing also included an examination of the modification indices to identify any specific observed indicator differences that should be tested in the second step. Although the model did not suggest average retell demonstrated DIF, retell was tested to be consistent with the research questions for this study. As shown in Table 4.10, there were no significant differences for any groups on average retell performance. The lack of differences in factor means for retell indicate the groups do not differ on intercepts.

Table 4.10

Influence of Student Characteristics on Retell Measure

	Estimate	S.E.
Average Retell on		
SES	0.010	0.062
BILINGUAL	-0.016	0.079
ELL	0.041	0.077
LEP	-0.048	0.082
SPECIAL EDUCATION	0.066	0.060
KBIT	-0.077	0.074

Because there were no significant differences among groups on average retell scores, the differences observed on the comprehension construct cannot be attributed to students' retell performance. Students who were classified as receiving free or reduced-price lunch, bilingual, English language learners, in special education, or below average on the K-BIT intelligence test performed significantly worse on the standardized measures of comprehension. Whereas on the retell measure, students in and out of these categorical groups all had comparable scores.

Summary

With a normally distributed sample of 311 seventh- and eighth-grade students, a three-factor model of reading competence converged and was, therefore, accepted. Although retell was only weakly correlated to the comprehension construct and to other standardized measures of comprehension, the model demonstrated adequate to mediocre fit ($\chi^2 = 97.316$ {32}; CFI = 0.958; TLI = 0.941; RMSEA = .081) with retell included. In contrast, one measure of word identification (the TMSFA Word Reading Fluency subtest) and one measure of fluency (TOSRE) had to be removed before the model would converge. Retell did, however, have a large residual variance (.938), suggesting it did not function well as a measure of comprehension as currently administered after a timed fluency test and with a demonstrated low inter-rater reliability.

Narrative retell scores were better predictors of comprehension than expository retell scores or the combination of all three scores, but average retell scores produced a more parsimonious model than when narrative retell scores alone were entered. In addition, retell did not demonstrate DIF when student characteristics (e.g., primary language, socioeconomic status, ability level) were entered as covariates, even though there were overall latent differences.

CHAPTER 5

Discussion

In this study, data from an extant database of seventh- and eighth-grade participants at a range of ability levels were used to model reading competence in a confirmatory factor analysis framework. By drawing from a diverse sample of adolescents, the research expands upon previous factor analyses conducted with third- and fifth-grade students (Shinn et al., 1992) as well as studies that included only those middle school students identified as struggling with reading (Buly & Valencia, 2003; Hock et al., 2009; Valencia & Buly, 2004). Knowing the factor structure of reading for a normally distributed sample of students in grades 7-8 and the specific role retell plays in that model of reading competence, could contribute to efficiently assessing middle school students and planning effective instruction or intervention.

The retell component of the TMSFA (Texas Education Agency, University of Houston, & The University of Texas System, 2008a) was a weak but significant contributor to the latent comprehension construct in a three-factor model of reading competence. Despite the existence of overall latent differences, average retell performance was not influenced by student characteristics. This chapter will discuss the findings with respect to each of the five research questions; the possible implications of the results for the administration, scoring, and use of retell protocols; the inherent limitations to the interpretation and generalizability of the findings; and potential areas for further research.

Findings with Respect to Research Questions

This study addressed the need for more data on the technical adequacy of retell, administered within an ORF approach, as a significant and efficient measure of adolescents' reading competence. Although ORF measures have been shown to be reflective of the overall reading ability of students in grades 1-5 (Burke et al., 2009; Fuchs et al., 2001; Spear-Swerling, 2006), research suggests ORF may not be a comprehensive index of reading comprehension for adolescents. Specifically, the correlation between ORF and reading comprehension is less robust for students above grade 5 (Schatschneider et al., 2004; Wiley & Deno, 2005), an age at which rate and accuracy scores begin to asymptote (Fuchs et al., 2001; Stage & Jacobsen, 2001). Moreover, teachers are reluctant to accept students' reading rate and accuracy as an indicator of how well a text was understood (Applegate, Applegate, & Modla, 2009; Goodman, 2006; Shinn et al., 1992).

The addition of a retell task to ORF might provide a solution to efficient progress monitoring of reading comprehension for middle school students; however, existing measures have not been validated for this purpose. By including the retell component of the TMSFA (Texas Education Agency, University of Houston, & The University of Texas System, 2008a) in a confirmatory factor analysis, the study sought to determine whether and to what extent the retell scores were indicative of seventh- and eighth-grade students' reading comprehension ability. Findings from each phase of the analysis will be discussed separately.

The factor structure of reading competence. The first research question concerned the fit of the theorized model of reading competence for middle school students to the data. Previous research has indicated a developmental difference in the number of latent constructs that comprise a student's reading ability. Data from students in first- (Burke & Hagan-Burke, 2007) and third-grades (Shinn et al., 1992) demonstrated a single-factor model in which ORF scores had the highest factor loading. In fourth- and fifth-grades, however, studies found that word-level skills (including ORF) were distinct from comprehension skills (Leach et al., 2003; Shinn et al., 1992). The literature on the age group in this dissertation study suggested decoding accuracy be separated from naming speed or text reading rate, resulting in three latent constructs: word identification, fluency, and comprehension (Fletcher et al., 2007; Jackson, 2005; Vukovic et al., 2004).

Results of the CFA indicate the three-factor model for seventh- and eighth-grades students was confirmed, although the fit of the model to the data might be considered mediocre (MacCallum et al., 1996). Unlike previous studies conducted with middle school students (Buly & Valencia, 2003; Hock et al., 2009; Valencia & Buly, 2004), scores from students at a range of ability levels, including those typically achieving in reading, were included in the analysis. This increases confidence in accepting the model; however, it is possible that the fit might have been improved if students were grouped by ability level (as will be discussed in a subsequent section) or if the model included a fourth latent construct (i.e., vocabulary).

Path analyses of predictor variables found vocabulary knowledge makes a large, significant contribution to reading comprehension (Cromley & Azevedo, 2007; Kintsch, 1988; Perfetti, 1985). The decision to test only a three-factor model of reading competence was theoretically-based in that vocabulary and comprehension have demonstrated such a strong, bidirectional relationship as to make distinctions between the two abilities difficult (Carlo et al., 2004; Leach et al., 2003; Snow, 2002; Valencia & Buly, 2004; Wagner et al., 2007). However, a principle component analysis conducted with ninth-grade students suggests vocabulary may be a distinct domain of reading that discriminates a small percentage of students (4%) who have adequate comprehension but low vocabulary knowledge (Hock et al., 2009). Because vocabulary was not tested as a separate latent construct, the results from the current study do not allow for conclusions as to whether a four-factor model of reading competence would be more parsimonious for seventh- and eighth-graders.

Retell by passage type. Previous research found students recall less information from expository than from narrative passages (Best et al., 2008). Therefore, the second research question in this dissertation examined model fit by comparing retell scores on a narrative passage, two expository passages, the combination of narrative and expository passages, and the average of the three retell scores. Results indicated that retell scores from the narrative passage were the best predictor of comprehension compared to scores from individual expository passages or the combination of all three scores. However, the average of the three scores produced a more parsimonious model than narrative retell alone.

Unlike the previous studies referenced above, the TMSFA retell scores are based on holistic evaluations of accuracy, completeness, and quality rather than on quantitative counts of pre-determined idea units. In addition, the TMSFA passages are not clearly distinguishable as expository but might be more accurately described as informational narratives. This is because their appearance is identical to the narrative passages. In other words, the passages labeled “expository” do not have subheadings or other features more reminiscent of subject matter textbooks. The content of the expository passages included in this study was descriptive and biographical (see Appendix B), so they did not place demands on awareness of more challenging text structures. According to Richgels and colleagues’ (1987) study of student retells with expository text, certain text structures, such as causation, are more challenging for middle school students to read and recall than other text structures such as cause-effect. Moreover, Zinar (1990) found students had poorer retell performance on expository passages when the relationship among the ideas was implicit than when it was explicitly stated. Hence, one hypothesis for explaining why the model was more parsimonious with expository retells included in the average score is that the results are a function of the type and structure of the expository text in the TMSFA. If students read passages with implicit causal relations as opposed to explicit descriptions, retells on those expository passages might have had different effects on existing model parameters.

Although not reflective of more traditional or complex expository text, the TMSFA passages labeled “expository” include more facts and require more understanding of history, geography, and culture than the passage labeled “narrative.” Therefore, a possible explanation for the difference in the factor loadings of the narrative versus the expository retell scores is that retell performance was influenced by student background knowledge. After vocabulary, background knowledge has been found to be the second largest contributor to ninth-graders' reading comprehension (Cromley & Azevedo, 2007). This would be consistent with Best and colleagues' (2008) finding that background knowledge was the best predictor of expository retell performance. Because a student's degree of background knowledge varies across passages on different topics, retell performance on individual expository passages is likely less indicative of overall reading competence than if the scores are averaged or are derived from a narrative passage in which prior knowledge was less relevant to understanding.

The extent to which retell contributes to comprehension, fluency, and word identification. Based on a CFA conducted by Shinn and colleagues (1992), it was expected that retell would not make a significant contribution to fluency or word identification. In addressing the third research question in this study, results revealed that the path coefficients to retell were, in fact, significant for all three constructs. However, the data suggest retell was most indicative of students' comprehension ability and least indicative of students' word identification ability. Because the analysis did not examine indirect pathways from retell performance to comprehension, any identified relationship might best be considered an indicator of literal comprehension ability as would be consistent with existing studies of retell (Best et al., 2008; Zinar, 1990).

Compared to the other three measures of comprehension included in the model, retell had, by far, the lowest factor loading and the highest residual variance. This suggests it is a poor indicator of the construct. However, it is worth noting that the model still converged with retell included. The same cannot be said of one measure of word identification (the TMSFA Word Reading Fluency subtest) and one measure of fluency (TOSRE); both were removed from the model when it was conceptually redefined after failing to converge in 50,000 iterations. Given the weak inter-rater reliability of the holistic scoring mechanism used in the TMSFA retell component and it administration following a timed fluency test, it is possible the data on the contribution of retell presented here is more a function of these particular scores than of the actual validity of "retell" as a dependent variable.

The Patterns of Associations among the Measures and Constructs. The fourth research question concerned the correlation of retell to standardized measures of the three latent constructs. Findings were in contrast to previous studies with students in and around the same age group that demonstrated moderate to high correlations between retell and measures of word identification, fluency, vocabulary, and comprehension (Carlisle, 1999; Fuchs et al., 1988; Hansen, 1978). In this dissertation study, retell was only weakly correlated to the other measures. However, it bore the strongest relationship to the best functioning predictors of comprehension (i.e., WJ-III Passage Comprehension and TAKS) and almost no relationship to measures of word identification.

As noted in the previous section, the weak relationship of retell to the other measures could be due to the inter-rater reliability of the holistic scoring mechanism because the studies that reported higher correlations were based on quantitative methods of scoring retells. Yet, the findings here are still noteworthy for two reasons. First, two measures of fluency, the TMSFA passage reading fluency subtest and AIMS reading maze, demonstrated a pattern opposite that of retell: the fluency measures were more associated with word identification measures than comprehension measures. Second, the two best predictors of comprehension were moderately correlated to measures of word identification; whereas, there was only one weakly significant correlation between retell and a measure of word identification (WJ-III Letter Word Identification). This seems consistent with the finding of Keenan and colleagues (2008) that retell was less sensitive to decoding ability than other standardized measures of reading comprehension.

Covariates. A number of studies have reported students with learning disabilities do not recall as much information as students without identified disabilities (Carlisle, 1999; Gambrell et al., 1991; Hansen, 1978; Horowitz & Samuels, 1985; McGee, 1982; Zinar, 1990). Other researchers have cautioned that socioeconomic status and cultural-linguistic differences might influence student performance on comprehension tasks, such as retell, that require oral language processing (Snyder, Caccamise, & Wise, 2005). Therefore, the final research question examined whether these student characteristics influenced retell performance. Despite overall latent differences on the covariates, retell scores did not exhibit any significant differences by group.

This is an important finding that suggests students who are from a lower socioeconomic status, speak a primary language other than English, are enrolled in special education, or have lower academic ability have significantly poorer performance on standardized measures of comprehension but not on retell. Previous studies of retell found ability differences when utilizing quantitative counts of pre-determined idea units because lower ability students and those with learning disabilities did not offer as much information unless specifically prompted to do so (Gambrell et al., 1991; Gambrell &

Jawitz, 1993; Zinar, 1990). In the current study, the subjective nature of the holistic approach to evaluating retells may have facilitated taking certain student characteristics into account. Rather than basing the score on a straightforward count of idea units, raters could draw on other impressions of quality and completeness that might have accommodated for what would otherwise be considered an insufficient response. However, this cannot be determined from available data.

Given that group differences were apparent for all three constructs, a conditional model that included the student characteristics as covariates might have resulted in a better fit to the data. That was not tested here because the research questions were specific to DIF on retell only.

Summary and Implications

Overall, the data on the retell component of the TMSFA indicate it currently lacks the technical adequacy to be a valid and reliable measure of reading comprehension for seventh- and eighth-grade students. If an assessment is no more valid than it is reliable, the primary concern is that the retell data used in the models tested here were based on a holistic scoring mechanism. Consistent with what has been previously reported (Gambrell & Jawitz, 1993; Klesius & Homan, 1985; Loyd & Steele, 1986; Pearman, 2008; Popplewell & Doty, 2001; Richgels et al., 1987), such an approach to evaluating students' responses had a rather low inter-rater reliability. Interestingly, however, scores from a single rater appeared to be rather consistently applied because the calculated inter-rater reliability was identical for each of the three passages. So, although two raters might disagree on how to score a response, there seems to be reliability with respect to how one rater applies the holistic criteria. Nevertheless, to make valid interpretations of students' scores, a more reliable method of scoring retell responses is needed.

The extant literature indicates a quantitative approach, such as counting the number of pre-determined idea units recalled, would improve reliability (Best et al., 2008; Gambrell et al., 1991; Gambrell et al., 1985; Horowitz & Samuels, 1985; McGee, 1982; van den Broek et al., 2001; Wright & Newhoff, 2001; Zinar, 1990). In fact, the finding from this dissertation study that retell was only weakly correlated to other standardized measures of reading is in contrast with the moderate to strong correlations found in previous studies conducted with students of the same age group using a quantitative method of scoring the retells (Carlisle, 1999; Fuchs et al., 1988; Hansen, 1978). Despite the weak correlations, the TMSFA retell component appears to have the potential to provide a different portrait of students' reading competence than that depicted by the ORF component, the TMSFA passage reading fluency subtest. In terms of factor loadings and correlations to other measures, the retell component was most closely associated with reading comprehension and least associated with word identification. On the other hand, with the exception of the TOWRE summary scores, the TMSFA passage reading fluency subtest was most associated with word identification and least associated with comprehension.

Among the latent constructs, fluency and word identification had the strongest relationship (.799); whereas, fluency and comprehension had only a moderate relationship (.640). In fact, comprehension was more associated with word identification (.722) even though the retell scores generally were not correlated to this construct.

Because the three-factor model of reading competence converged, this could still be considered supportive of the notion that the role of decoding accuracy in comprehension diminishes in adolescence (Gough et al., 1996; Keenan et al., 2008) as compared to its contribution in a one-factor model for younger students (Shinn et al., 1992). These findings also seem to confirm previous research that found a weaker relationship between fluency and comprehension above grade 5 (Schatschneider et al., 2004; Wiley & Deno, 2005).

Word-level deficits are the primary focus of early identification and prevention of reading disabilities but are considered distinct from comprehension deficits in students with learning disabilities (Lyon et al., 2001). About 25% of elementary children provided with intensive, explicit, and systematic instruction in auditory discrimination, phonics, and word identification still demonstrate persistent difficulties beyond the elementary years (Juel, 1988; Torgesen et al., 2001; Velluntino et al., 1996). Word-level reading disabilities, or dyslexia (Fletcher et al., 2007), can make it difficult to accurately assess reading comprehension, particularly when the instruments are timed (Catts et al., 2002). If retell can provide unique information on the reading skills of adolescents in special education, teachers and reading interventionists would be better able to plan targeted instruction in the appropriate areas.

Developing a retell component with better reliability and better functioning as a measure of comprehension (i.e., higher factor loading and lower residual variance), might make it an efficient compliment to ORF in monitoring the overall reading progress of adolescents with reading difficulties. Specifically, retell might more accurately reflect the understanding of students in special education than standardized reading comprehension instruments or curriculum-based measures that assess the number of words read correctly per minute. Whereas students in special education and those identified as “below average” in ability performed significantly worse on standardized reading comprehension and ORF assessments, there were no observed differences in their retell performance. Additional research is needed to determine whether a more technically sound retell instrument could help distinguish students who are dysfluent readers but adequate comprehenders.

Limitations

In addition to the overarching limitation imposed by the weak inter-rater reliability of the scoring mechanism, there are several other characteristics of the study to consider when interpreting the results. Each will be addressed in the following sections.

Timed tests. Nearly all the data were derived from assessments administered under timed conditions, which place added pressure on speed of reading (Cutting & Scarborough, 2006; Jackson, 2005). This created a conceptual difficulty in specifying the baseline model in that it was difficult to separate measures of reading accuracy (the word identification construct) from measures of reading rate (the fluency construct). Hence, there were initially cross loadings that were later parsed to individual constructs, and measures that had to be moved around within the model or taken out of the model altogether. What is not known is how much the timed nature of the assessments might be influencing the correlations among the measures and constructs. For example, it is possible that fluency might have a weaker relationship to both word identification and comprehension if all the predictors for the latter two constructs were un-timed and, therefore, more inherently independent of reading rate.

Although students were not limited in the amount of time they had to produce their retell responses, the retells were based on the amount of text read in the one-minute allotted for the TMSFA passage reading fluency subtest. For some students, this could have been one or two paragraphs of the passages. Raters were trained to evaluate the responses against only that information read, and the lack of group differences on retell performance suggests this was carried out. However, the different lengths of text associated with each retell response could be contributing to the low inter-rater reliability. The first scorer was present with the student during testing and knew exactly at what point in the passage each student ended in the minute timeframe. The second rater relied upon transcribed responses and the recorded words correct per minute, which may not be as accurate as in-the-moment scoring.

Scoring issues aside, it is important to remember that retell as utilized here occurred within an ORF approach. This is consistent with the procedures in some previous studies (Burke & Hagan-Burke, 2007; Riedel, 2007; Roberts et al., 2005). However, many other studies provided students an unlimited amount of time to read and/or allowed students to read the passages silently (e.g., Best et al., 2008; Doty et al., 2001; Fuchs & Fuchs, 1992; Gagne, Bing, & Bing, 1977; Gambrell et al., 1991; Mason et al., 2006; Pearman, 2008; Richgels et al., 1987; van den Broek et al., 2001). The outcomes under those conditions could be different than what is reported for the retell protocol used in this study.

Social validity. One of the rationales for adding a retell component to an ORF assessment is to address the issue of social validity (Roberts et al., 2005). Some teachers and researchers have expressed concern that the number of words a student reads correctly in a minute is not truly reflective of whether or not that student understood the text (Applegate, Applegate, & Modla, 2009; Goodman, 2006; Shinn et al., 1992). This is related to Messick's (1989) validity model that includes considerations of relevance and utility with construct validity. Unfortunately, no social validity data was available for analysis in this study. Information on how accurate the students' teachers thought the retell scores were might have helped to better explain the inter-rater reliability problem or the finding of no group differences on retell performance. Moreover, data on teachers' confidence in the retell scores versus the ORF scores as indicators of comprehension would help determine whether improving the retell component was worth the challenge.

Sample characteristics. The extant database used in this study was compiled from a purposefully selected sample of seventh- and eighth-grade students. It represents a population that was predominately economically disadvantaged and of African-American or Hispanic heritage. This presents two limitations on the results.

Generalizability. Even though the sample was normally distributed on all measures of reading and intelligence, the findings presented here might not generalize to other settings in which those administering and scoring retells were accustomed to working with students from different backgrounds. Just as the results are specific to the format of the retell used to collect the data, they are also specific to the population from whom the responses were elicited. From what can be determined, the extant literature on retell is reflective of a diverse group of students, but more were from lower grade levels (Reed & Vaughn, manuscript under review). The few studies that included students in grades 7-8 typically had few participants or represented a specific group (e.g., students with LD, middle class). This study, therefore, adds knowledge to the field about the retell performance and model of reading competence among high poverty, high minority students in the middle grades who were at a range of ability levels.

Student background knowledge. The other limitation imposed by the sample characteristics concerns the amount and kind of background knowledge the participants brought to the assessment tasks. Previous research has demonstrated a negative relationship between poverty and students' vocabulary (Hart & Risley, 1995), content knowledge (Vellutino et al., 1996), and cognitive and verbal ability (Smith, Brooks-Gunn, & Klebanov, 1997). Structural equation models have consistently found that background knowledge makes a significant contribution to reading comprehension and is correlated to vocabulary knowledge (Cromley & Azevedo, 2007; Kintsch, 1988; Perfetti, 1985).

Because background knowledge also is believed to be a significant predictor of retell performance on expository passages (Best et al., 2008), it is possible that the results presented here are specific to the predominately economically disadvantaged sample. Two of the three passages on which retell responses were gathered were labeled "expository." As previously discussed, the passages were more similar to informational narratives, but they still required a greater depth of knowledge about history, geography, and culture than the passage labeled “narrative.” For students with more background knowledge, these passages might have been better predictors of comprehension.

Recommendations for Future Research

Questions still remain about the reliability and validity of retell as a measure of adolescents’ reading comprehension that would compliment traditional ORF progress monitoring instruments. In its current state, the retell component of the TMSFA was not technically adequate, but it might reflect the holistic approach to scoring responses or the context of occurring within an ORF approach rather than the utility of retell in general. Future research might attempt to replicate the CFA with retells scored using a more reliable mechanism. A quantitative approach to evaluating responses might improve the functioning of retell in the model and reveal stronger correlations to the other measures. This would also allow for a comparison of the quantitative and holistic/qualitative scoring mechanisms with respect to potential covariates. If a quantitative method improves the functioning of retell at the expense of introducing differential item functioning, other studies might explore whether these group differences can be mitigated with follow-up prompting to elicit more of the desired information.

In addition, future research might compare retell within ORF to retell administered after reading completed silently and/or for an unlimited amount of time. Replicating the analysis by including retell condition as a potential covariate would yield important information about the optimal administration procedures.

Of course, replications of the CFA conducted here assume that the specified three-factor model of reading competence is most parsimonious for students in seventh-

and eighth-grade. Because the model demonstrated mediocre to adequate fit, it might first be necessary to test different models. For example, fit could be improved by making the model conditional with student characteristics as covariates. Alternatively, reading competence might be better modeled with four latent constructs: word identification, fluency, vocabulary, and comprehension.

To address questions of social validity (Kazdin, 1977; Wolf, 1978), future research should gather information on teachers' perceptions of standardized comprehension test results, ORF scores, and retell ratings. When presented with various data on students' reading performance, it would be important to know how much credence teachers give to each type of test. Presumably, the data teachers believe the most will serve as the basis for the instructional decisions they make. It might not be worthwhile to pursue improvements to retell tasks if teachers already had confidence in and were relying upon other more psychometrically sound measures. Conversely, if teachers are disregarding data from instruments with high technical adequacy in favor of a retell, it would be critical to advance this line of research and ensure more valid information was available.

As with any reading assessment, the value of a retell task lies in what it can reveal about a students' abilities that would be useful for planning and evaluating instruction. Even if subsequent studies can substantiate that a reliably scored retell is a strong predictor of comprehension that is trusted by teachers, more guidance is needed in how to use the retell scores to make instructional decisions. It is not clear how teachers would interpret retell responses to group students, plan targeted skills instruction beyond retelling information, or connect to inferential comprehension. In summary, there is a great deal yet to learn about the utility of retell assessments.

Conclusion

The purpose of this study was to examine the validity of the retell task included in the TMSFA within a confirmatory factor analysis (CFA) framework. Retell made a small but significant contribution to comprehension in a three-factor model of reading, was more closely associated with other measures of comprehension than of fluency or word identification, and did not exhibit differential item functioning by student characteristic (e.g., socioeconomic status, primary language, ability level). However, its low factor loading, high residual variance, and low inter-rater reliability make it a questionable measure of the construct. This study contributes to the understanding of reading competence in the middle grades and how to gauge students' comprehension ability.

Figure 3.1

Baseline Model

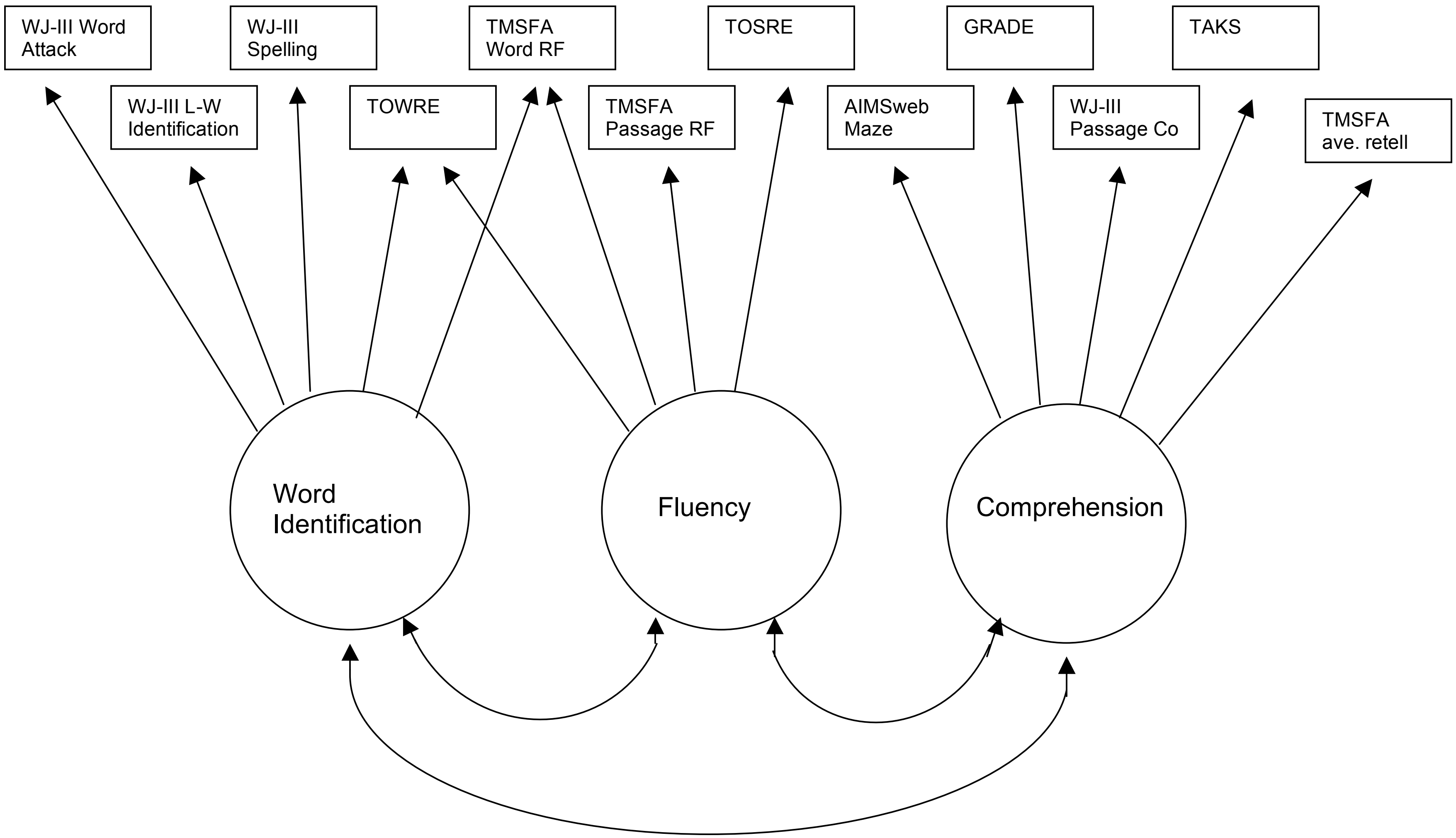


Figure 4.1

Revision 1

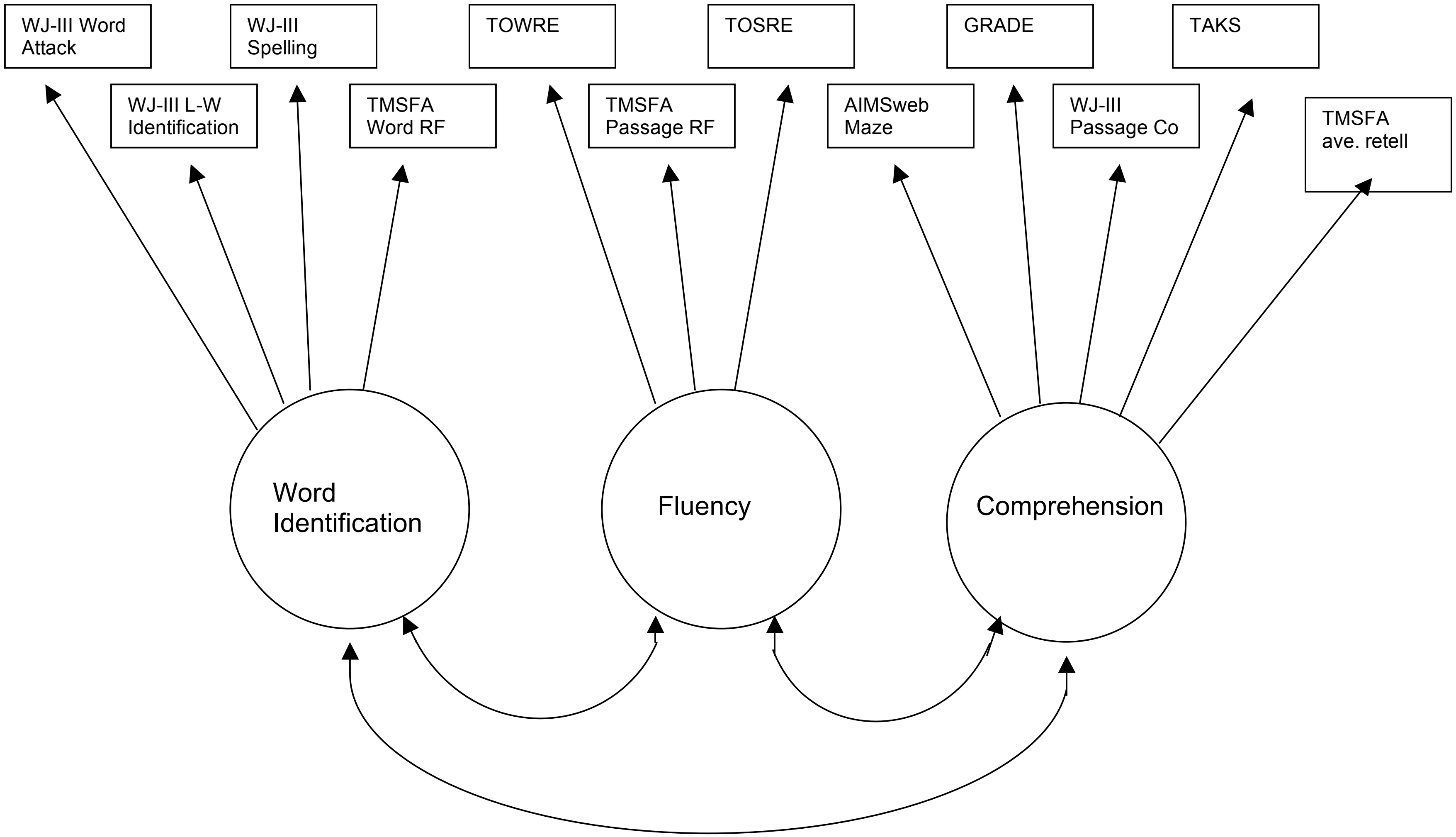


Figure 4.2

Final Model

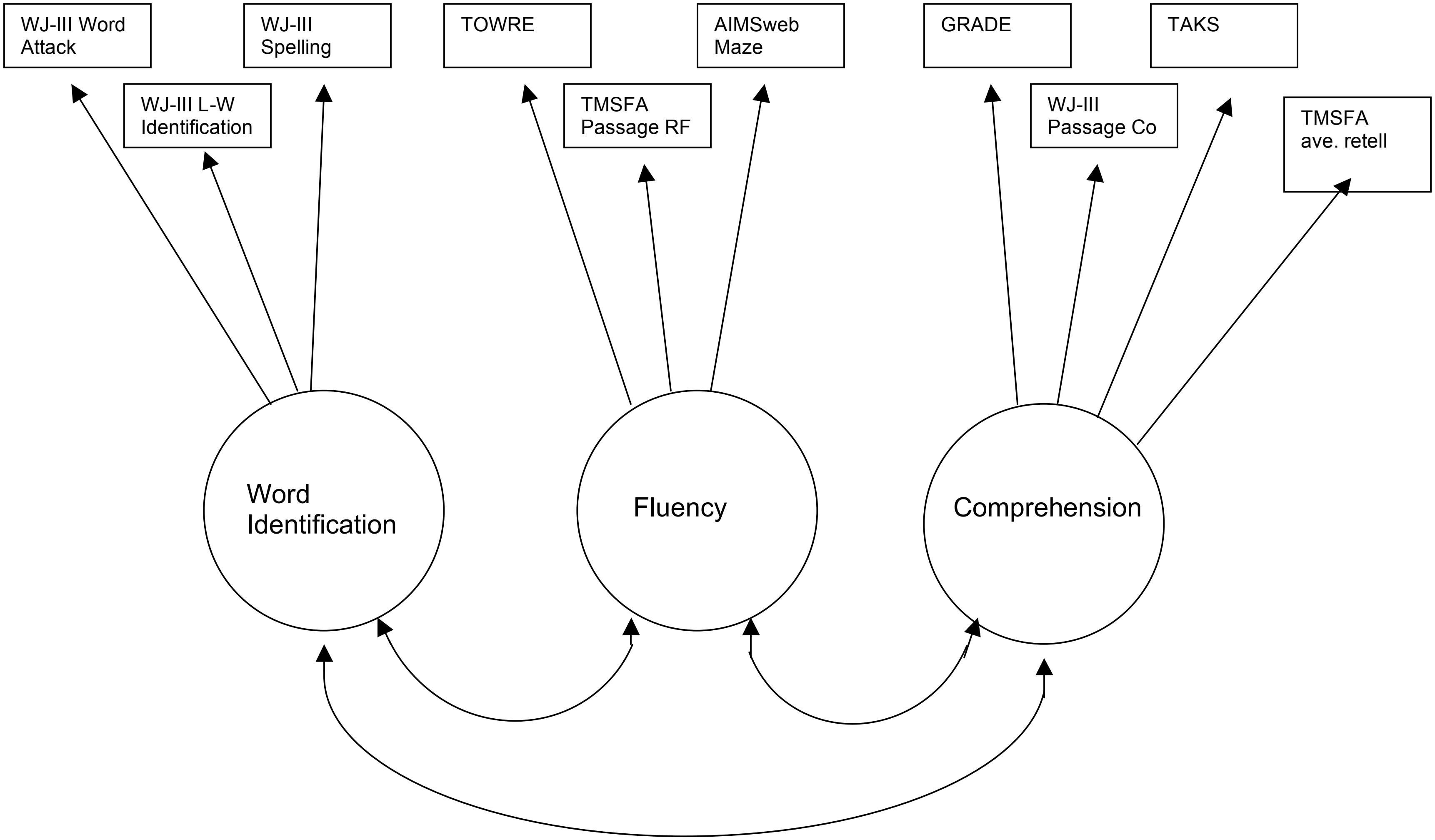
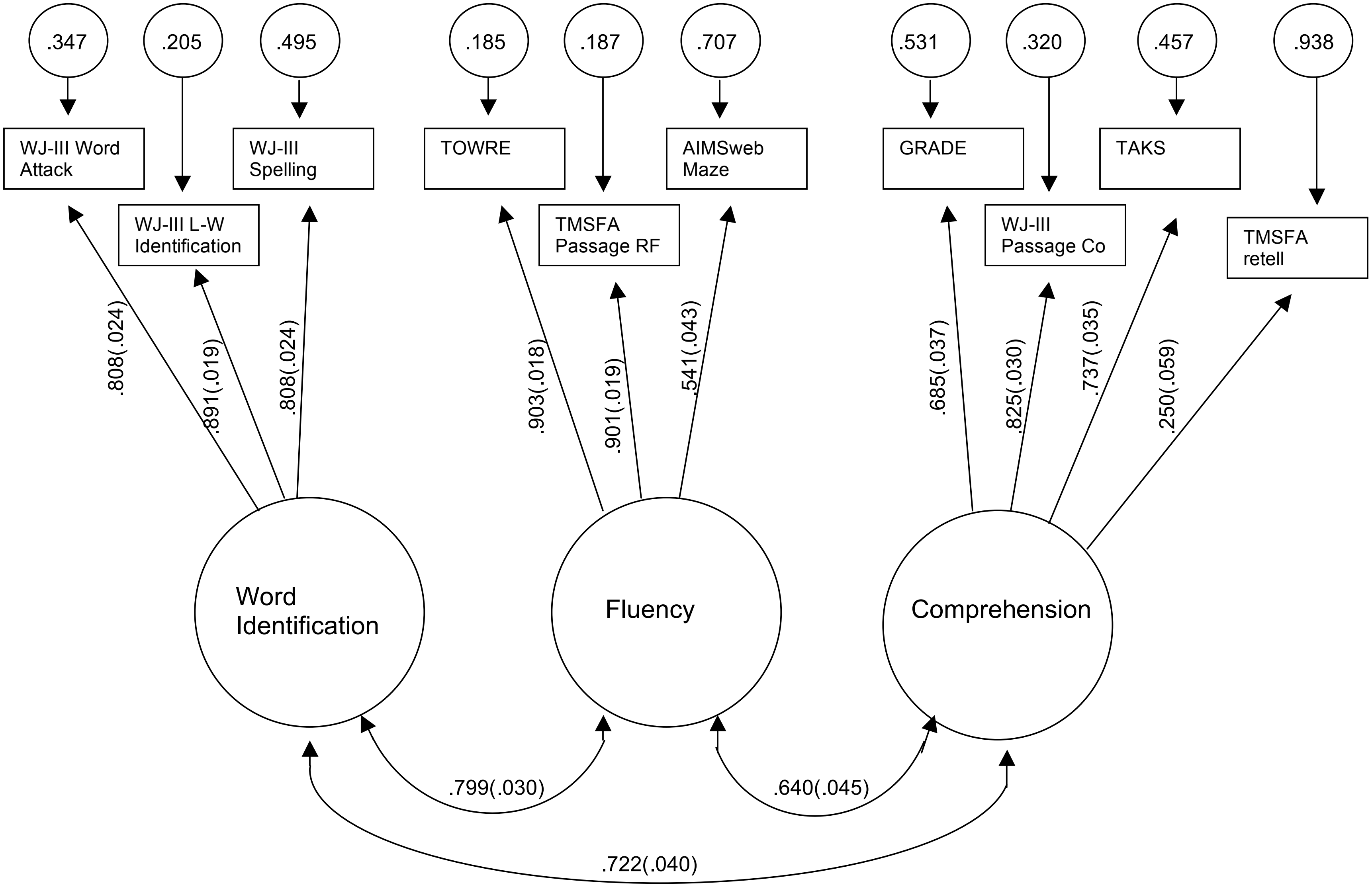


Figure 4.3

Final Model: Estimates, Standard Errors, and Residual Variance



Appendix A: Retell Synthesis (Reed & Vaughn, manuscript under review)

Retell as an Indicator of Reading Comprehension

Studies investigating the skill deficits of those who struggle with reading indicate that word identification, fluency, and comprehension are often distinct categories of ability (Catts, Adlof, & Weismer, 2006; Fletcher, Lyon, Fuchs, & Barnes, 2007; Valencia & Buly, 2004). Students may exhibit difficulty in only one domain (identified by Catts et al., 2006, as *specific deficit in word reading* or *specific comprehension deficit*), or they may struggle with a combination of skills (referred to as *mixed deficit*). Regardless of the number or type of reading abilities concerned, all affected students will demonstrate poor understanding of text. This is often interpreted as consistent with the *simple view* of reading (Gough & Tunmer, 1986), which conceives of reading comprehension as a result of both decoding and language comprehension. However, there is some evidence that unique variance in reading comprehension ability is also contributed by reading speed (Cutting & Scarborough, 2006), verbal ability (Savage, 2006), relevant background knowledge in expository texts (Best, Floyd, and McNamara, 2008); or awareness of the relationship and relative importance among the ideas (Carlisle, 1999). Moreover, it is believed the contribution of decoding diminishes somewhat as students become older (Keenan, Betjemann, & Olson, 2008) and better able to rely upon compensatory strategies, such as context clues (Savage, 2006).

Given the potentially large number of component skills, assessing the reading comprehension of students is anything but “simple.” An instrument designed to measure only one type of ability (e.g., word identification or vocabulary knowledge) might fail to identify those students whose reading difficulty rests largely in another domain. Similarly, instruments of overall comprehension are problematic in that they do not measure equivalent cognitive processes (Cutting & Scarborough, 2006; Keenan et al., 2008; Spooner, Baddeley, & Gathercole, 2004), particularly if they differentially employ narrative and expository texts (Best et al., 2008).

Rationale and Research Questions

It has been suggested that a retell prompt might be added to an oral reading fluency (ORF) measure as a means of improving the validity of the assessment without diminishing its efficiency (Roberts, Good, & Corcoran, 2005). In comprehension research, the skills of retelling, recalling, summarizing, and paraphrasing are considered distinct skills that require differing levels of complex thought and different degrees of telling or transforming knowledge (Kintsch & van Dijk, 1978; Scardimalia & Bereiter, 1987). Within studies examining retell as a measurement tool, however, these skills are treated almost interchangeably (Duffelmeyer & Duffelmeyer, 1987). Depending upon the instrument or study, “retell” and “recall” could be used to elicit main ideas, summaries of the content, or a thorough restatement of the passage. In the most common approach, students are asked to read a passage, either silently or orally, and are then prompted to tell or write about the passage in their own words without referring back to the text.

Retells are among the more popular elements of reading comprehension assessment (Fuchs et al., 1988; Nilsson, 2008; Talbott, Lloyd, & Tankersley, 1994), but they have several limitations. Notably, students with learning disabilities (LD) tend to perform more poorly on retell tasks than students without LD, even after controlling for topical vocabulary and passage comprehension (Carlisle, 1999). Hence, it is possible that retell could not accurately convey a student's comprehension. There are several possible explanations for this. To retell a passage verbally or in writing, the student must be able to recall information, organize it in a meaningful way, and possibly draw conclusions about the relationships among the ideas (Klingner, 2004). Producing the retell is highly dependent upon the student's productive language abilities (Johnston, 1981). In fact, oral retell performance reliably differentiates adults with and without aphasia, an impairment in the ability to produce or comprehend language resulting from brain injury (Ferstl, Walther, Guthke, & Yves von Cramon, 2005; McNeil et al., 2001; Nicholas & Brookshire, 1993).

Moreover, the quality, accuracy, and completeness of students' written retells are related to their transcription fluency, or the number of letters the students can write in one minute (Olive & Kellog, 2002; Peverly et al., 2007). Some have suggested that assessing comprehension with open-ended questions, such as a retell prompt, makes it difficult to distinguish among difficulties at the level of input, retrieval, expression, or some combination thereof (Johnston, 1981; Spooner, Baddely, & Gathercole, 2004). Others have cautioned that socioeconomic status and cultural-linguistic differences might influence student performance on comprehension tasks that require oral language processing (Snyder, Caccamise, & Wise, 2005). Unfortunately, no known studies have explored this with respect to students' retell performance or the teachers' judgments of students' retell ability.

Differences in retell performance might also be due to maturation and/or measurement artifacts. Results of a study with third- and fifth-grade participants, indicated that only the older pupils benefited from practice effects (Otto, Barrett, & Koenke, 1968). This is notable in that, at the grade levels where studies suggest students' ORF results begin to asymptote (Fuchs et al., 2001; Stage & Jacobsen, 2001), their retell ability improves. Familiarity with the content of the passage, however, seems to benefit the retell performance of students across grade levels (Leslie & Caldwell, 2006; Otto et al., 1968). Similarly, text type was found to be influential for students at various grade levels, who recalled significantly fewer ideas from expository versus narrative selections (Leslie & Caldwell, 2006).

There is, however, disagreement as to whether the length of text can influence retell performance. Some researchers believe that construct validity is only possible if the measure relies upon "selections of sufficient length and complexity to allow children to make constructive connections across text, similar to texts encountered in classrooms" (Snyder, Caccamise, & Wise, 2005, p. 40). Whereas, other researchers believe that longer, more complex texts reduce the richness of the retell and encourage students to merely provide a main idea or gist of the passage (Leslie & Caldwell, 2006).

In addition to measurement artifacts, there are concerns about the psychometric properties of retells. Reportedly, there is no uniform scoring procedure across instruments (Nilsson, 2008), and the inter-rater reliabilities are often weak (Klesius & Homan, 1985). These concerns combined with large score fluctuations have caused retell tasks to be considered unsatisfactory for monitoring student performance over time (Fuchs & Fuchs, 1992). Yet, retell tasks remain an appealing compliment to ORF measures due to their efficiency, equivalency of format across passages, reliance on active reconstruction of text, and relevancy to comprehension instruction (Hansen, 1978; Roberts et al., 2005). In addition, an informal assessment using a retell task was shown to be much less sensitive to students' decoding ability than other standardized measures of comprehension (Keenan et al., 2008). A retell component, therefore, has the potential to detect other instructional areas of need that might be missed by the ORF measure alone. This information would be highly useful in planning reading interventions.

However, no systematic review of the practice has been conducted to determine if a retell component contributes unique, valid, and reliable information about students' reading comprehension. Therefore, this descriptive synthesis seeks to address the following questions: (a) What existing research has examined the validity of retell as a comprehension measure?, (b) How have existing assessments of reading comprehension incorporated a retell procedure?, and (c)What is the reliability and validity of the retell component in existing assessments?

Method

To identify studies of the reliability and validity of retell measures, the Academic Search Complete, PsycINFO, ERIC, and MEDLINE electronic databases were searched using the following descriptors: *retell** OR *free recall* OR *main idea* AND *read** *comprehen**. No limitation was set on the initial date of publication because there was no reason to believe that the age of the study would be relevant to ascertaining the technical adequacy of a retell protocol. However, a search end date of 2008 was imposed. Despite reports that “an overwhelming number of studies investigating reading comprehension have used free recall as a dependent variable” (Gambrell, Pfeiffer, & Wilson, 1985, p. 216), these were not easily identifiable in electronic searches because “retell” and “recall” were infrequently named in abstracts or listed among the key words. To locate as many potential studies as possible, an ancestral search was conducted using the reference lists of articles and technical reports on reading comprehension assessments. The initial search netted approximately 300 abstracts. Based on the recommendations of a reviewer, a second wave of searching was conducted to more thoroughly ensure a complete examination of the literature.

All identified abstracts identified were evaluated on the basis of the following criteria: (1) Article was published in a peer-reviewed journal. Dissertations and conference papers were excluded due to difficulties in reliably obtaining such manuscripts, particularly those dating back more than 10 years; (2) Participating students were in grades K – 12. Studies with older or younger students were included if they also sampled students within the target (school-age) range; (3) The language in which participants were tested was English; (4) Students in all conditions were assessed using connected text as opposed to graphic displays, wordless picture books, rebuses or other symbolic representations; (5) Participants were not identified on the basis of sensory impairments; and (6) Results reported sufficient information on the reliability, validity, and/or utility of retell as an indicator of reading comprehension. A total of 26 studies were judged to meet all criteria for inclusion in the synthesis.

Existing assessments that include a retell measure were also identified in the ancestral search of articles on reading comprehension assessments. In addition, the databases of test publishers (e.g., ProEd, Pearson, McGraw Hill, Kendall Hunt) were manually searched for Informal Reading Inventories (IRIs), which the extant literature indicated were the most common type of comprehension assessment to include a retell component. The instruments identified were evaluated on the basis of the following criteria: (1) Measure was designed for students in grades kindergarten - 12. Instruments intended for use with younger students or adults were included if the assessment is also intended for students within the target range (e.g., *BADER Reading and Language Inventory,* Bader & Pearce, 2009; *Classroom Reading Inventory,* Silvaroli & Wheelock, 2004); (2) Measure included a stated protocol for administering either an oral or written retell; (3) Measure is not tied to a commercial reading program (e.g., *Houghton Mifflin Leveled Reading Passages Assessment* in the Houghton Mifflin reading series) unless the instrument has been used in a study of retell identified for inclusion in the first part of this review (e.g., *Vital Indicators of Progress* in the Voyager Passport reading intervention); (4) Measure is commercially or publicly available in all states. A total of 12 assessments were judged to meet the criteria for inclusion in the synthesis.

Data Analysis

Coding procedures. Studies of retell measures were coded for elements pertinent to this descriptive review. The code sheet included the grade level(s) and characteristics of participants, whether the passages were read orally or silently, the purpose of including a retell measure in the study, whether the retell was provided verbally or in writing, the initial prompt given to students as well as any follow-up prompting, the scoring procedure, and findings related to the reliability, validity, and/or utility of the retell measure. The information from all code sheets was organized in Table A1 to summarize the studies. To facilitate the interpretation of the findings, studies were grouped according to the purpose for which the retell measure was included (validation study, reliability study, comprehension outcome measure).

A different, but related, code sheet was used to analyze existing assessments with a retell component. This sheet included the grade level(s) for which the instrument is intended, whether the retell is to be provided verbally or in writing, genre of stimuli (i.e., narrative, expository, or both), whether the retell is asked after silent reading or oral reading, the initial prompt given to students as well as any follow-up prompting, the scoring procedure, descriptions of the norming sample, and the reported reliability and validity of the retell portion of the instrument. The information from all code sheets was organized in Table A2 to summarize the existing assessments incorporating retell measures.

Results

Retell Study Features

Of the 26 studies that met the selection criteria (summarized in Table A1), less than half (n = 11) were published within the last decade (1998 – September 2008), a time during which studies of ORF measures have proliferated. The remaining 15 studies were published in a two-decade period spanning from 1977 to 1997, with 11 of those appearing in journals between 1982 and 1992.

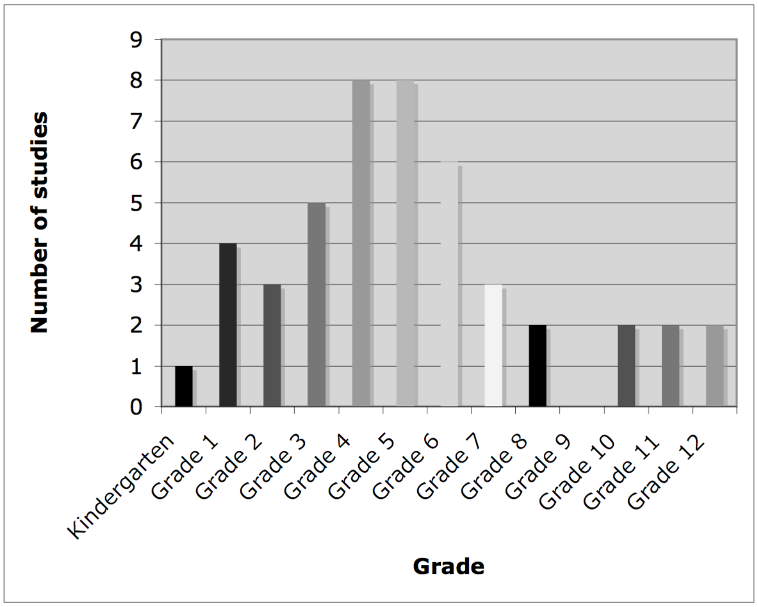
Sample characteristics. Although a total of 3,424 students participated in studies of retell measures, this sum is inflated by a single study that included 1,518 students (Riedel, 2007). Excluding that study, participant counts ranged from 9 (Mason, Snyder, Sukhram, & Kedem, 2006) to 240 students (van den Broek et al., 2001). The overwhelming majority of studies (n = 20) had less than 100 participants. A variety of ability levels were represented in the aggregate data. However, individual studies might have focused solely on students with disabilities (e.g., Fuchs & Fuchs, 1992; Fuchs et al., 1988) or students considered average or above-average readers (e.g., Gagne, Bing, & Bing, 1977; Gambrell, Pfeiffer, & Wilson, 1985; Loyd & Steele, 1986; Rasinski, 1990).

The selection criteria for this review allowed for students in kindergarten through grade 12, and only grade 9 was not included in any of the studies identified. Grades 4 and 5 were included more often across studies; however, first-graders represent the single largest population in the aggregate data due to the large sample in the Riedel (2007) study. Twelve studies targeted multiple grade levels, so the data depicted in Figure A1 reflect overlapping studies.

Most studies included comparable numbers of males and females, with the exception of one study that included only boys (Fuchs, Fuchs, & Maxwell, 1988). Other student characteristics were less consistently reported across studies. Some (n = 6) did not report the ethnic composition of participants. Twelve studies reported information on students' ethnicities and backgrounds, reflecting a wide range in the proportion of study participants from diverse populations. Two studies reported predominately (85% or greater) Caucasian samples; 6 studies had between 30 and 50% ethnically diverse samples; 2 studies referred more generally to students being of diverse backgrounds; and 2 studies had predominately African-American participants. Only 1 study (Riedel, 2007) referred to the inclusion of, at least, some English language learners (ELLs). Whereas, two studies (Gambrell, Pfeiffer, & Wilson, 1985; Pearman, 2008) excluded ELLs.

Figure A1

Number of Studies in Which Each Grade Level Was Included



Purpose for including retell in the study. Only 8 of the 26 studies specifically sought to determine the criterion, concurrent, or predictive validity of a retell measure, usually by correlating it to other formal and informal assessments of reading. Many of these 8 studies also provided data on inter-rater reliability for the scoring procedures, but three other studies focused more directly on issues related to the reliability of retell measures. The latter studies examined whether practice with retell (Gambrell et al., 1991) or a stated goal or instructional focus for reading (Gagne et al., 1977; Gambrell & Jawitz, 1993) influenced students' retell performance. The remaining studies included in this review did not directly address issues of the validity or reliability of retell instruments. Instead, retell protocols were implemented as a means of assessing students' comprehension. Eight of those studies examined the influence of text genre, particularly elements or structure of and strategies for reading expository text, on students' comprehension as determined by retell performance. Four studies (Doty, Popplewell, & Byers, 2001; Kouri & Telander, 2008; Pearman, 2008; Wright & Newhoff, 2001) used retell to determine whether there was a difference in students' comprehension of stories delivered in different formats (electronic vs. print, sung vs. told, read vs. told), and one study (van den Broek et al., 2001) assessed students' retell performance when provided probing questions on the causal relations in a story. The final two studies in the category of "comprehension outcome measure" examined the influence of instruction or practice on retell performance (Gambrell et al., 1985; Popplewell & Doty, 2001).

Retell measure format. The format of the retell measures employed in the studies differed in three primary ways. First, students could have been reading orally (n = 4), silently (n = 7), or in combination ([n = 3] Fuchs et al., 1988; Mason et al., 2006; Pearman, 2008). In some studies students listened to the teacher or examiner read ([n = 2] Horowitz & Samuels, 1985; Moss, 1997), and in others the author(s) did not identify the type of reading conducted prior to the administration of the retell ([n = 2] Carlisle, 1999; Loyd & Steele, 1986). The second variation in the format concerned the type of text read prior to the administration of the retell. Passages could have been expository (n = 8), narrative (n = 5), or both (n = 1). Four of the studies did not specify the text genre. Finally, students could have been asked to provide their retell orally (n = 10), in writing (n = 5), or both orally and in writing (n = 3). Table A3 depicts a matrix of these format variations to better depict the types of studies conducted.

Findings from Retell Studies

Correlations of retell to other reading assessments. Seven studies representing culturally and economically diverse student populations provided the correlation of retell scores to other measures of reading ability. The strength of the correlations discussed in this section will be judged conservatively using the following scale of absolute correlation coefficient values (Williams, 1968): (a) 0.00 – 0.30: weak; almost negligible relationship, (b) 0.30 – 0.70: moderate correlation; substantial relationship, and (c), 0.70 – 1.00: high/strong correlation; marked to perfect relationship. The more conservative estimations of the strength of correlation was used here because the study was formative. A more stringent parameter would increase the confidence that the data represents reality.

Four studies identified were specifically designed as validity studies, two examined influences on students' expository and/or narrative text comprehension, and one compared students' retell scores to their scores on open-ended, factual questions. The results spanned the grade levels represented in the corpus of studies and demonstrated a rather consistently moderate correlation between recall and assessments of overall reading ability, letter-word identification, academic knowledge, vocabulary, reading comprehension, and fluency. For the large sample of first-grade participants (Riedel, 2007), oral retell results were more moderately correlated (*r* = .39 - .69) to the vocabulary and comprehension subtests of two standardized measures of reading, GRADE and TerraNova. A study of third-graders' comprehension of narrative versus expository text comprehension (Best, Floyd, & McNamara, 2008), revealed that free and cued oral recalls of both narrative and expository text were moderately correlated (*r* = .36 - .58) with the Woodcock-Johnson academic knowledge test. Narrative free and cued oral recall, as well as expository free oral recall, were moderately correlated with the Woodcock-Johnson letter-word identification test (*r* = .48 - .64).

One exception to the pattern of correlations was found in a study of third- and fifth-graders (Rasinski, 1990) where oral retell was not significantly correlated with researcher-developed measures of phrasing ability. Retell was, however, moderately correlated with both miscue and reading rate (*r* = .38 - .52). It should be noted that this is the only study for which the retell scoring procedure could not be determined, so the basis of the correlation calculation is unknown. For all other studies reporting correlation data, retells were scored by a numerical count of the words or pre-determined idea units/propositions the student included (see section on inter-rater reliability for more information).

Stronger correlations between retell and fluency were found in the Fuchs et al. (1988) study with slightly older students. Retell scores of fourth- through eighth-graders were highly correlated to an ORF measure (mean *r* = .75) and moderately to highly correlated with the Stanford Achievement Test (SAT-7) reading comprehension and word study subtests (*r* = .47 - .82). . This is one of only two studies that incorporated both oral and written retells, so it is noteworthy that the researchers found consistently and significantly higher correlations for the written recalls than those for oral recalls. Yet, ORF scores were more highly correlated with the SAT-7 than any of the other measures included in the study. Moreover, ORF had higher correlations with the SAT-7 reading comprehension subtest than the word study subtest.

In another study of upper-middle grades students (Carlisle ,1999), oral recall scores of the sixth- and eighth-grade participants were moderately to highly correlated to scores on researcher-developed sentence verification (*r* = .50 - .74) and science vocabulary (*r* = .49 -.51) tests. Results were similar in a study of fifth- and sixth-grade students (Hansen, 1978). Spearman's rank correlation coefficient revealed the proportion of idea units recalled was moderately to highly correlated with performance on open-ended, factual comprehension questions (ρ = . 46 - .77).

The study with the oldest students (Loyd & Steele, 1986) found weak to moderate correlations between eleventh- and twelfth-graders' written recall of idea units and SRA reading comprehension and language arts mechanics scores (*r* = .28 - .56). Holistic coherence scores on those written retells were also all in the weak to moderate range (*r* = .11 - .39). In sum, across all grade levels and test types in the 6 studies providing validity data, retell measures tended to be moderately correlated with both formal and informal assessments of reading ability.

Predicting and monitoring student progress in reading comprehension. Six of the 26 studies provided data on the predictive validity of retell measures or the adequacy of retell scores for tracking student progress over time. For first-graders (Riedel, 2007; Roberts et al., 2005), results indicate that ORF scores are the best predictor of reading performance. Overall, adding oral retell scores only improved the predictive accuracy by 1% or less than ORF alone. For some students, however, retell performance was notably inconsistent with their ORF performance. It is important to note that in neither of the first-grade studies was it possible to determine whether narrative or expository passages were used. The measures include both genres, but the particular selections used as stimuli in the research were not specified.

In a study comparing third-graders' oral recall of narrative and expository passages (Best et al., 2008), decoding skill was the strongest predictor of narrative recall, but background academic knowledge was the stronger predictor of expository recall. In addition, Shinn et al. (1992) found the residual variance of written retells for narrative passages to be so high (74%) that "they did not function well as measures of reading constructs for fifth-grade students" (p. 470). A one-factor model of narrative text reading was most parsimonious at grade 3, with ORF demonstrating the highest factor loading (.90). At grade 5, a two-factor model of narrative text reading was most parsimonious, and ORF no longer demonstrated the highest factor loading. In the two-factor model, ORF loaded on decoding, and written retell loaded on reading comprehension.

Only 2 studies explored the consistency or stability of students' retells, which would indicate the adequacy of such measures for tracking student progress. Fuchs and Fuchs (1992) found that a written retell measure administered to fourth- and fifth-graders twice weekly over 15 weeks produced instable scores which, when graphed for monitoring purposes, produced small average slopes in relation to the average standard error of estimate. Therefore, the researchers concluded the retells (scored quantitatively) were difficult to use for interpreting students' growth in performance. In a study of fourth-graders, oral retell scores were inconsistent across the multiple baseline probes administered over a 26-week period of multiple strategy instruction related to retell (Mason et al., 2006). The results of these studies reflect a narrow range of grade levels (4 – 5) and a limited number of participants (n = 47) from the aggregate sample of studies included in this review. In fact, no studies of retell measures for the purposes of predicting or monitoring progress were conducted with students above grade 5.

Inter-rater reliability. Only 1 of the 26 studies (Rasinski 1990) did not specifically describe how the retells were scored. A total of 18 studies provided inter-rater reliabilities with an overall range of 72% to 100% agreement. Higher agreements were noted for some written retells (Fuchs & Fuchs, 1992; Fuchs et al., 1988; Loyd & Steele, 1986; Mason et al., 2006; van den Broek et al., 2001) and for scoring procedures that relied upon the number of pre-determined idea units, story structure elements, or propositions recalled in oral retells (Best et al., 2008; Gambrell et al., 1991; Gambrell, Pfeiffer, & Wilson, 1985; Horowitz & Samuels, 1985; McGee, 1982; van den Broek et al., 2001; Wright & Newhoff, 2001; Zinar, 1990). Lower inter-rater reliabilities (generally below .90) were noted for scale scores of writing coherence (Loyd & Steele, 1986) or of the match between the composition's organizational structure and that of the text (Richgels, McGee, Lomax, & Sheard, 1987); holistic scores of orally recalled story elements (Gambrell & Jawitz, 1993; Pearman, 2008; Popplewell & Doty, 2001); and holistic scores of overall retell quality (Mason et al., 2006).

Indeed, the most common method for scoring students' retells involved numerical counts of words, idea units, propositions, or story elements. Quantitative procedures were used in 19 of the studies, and Fuchs et al. (1988) found no significant differences among scoring by number of words, percent of content words matching original text, or percent of predetermined idea units. This is particularly noteworthy because the study employed both written and oral retells after both oral and silent reading. However, only narrative passages were administered, and the students were allowed 10 minutes to respond with repeated prompting if they paused for 30 seconds yielding a longer period that involved more examiner cuing than was reported in other studies of oral retell.

What was not addressed in the studies was interpretation of the numerical counts. In some cases, the counts were converted into a proportion of idea units recalled (e.g., Best et al., 2008; Gambrell et al., 1991; Hansen, 1978; McGee, 1982; Richgels et al., 1987; van den Broek et al., 2001; Zinar, 1990). However, little guidance was provided for making conclusions about what a desirable percentage of recalled idea units might be, or what percentage might indicate comprehension difficulty. Hansen (1978) noted that even on-grade-level readers recalled only about one-third of the idea units. In comparing third- and fifth-grade students, McGee (1982) found on-level third-graders recalled, on average, less than 20% of the main ideas and less than 30% of the details. Whereas, average achieving fifth-graders recalled, on average, about 50% of the main ideas but less than 40% of the details. Fifth-grade students identified as below-level readers recalled about 30% of both main ideas and details.

In all studies with quantitative scoring techniques, inter-rater reliability was based on the count itself, not on a translation of the tally or proportion to categories of "better" or "weaker" reading comprehension skill. No studies using either quantitative or holistic approaches to scoring examined teacher or student factors that might influence the scoring and/or interpretation of results.

Prompt variation. Although only two studies addressed the influence of a stated reading goal or focus on students' retells (Gagne et al., 1977; Gambrell & Jawitz, 1993), there was remarkable variety in the prompts and other cuing provided to students across all studies. Less than half (n = 10) of the studies provided complete verbatim accounts of the initial prompt and any follow-up prompting or cuing used to elicit the retell from students. From what could be determined, students might have been told simply to compose a summary from memory (Fuchs & Fuchs, 1992; van den Broek et al., 2001), recall as much as they can about what they read (Carlisle, 1999; Fuchs et al., 1988; Horowitz & Samuels, 1985; Rasinski, 1990; Riedel, 2007; Zinar, 1990), paraphrase the passage in their own words (Loyd & Steele, 1986), retell the passage (Hansen, 1978; Shinn et al., 1992; Wright & Newhoff, 2001), retell the story as if telling it to a younger student (Gambrell et al., 1991), list 10 facts from the passage (Gagne et al., 1977), tell a friend as many details as possible about what they read (Best et al., 2008), write the story or retell the book as if telling it to someone who never heard it (Gambrell & Jawitz, 1993; Kouri & Telander, 2008; Moss, 1997), tell everything they learned as if telling someone who knew nothing about the topic (Mason et al., 2006), tel or write everything they could remember about the passage (McGee, 1982; Richgels et al., 1987), retell all the important ideas (Gambrell et al., 1985), or retell the story, what it was about, and what they remembered about the events (Doty, Popplewell, & Byers, 2001; Popplewell & Doty, 2001). Some studies even allowed the examiner to select among different prompts (Pearman, 2008).

Despite some indication that the term “retell” has been used more frequently by researchers when referring to application with narrative text and “recall” more often with expository, the terms were inconsistently applied in the identified studies when prompting students. The predominant verb used to elicit responses was “tell,” regardless of passage type. When considered in light of how students’ responses were evaluated (see Table A1), the distinctions among *retelling*, *recalling*, *summarizing*, *paraphrasing*, and *identifying the main ideas* are even less distinct.

The numbers and types of allowable follow-up prompts across studies further increase the variability among the procedures employed, including 11 studies without descriptions of follow-up prompting. Based on the 10 studies that did provide this information, students might have been asked scripted questions based on the reading (Doty et al., 2001; Fuchs et al., 1988; Gambrell et al., 1991; Gambrell et al., 1985; Wright & Newhoff, 2001), cued to the major headings in the passage (Rasinski, 1990), encouraged to tell more (Carlisle, 1999; Kouri & Telander, 2008; McGee, 1982; Pearman, 2008; Popplewell & Doty, 2001; Shinn et al., 1992), cued to the major sections of the text (Best et al., 2008), or specifically probed about pre-determined propositions not freely recalled (Zinar, 1990). In some studies, students were both encouraged to tell more and asked scripted follow-up questions (e.g., Hansen, 1978). As with the initial prompts, examiners might also be allowed to select among different follow-up prompts (Moss, 1997).

The number of combinations of initial and follow-up prompts could not be accurately determined given the lack of specific information in many articles, but the number of different means used to elicit the retell data is believed to exceed the total number of studies in this review. If the instructions provided to students prior to reading the passages were also included in this analysis, the variation would be even greater. The results from Gagne et al. (1977) suggest this inconsistency can significantly influence retells. Students in grades 10 through 12 who were provided different reading goals in their preliminary instructions, but the same retell prompt (i.e., write down the first 10 facts that can be remembered from the passage), produced the same amount of information but with qualitatively different content. Students told to read an expository text for discrete and sequential facts about a single topic almost exclusively recalled explicit facts on the topic. Whereas, students told to read the same expository text for 2 to 3 non-sequential, descriptive attributes of a topic almost exclusively recalled attributes.

Similarly, Gambrell and Jawitz (1993) reported that fourth-grade students given instruction to construct mental images and attend to text-relevant illustrations recalled more story propositions and story structure elements than students told simply to "read to remember." Within the story structure elements, the combined imagery and illustrations group performed significantly better on setting, characters, and plot than the control group, and significantly better on characters than the illustration only group. In addition, the combined treatment group students were more likely to provide complete retells, performed better overall on cued recall questions, and better on text-implicit cued recall questions than students in the three other groups. For text-explicit questions, the combined imagery and illustrations group performed significantly better than the “read to remember” control.

Content differences in responses were identified by teachers of fourth- and fifth-graders who expressed concern that scoring students’ retells by word counts did not reflect differences in the quality of the written recalls (Fuchs & Fuchs, 1992). In this latter study, it could not be determined whether students were provided a consistently worded prompt to compose a summary of the passage. No studies explored teacher or student characteristics that might influence the delivery of, or response to, retell prompts.

Other measurement artifacts. Collectively, the greatest number of the 26 studies (n = 11) explored issues related to the testing conditions that might influence student performance. Four of those were concerned with the influence of text genre, particularly the difficulty some students might exhibit with expository passages. Although children as young as first-grade (Moss, 1997) were able to accurately and completely provide main ideas and details in informational trade books, retell information in the proper sequence, and summarize what was most important about what they read, it was reported that student’s responses varied widely. When comparing recall of expository texts with that of narratives, Best and colleagues (2008) found that third-grade students recalled significantly more pre-determined propositions in narratives (10 – 15 versus 4 – 7 in expository text). With neither genre did students include many inferences (1 – 3%).

Similarly, fifth-graders were more likely to include explicitly-stated causal information from expository texts than when the causal information was implicit (Zinar, 1990). Students in that study who were identified as having comprehension difficulties did not include any causal information in their free recalls, but they included comparable amounts of causal information as their higher ability counterparts when probed. As in the Best et al. (2008) study, having students freely recall information from the passage did not produce as much acquired information as when students were specifically cued to provide information, including inferences, they initially left out of their retell. Hence, the use of specific follow-up prompting influenced student performance in quantitative as well as qualitative ways, particularly for students otherwise considered to have difficulty with reading comprehension (Zinar, 1990).

It is also important to note that the type of relationship among ideas targeted in the previous study (i.e., causal) was found by Richgels and colleagues (1987) to be the most challenging organizational pattern for students of all abilities to detect and apply. When probed on their awareness of four expository text structures (collection, comparison-contrast, causation, problem-solution) and recall of texts written in those structures, sixth-graders were most aware of and able to convey information from the comparison-contrast structure. Conversely, students were least aware of or able to produce compositions in the causation structure. The more aware students were of a text's structure, the more likely they were to understand and remember that text as reflected in their written recalls. Furthermore, students demonstrated better-organized recalls in response to passages they read than to structured discussions in which they participated without the aid of a written text or other guide.

The issue of delivery formats for content to be retold was examined more specifically in three studies utilizing only narrative stories. Doty and colleagues (2001) compared second-grade students' retell performance when reading from an electronic medium versus a traditional print book. Research with a small sample of students found no significant differences in students' oral retellings of print versus electronically-based stories. Pearman (2008) found similar results with second-graders. However, when students were separated by reading ability (high-, medium-, low-proficiency), low reading proficiency students' mean retelling scores were significantly higher on electronically-based stories where students could access other supports such as labels, vocabulary definitions, and pronunciations. Changing the delivery format by adding a melody line, so that stories are sung rather than spoken, did not show more promise than the electronic formats. Kinder- and first-grade students demonstrated no significant differences in retell, reading comprehension questions, or mean length of utterance when stories were sung or spoken to them (Kouri & Telander, 2008). Students included a greater number of different words (a higher type-token ratio) when retelling sung stories, but they had greater attention and on-task behaviors when listening to spoken stories.

The 5 studies that explored the influence of instruction in or practice with retelling had somewhat conflicting results. Second-grade students identified as high-, medium-, and low-proficiency readers all demonstrated no significant difference between mean scores on a first- versus second- administration of an oral retell measure (Pearman, 2008). However, second-grade students, who were accustomed to providing retells when conferencing with their teachers about the stories they are reading, performed significantly better on a retell assessment than students who did not practice retelling as part of their literacy instruction (Popplewell & Doty, 2001). Fourth-grade students provided multiple strategy instruction in elements of oral and written retelling demonstrated some improvement in the number of main ideas included (Mason et al., 2006). Although, the improvement was not evident in all of the 9 participants, and those students who did show progress were still inconsistent in the number of main ideas they included. Similarly, fourth-grade students provided opportunities to practice identifying the important ideas and supporting details in passages performed significantly better on written and oral retell tasks than students who practiced illustrating the important ideas (Gambrell et al., 1985). Moreover, the students who practiced retelling had significantly higher free recall scores 2 days after the treatment as compared to the immediate free recall scores of the students who were in the comparison group and practiced illustrating.

Besides the age difference of the participants in the grade 2 and grade 4 studies, there was also a difference in the genre of text. The second-grade participants (Pearman, 2008) were reading narrative passages; whereas, the fourth-graders were reading expository (Zinar, 1990) or informational narrative (Gambrell et al., 1985) passages. In a separate study of grade 4 students (Gambrell et al., 1991), practice effects were also evident in students' oral retells of narrative stories, as well as their ability to answer cued-recall questions. Therefore, the data seem to indicate that the inconsistency in results might be attributable to developmental differences more so than text type. Unfortunately, this cannot be concluded with confidence because none of the available studies examined practice effects at different grade levels with both narrative and expository passages.

Developmental trends were also noted in a study of the effects of causal relation questions on students' written recall performance (van den Broek, 2001). When comparing the performance of fourth-grade, seventh-grade, tenth-grade, and undergraduate college students, younger students tended to recall less information than did older students. In addition, the school-age students generally recalled significantly less information when provided questions during and after reading, with the youngest students showing the most severe impairment in recall with questions used during reading. In contrast, the college students benefited from the inclusion of causal relation questions and recalled significantly more information when provided the questions during reading. Students of all ages included in their recalls significantly more story propositions that were also needed to answer the questions, so memory of and attention to information was universally heightened by the nature of the questions asked during or after reading. Students in grade 10 and college recalled similar amounts of information not specifically probed in the causal relation questions as did students who were not provided any questions. Students in grades 4 and 7 recalled significantly less information not specifically probed in the questions than students in the comparison. Hence, it seems students' sensitivity to potential measurement artifacts varies with age or developmental level. It cannot be determined from available data whether students' cultural-linguistic backgrounds are related to any variations in retell performance.

Ability differences among student participants. Interactions between ability and retell outcomes were addressed in 8 of the 26 studies of retell measures. All utilized only one type of text (narrative or expository exclusively); therefore, no data are available on ability interactions with text type. The youngest participants ([grade 2] Pearman, 2008) were categorized as having high-, medium-, and low-reading proficiency and were assessed with a retell protocol after reading traditional print and electronically-based stories. Although there were no differences in retell performance on the two text formats between students classified as high- and medium-proficiency, students with low-reading proficiency performed significantly better on the retell measure when reading electronically-based stories with hyper-textual supports in the form of labels, vocabulary definitions, and pronunciations of words or segments of text.

When only reading traditional print narratives, fourth-graders classified as proficient- and less-proficient readers made similar improvements in their abilities to answer cued-recall questions and to recall text-based propositions, themes, and plot episodes after four testing sessions (Gambrell et al., 1991). However, only the proficient-readers included significantly more appropriate elaborations with practice.

A comparison of the retell performance of students in grades 5-6 with and without LD (Hansen, 1978) found students with LD included significantly fewer idea units. Both groups accurately retold just over one-third of the total propositions when reading instructional-level material, had similar amounts of “other” information, and included few inaccuracies (mostly isolated, specific details). Students without LD had more partially-correct propositions and recalled significantly more super-ordinate propositions. However, both groups included similar amounts of subordinate details.

Similarly, Zinar (1990) reported that fifth-graders with higher comprehension ability freely recalled significantly more pre-determined propositions than students identified as having low comprehension. In addition, high comprehenders were more likely to include explicitly-stated causal information; whereas, low comprehenders did not include any causal relationships unless probed. Then, low comprehenders included similar amounts of causal information and similar amount of pre-determined propositions as the high comprehenders. Low comprehenders seemed to understand the expository passages just as well as the students considered to have better reading ability, but the former students did not offer as much information unless specifically probed. They did not offer any more non-target information than the high comprehenders, rather the low comprehenders just did not say as much.

This consideration of target/significant and non-target/less significant information from the passage was explored further in Carlisle's (1999) study, which scored students' retells not only by the number of words and idea units included, but also by the importance or centrality of the ideas. Even after controlling for students' scores on researcher-developed sentence verification and science vocabulary tests, sixth- and eighth- grade students with learning disabilities (LD) still performed more poorly on recall than their peers without LD. Both ability groups included similar numbers of ideas and total words. However, the students without LD had better constructed and elaborated oral recalls of the expository passage. Among the better readers, a significantly greater proportion of their overall scores were attributable to main ideas, as opposed to the subordinate details. The follow-up prompting in this study was not specific to the missing information as was the case in the Zinar (1990) study with fifth-graders, so it is not possible to determine if these results confirm or contrast with the earlier study.

These results are consistent with a comparison of fifth-grade on-level, fifth-grade below-level, and third-grade on-level readers when providing retells for an expository passage written on the third-grade level (McGee, 1982). Although there were no significant differences among the groups on the number of subordinate ideas recalled, the better fifth-grade readers included a greater proportion and more total ideas than their peers reading below grade-level. Below-level fifth-graders recalled a greater proportion and more total ideas than third-grade on-level readers. As in the Zinar (1990) study, McGee (1982) found that students' sensitivity to the organizational structure of information in the text was related to their retell performance. Fifth-grade better readers were more likely to match the organization of their response to the structure of the passage read and include more super-ordinate ideas. Fifth-grade below-level readers demonstrated only a partial match to the structure of the text and included similar amounts of super- and sub-ordinate ideas in their recalls. Third-grade on-level readers, however, responded in list-like fashion with no match to the text's structure and included a greater proportion of subordinate ideas. McGee speculated that the differences in performance could be related to the degree of difficulty the expository text presented to students. Fifth-grade better readers not only found the text (written on a third-grade level) easier, but were also more likely to have the requisite background knowledge and experience with expository text.

Similarly, Horowitz and Samuels (1985) examined the recalls of sixth-grade students classified as "poor" and "better" readers when listening to and reading expository passages. Retells were scored with respect to the number of idea units and the rank of those ideas in the text hierarchy. The results did not differentiate between lower- and higher-order information, and follow-up prompting was not specific to missing information. Overall, poor readers performed better when listening to text, and better readers demonstrated significantly higher recall than their lower ability counterparts when reading text. When retell results were disaggregated by the level of text difficulty, both better and poor readers performed better when listening to easier texts. However, the two ability groups had no significant within group difference between listening and reading recall with more difficult texts.

In contrast, Wright and Newhoff (2001) did not report significant differences among the retell performance of students in grades 3-7 with and without language-learning disabilities (LLD) when reading or listening to narrative stories with a difficulty level that does not exceed the students' oral vocabulary or identified reading level. However, students with and without LLD did perform significantly better on inferential comprehension questions when the stories were read to them. In comparing the retell performance of students with LLD, those without LLD matched by chronological age, and those without LLD matched by language ability, the chronological-age-matched group produced more sentences, more verbatim information, and retold significantly more story grammar parts than the other two groups. There were no significant differences between the retell performance of students in the LLD and language-ability-matched groups. The researchers noted that age-matched students generally provided a longer retell, thus giving themselves more opportunity to include story components. As there was no follow-up prompting described for the retell portion, it is possible that students in the other groups might have provided more story components had they been specifically prompted as in the Zinar (1990) study.

Across the 8 studies, students who are considered to be struggling with reading performed more poorly than average achieving or better readers when the retell protocol was administered in a more traditional format (i.e., with print-based passages read independently by the student and assessed with a generic recall prompt). Because the former students have previously exhibited difficulties, it is, perhaps, not surprising that they would perform better on a retell comprehension measure when they receive some assistance with reading the passages – either through electronic hypertext or from the teacher reading the passage aloud. The more compelling data suggest that these younger and middle grades students may not retell as much as they actually do comprehend unless they are specifically cued to provide missing information. However, they still do not provide the degree of elaboration or strength of retell construction that was provided by better readers.

Features of Existing Retell Measures

All but 1 (Karlsen & Gardner, 1996) of the 12 retell measures included in this review has undergone revision and republishing within the last 6 years. Publishers of half the measures have released new editions between 2007 and 2009. The measures are evenly divided between those that are appropriate for preK/K/1 – grade 12, and those that are intended for preK/K/1 – grade 6/8/9. In other words, where an instrument excludes grade levels, those grades considered not appropriate for use with the assessment are most likely in the high school years. In contrast, two of the assessments can be used through adulthood (Bader & Pearce, 2009; Karlsen & Gardner, 1996). Specific information about each study's features is recorded in Table A1, and the information is summarized in the following sections to better explain the findings.

Norming sample characteristics. Although 8 retell measures reported at least some information on the norming samples of students, only 1 had a large and diverse sample that represented the full span of grade levels for which the assessment was intended (Applegate, Quinn, & Applegate, 2008). A second measure reported a more limited sample of students identified in grade groupings (i.e., elementary, middle school, secondary, adult) for the reliability study, but did not utilize all grade levels for the validity study and did not report student ethnicities (Bader & Pearce, 2009). A third measure reported employing a diverse sample representative of all grades, but did not make it clear whether that sample was administered the optional retell subtest (Karlsen & Gardner, 1996). Similarly, a fourth measure had a large and diverse sample of all grade levels excluding the youngest (preK) and oldest (grade 9) for which the instrument is intended; however, the retell measure was not separated from the overall analysis of the assessment in the reliability study and reported no validity study (Cooter, Flynt, & Cooter, 2007).

The remaining 4 measures included only a single grade (Good & Kaminski, 2002b; Johns, 2008) or a small span of grades out of all those for which the assessment is intended (Beaver, 2003; Leslie & Caldwell, 2006). Among those 4 measures, one only conducted a reliability study (Johns, 2008) and another only reported the norming sample for the criterion validity study (Leslie & Caldwell, 2006). Bilingual students were reported in one measure's reliability study sample, but not the validity study sample (Beaver, 2003; Beaver, 2006). Overall, few existing retell measures reported information about the norming sample demographics suitable for determining the generalizability of results across students of different ages and backgrounds.

Existing retell measure format. Nearly all (n = 11) the instruments ask students to provide an oral retell. One measure allows the examiner a choice in requesting an oral retell or a written retell, depending on whether the subtest is administered individually or to groups (Karlsen & Gardner, 1996). More variation occurs in the type of reading conducted prior to administering the retell prompt. Five of the instruments allowed either the student or examiner to choose whether the student read the passage orally or silently. The two instruments that require the students to read the passages orally were developed by the same researchers and are reportedly "parallel" measures (Good & Kaminski, 2002a; Good & Kaminski, 2002b). The remaining 5 measures each employ a different protocol: silent reading and listening comprehension (Bader & Pearce, 2009); listening comprehension plus oral reading for younger students, and silent plus oral reading for older students (Beaver, 2003); silent reading only (Cooter et al., 2007); oral reading and silent reading (Johns, 2008); and listening comprehension plus silent reading (Silvaroli & Wheelock, 2004).

What was difficult to classify in the existing retell measures was the type of text utilized. Most measures (n = 7) describe the stimuli as consisting of both narrative and expository passages. One of the 12 measures includes predominately narratives with a set of alternative materials containing informational and expository passages (Beaver, 2003). Two measures only utilize narratives (Karlsen & Gardner, 1996; Sivaroli & Wheelock , 2004), and two measures provide only narrative stories at lower grade levels with an ill-defined mixture of narrative and expository texts at upper grade levels (Bader & Pearce, 2009; Roe & Burns, 2007). In general, however, the labels provided by the authors are somewhat misleading because informational narratives may be treated as "stories" in one assessment (e.g., Bader & Pearce, 2009) or expository text in another (e.g., Johns, 2008).

Findings from Existing Retell Measures

The 12 existing retell measures identified for this review were analyzed for elements pertinent to the research questions. Specifically, the measures were analyzed for 1) the way in which the retell is prompted; 2) the way in which the retell is scored; 3) the established reliability of the instrument; 4) the established validity of the instrument. Each of these elements will be addressed in the following sections.

Retell prompt. The prompts in the existing retell measures vary along two continua: the initial prompt provided to students and the follow-up prompt given when the students pause or fail to provide certain information. Although the measures do not report using exactly the same wording, there are some patterns among the ways in which students are prompted to retell the passage(s). Almost half (n = 5) the instruments use some form of "tell me about," and as many instruments use the word "retell" when initially instructing students to recall information. Only one measure provides no specific prompt, but allows students to freely recall what they read based on teacher modeling provided before the retell is administered.

All instruments have some mechanism for encouraging students to provide additional information and most (n = 7) allow for specific cuing to missed information, usually through scripted questions. Table A4 provides a matrix to depict the combinations of initial and follow-up prompts included in the existing retell measures. Eight different combinations of prompts are evident among the 12 instruments.

Retell scoring procedure(s). Only two of the 12 measures utilize the same scoring procedures, but that is largely because those instruments were developed by the same researchers and are described as "parallel" to each other (Good & Kaminski, 2002a; Good & Kaminski, 2002b). The procedure is among the least complex in that retell scores are based solely on the total number of relevant words the student uses. The other relatively straightforward method for scoring retells involves a count of the relevant pre-determined idea units or story propositions that students include in their recall (Cooter et al., 2007). Variations on this approach are used in 3 other instruments included in this review. Two measures employ scoring procedures that weight the included ideas/elements based on the examiner's estimation of the students' overall understanding of the topic (Applegate et al., 2008) or the overall quality of the retell (Leslie & Caldwell, 2006). Another instrument requires continued prompting until students' provide a minimum number of ideas and, then, considers the examiner's judgment of whether the retell was organized (Bader & Pearce, 2009).

These more subjective judgments or rankings by the examiner resemble the predominant scoring method utilized by half of the instruments (n = 6). The rubric scores have a variety of scales: 6-24 based on specificity, order, depth of interpretation, and relation of free to prompted recall (Beaver, 2003; Beaver, 2006); "none" to "thorough" for story elements and reading processes (Karlsen & Gardner, 1996); 1-3 on story elements and 1-5 on guiding questions (Roe & Burns, 2007); 1-3 on categories of information and "excellent," "needs assistance," or "inadequate" overall comprehension (Silvaroli & Wheelock, 2004); and "all," "some," or "none" for elements in addition to an estimation of the overall quality and adequacy (Woods & Moe, 2007). One instrument offers examiners four different options for scoring the retell, with different scales or classifications (Johns, 2008).

The scoring procedures currently being used in retell measures are somewhat in contrast to the methods used in the research studies reviewed in the previous section. Purely numerical counts of pre-determined idea units were more frequently used in the research, but holistic and rubric scores are more common in existing instruments, including those used in combination with tallies of idea units and story elements. Nonetheless, the reported inter-rater reliabilities in existing retell measures are consistent with those reported in the research.

Established reliability of existing retell measures. Authors and publishers of existing retell measures were more likely to report the inter-rater reliability of the instruments than any other type of established reliability (e.g., alternate form or test-retest reliability). Half of the instruments (n = 6) provide information on the agreement of different scorers. As was evident in the research on retells, higher inter-rater reliabilities were reported in 3 of the instruments that score retells on the number of pre-determined idea units a student includes in the recall ([.90 - .98+] Applegate et al., 2008; Bader & Pearce, 2009; Leslie & Caldwell, 2006).

Only two measures that score retells holistically or with a more subjective scale provided inter-rater reliabilities (Beaver, 2003; Beaver, 2006; Johns, 2008). These were lower (.74-.81) as is consistent with what was reported in the research studies. A third measure utilizing holistic scores reported "some variation" in scoring but "great consistency" determining the overall reading level of students; however, the authors did not quantify the percent agreements among scorers to define their descriptors (Woods & Moe, 2007).

The second most common type of reliability reported among the existing measures was passage equivalency or alternate form reliability. Five measures provided data that ranged from a low of .57 (Good & Kaminski, 2002b) to a high of .90 (Leslie & Caldwell, 2006). Only 2 of the 12 instruments reported test-retest reliability data, and neither reported alternate form reliability (Beaver, 2003; Beaver, 2006; Cooter et al., 2007). Test-retest reliability ranged widely from .67 to .93 in the measure incorporating both narrative and expository text (Cooter et al., 2007) and were in the .90 range for the measure that primarily utilizes narrative stories (Beaver, 2003; Beaver, 2006).

The remaining reliability data included an estimated reliability of 3 passages for the retell fluency measure (.80) based on the Spearman-Brown prophecy formula (Good & Kaminski, 2002b); the percent agreement (66%) on reading instructional level between the reading inventory and a clinician-constructed inventory (Johns, 2008); and the internal consistency of the overall reading comprehension portion of the instrument which included the retell protocol as an optional component ([r = .79 - .97] Karlsen & Gardner, 1996). Only one measure provided information to establish the reliability of the pre-determined idea units used to score students' retells (Leslie & Caldwell, 2006). Two measures reported no reliability data (Roe & Burns, 2007; Silvaroli & Wheelock, 2004). These same instruments provided no validity data either.

Established validity of existing retell measures. Remarkably little work has been done to establish the validity of the existing retell measures. Five of the 12 instruments reported no information on validity; however, 2 of those measures included correlation data in sections of the technical manuals labeled as "reliability" that was similar to what other measures reported in sections labeled "validity" (Johns, 2008; Leslie & Caldwell, 2006).

Four measures provided correlations among test components as validity data. Although the results were somewhat consistent in indicating moderate correlations, some measures lacked specific information or a broader sample that would increase the confidence in and generalizability of the data. A moderate correlation (r = .51) was reported between the retell score on the *Critical Reading Inventory* (Applegate et al., 2008) and the total comprehension score on narrative passages, but a less robust correlation (*r* = .43) was reported for informational passages. Leslie and Caldwell (2006) reported the retell component of the *Qualitative Reading Inventory* (*QRI-4*) was correlated with prior knowledge scores from kindergarten through upper middle school, but no coefficients were provided. In addition, the overall reading comprehension score was correlated with word identification and rate at preK, second-, third-, and fourth-grades, but no information on the complete norming sample and no coefficients were provided. With a limited sample of first-graders, the average retell fluency score on the *Vital Indicators of Progress* ([*VIP*] Good & Kaminski, 2002b) was moderately correlated (*r* = .61) with the oral reading fluency average. Finally, the continuity of the *Stanford Reading Diagnostic Test* (*SDRT*) across grade levels was established with moderate to strong correlations between corresponding subtests (*r* = .59 - .87), but the optional retell subtest was not disaggregated in the data.

Test developers often provided information on only one type of validity (e.g., concurrent, predictive, construct, or criterion validity), and rarely did two measures include data on the same type. The developers of the *SDRT* sought to establish the instrument's construct validity (how accurately the test measures the construct of reading and academic performance) by correlating results to scores on a standardized measure, the *Otis-Lennon School Ability Test*. In contrast, researchers of the *VIP* correlated results to scores on a standardized measure of general reading achievement, the *Broad Reading Cluster*, in order to establish the *VIP*'s predictive validity (how accurately the test represents students' future reading ability or performance). Despite the different purposes, results in neither validity study were highly encouraging. Correlations between the *SDRT* and the *Otis-Lennon* for a large sample of students in grades 2 through 12 were reported from a moderate .43 to a strong .95, a wide range without disaggregated data on the optional retell subtest. The correlation of the *VIP* with a limited sample of first-graders was a moderate .51, but the retell measure only explained an additional 1% of the variance in the *Broad Reading Cluster* results compared to the variance accounted for by ORF scores alone (Roberts et al., 2005).

The assessment labeled as “parallel” to the *VIP*, the *Dynamic Indicators of Basic Early Literacy* ([*DIBELS*] Good & Kaminiski, 2002a), provided data on criterion-related validity. Consistent with the *VIP* data, the correlation between the *DIBELS* retell component and the *Oregon State Assessment Test* was a moderate .50. However, the test publishers did not directly report the norming sample or the percent variance explained by the *DIBELS* retell. In addition to predictive validity, information was provided on the measure’s concurrent validity (how accurately the test represents the student’s current level of reading ability or performance). The correlation between DIBELS and ORF scores was, again, reported as moderate (*r* = .59), with no immediately available information on the norming sample.

The developers of both the *Developmental Reading Assessment* ([*DRA*] Beaver, 2003; Beaver, 2006) and the *QRI-4* provided results on the correlation of their measures to the *Iowa Test of Basic Skills* (*ITBS*). Data for the *QRI-4* were used to establish the instrument’s criterion validity; whereas, the developer of the *DRA* did not specify what type of validity the data were to establish. As with the intra-correlations of test components reported earlier, results were similar but lacked specific information on the norming samples or were based on samples that did not reflect the full spectrum of grade levels for which the assessments are intended. The *DRA* was moderately correlated (*r* = .68 - .83) with *ITBS* grade-equivalent scores and national curve equivalents as well as Lexile measures. However, only students in grades 1, 2, and 3 participated in the validation studies. Interestingly, the developers of the QRI-4 did not administer the ITBS to students in grades 1 through 3 but, instead, administered the *California Achievement Test* for these lower grade levels.

Correlations between the *QRI-4* and the *ITBS* (for grades 3-8) or the *California Achievement Test* (for grades 1-3) were reported in a wide range, with some non-significant findings and inconsistent results on narrative versus expository passages in the *QRI-4*. For narrative text, correlations ranged from a weak and non-significant .27 at grade 6 to a strong .85 at grade 1. For expository text, correlations ranged from a weak and non-significant .28 at grade 7 to a moderate .55 at grade 9. The norming sample was reported as including students in grades 1 through 8, so it is unclear how the results for the grade 9 students were obtained. The *QRI-4* is intended for use through high school. Test developers also reported a moderate correlation (*r* = .75) between the *QRI-4* and the *Woodcock Reading Mastery* passage comprehension subtest, but did not specify the type of validation study conducted or the norming sample on which the results were based.

The developer of the *DRA* established content validity by reporting that 89% of the teachers at the test development site (n = 84) agreed that the measure was helpful in evaluating students' reading progress, and 82% agreed that the *DRA* was helpful in determining instructional goals. The only other instrument reporting similar data was the

BADER Reading and Language Inventory (Bader & Pearce, 2009) reporting a high correlation between *BADER* scores to school reading specialists' judgments of students' reading level (*r* = .93) and to classroom teachers' judgments of students' reading levels (*r* = .89).

Discussion

This descriptive review sought to determine if a retell measure contributes unique, valid, and reliable information about students' reading comprehension.

What existing research has examined the validity of retell as a comprehension measure?

Results from the 26 studies reviewed here indicate that retells tend to be moderately correlated with standardized and researcher-developed measures of reading ability across grade levels and other demographic variables. Consistent with other research on the technical adequacy of ORF measures (Jenkins & Jewell, 1993; Spear-Swerling, 2006), ORF scores in the studies included in this descriptive synthesis were more strongly correlated with the other measures of reading and accounted for more variance than retell scores through eighth-grade (Fuchs et al., 1988). However, there was some confirming evidence that ORF is less of a factor in students' reading comprehension by grade 5 (Shinn et al., 1992). Above this age, ORF as a measure of reading progress begins to asymptote (Fuchs et al., 2001; Stage & Jacobsen, 2001) and the correlation between ORF and standardized measures of reading emerges as less robust than for younger students (Schatschneider et al., 2004; Wiley & Deno, 2005). It is also around fifth-grade where retells show more sensitivity to practice effects (Gambrell et al., 1991; Gambrell et al., 1985; Zinar, 1990). Despite evidence that retell scores are instable for fourth- and fifth-graders (Fuchs & Fuchs, 1992; Mason et al., 2006), it is not yet clear whether retell performance is a valid and reliable means of predicting and monitoring student progress above the grade 5 fulcrum.

One study in the review indicated that written retells might be more adequate indicators of reading ability than oral retells (Fuchs et al., 1988). However, the correlation of the written retell score to ORF and standardized measures of reading was still within the moderate range. Moreover, scoring oral retells through a quantitative analysis of predetermined propositions or idea units (as opposed to holistic ratings) produced inter-rater reliabilities comparable to those for written retells. Hence, it may not be of practical significance to require written responses from students, especially given the time efficiency of scoring oral responses at the moment they are elicited.

What seems the more critical aspect of using retell protocols is defining the expectation for what information they can provide about students' comprehension. Students of all ability levels were not likely to spontaneously include inferences or implicit information in their recalls (Best et al., 2008; Zinar, 1990), and follow-up prompting was necessary to improve the recall of targeted information (Gambrell et al., 1991; Best et al., 2008; Zinar 1990) or elicit inferential information (Gambrell et al., 1985). This was particularly true when using expository texts and when assessing students with LD or other reading difficulties, who needed more textual support, practice, and cuing to produce retells on par with better readers. Even after controlling for explicit vocabulary knowledge, students with LD produced more poorly constructed and less well elaborated recalls than students without disabilities (Carlisle, 1999). On average, students provide between 30-40% of the important information in narrative and expository passages (Hansen, 1978; McGee, 1982). With greater awareness of text structures, students still recall less than 55% of the main ideas in short passages specially written to present a consistent and recognizable organizational pattern (Richgels et al., 1987).

Consequently, it seems unrealistic to include a retell measure for the purpose of adding richness to an assessment of students' reading comprehension as some researchers have suggested (Leslie & Caldwell, 2006; Gambrell et al., 1985; Klingner, 2004). It is worth reiterating that subjective estimations of the quality, coherence, and completeness of retells are not as reliably scored (Gambrell & Jawitz, 1993; Loyd & Steele, 1986; Mason et al., 2006; Pearman, 2008) as numerical counts of explicit idea units. Although one study reported 100% inter-rater reliability on scoring inferential cued follow-up questions (Gambrell et al., 1985), no samples of the questions or responses were provided to substantiate the depth of processing required in the literal versus inferential questions. It is uncommon to have perfect agreement on implicit answers or explanations, and even competent adult readers have demonstrated difficulty monitoring their inferential comprehension when responding to open-ended questions (Pressley, Ghatala, Woloshyn, & Pirie, 1990). It would seem best not to expect retell measures to detect more advanced comprehension skills.

Further support for this can be derived from the lack of consistency in and influence of the retell prompt. Prompts and the expectations for student responses interchangeably apply the terms retell, recall, summarize, and paraphrase. However, these do not measure equivalent cognitive processes (Cutting & Scarborough, 2006; Keenan et al., 2008; Kintsch & van Dijk, 1978; Scardimalia & Bereiter, 1987). Greater attention was paid to reporting the scoring procedures employed in the studies than the procedures for obtaining the grist of what was scored. Although existing retell measures reportedly lack of uniform scoring procedures (Nilsson, 2008) and demonstrate weak inter-rater reliabilities (Klesius & Homan, 1985), the studies of retell in this review were found to have more commonality in scoring and consistency in inter-rater reliabilities than for retell prompts.

This is an important weakness in the extant research on retell measures because findings suggest that variations in the wording of a question or prompt (Fuchs & Fuchs, 1992; Gagne et al., 1997; Seifert, 1994) or in the administration procedures surrounding the prompt (Cordon & Day, 1996; Gambrell & Jawitz, 1993; van den Broek et al., 2001) can substantively alter both the quantity and the quality of participants' responses. As indexes of quantity and quality are the means by which retells are scored, insufficient reporting of the prompts employed and a paucity of data on the outcomes associated with different prompts substantially reduce the confidence with which interpretations can be made about the validity, reliability, and utility of retell measures.

How have existing assessments of reading comprehension incorporated a retell procedure?

The retell assessments reviewed for this paper commonly allowed oral or silent reading options, but such combinations were rarely reported in studies of retells (n = 3).

Only 2 of the existing assessments required oral reading only, and these parallel measures included an ORF component (Good & Kaminski, 2002a; Good & Kaminski, 2002b). Research has not yet addressed whether reading orally or silently produces significant variation in student retell performance, so it is not possible to draw conclusions about whether allowing students an option is a strength or weakness of the existing measures.

There is more evidence that passage type and difficulty can affect student performance (Best et al., 2008; Francis et al., 2008; Leslie & Caldwell, 2006; Otto et al., 1968; Richgels et al., 1987), but the ways in which narrative and expository texts are defined and incorporated is inconsistent across assessments. It seems more likely that narratives predominate in the existing measures because even where both genres are included, the expository passages are often in optional sections or are not clearly distinguishable in the set of stimuli. Expository text is more challenging for students and more prevalent in the middle grades and high school where the research indicates that retell measures might become a more valuable tool for gauging student comprehension performance, so many existing assessments do not go far enough in providing materials that could be considered authentic or reflective of the reading demands confronted by adolescents. In fact, half the measures are not designed for grades 10 through 12, and some have only added grades 6 through 9 in more recent editions. It could be that the use of these primarily informal instruments has not yet "come of age" for middle and secondary schools.

If the assessments are designed more with younger elementary students in mind, it might also explain why all but one of the measures included in this descriptive synthesis asks for an oral retell. This is the one feature that was most consistent across the 12 assessments. Because the studies of retells did not provide enough compelling data to indicate that written retells would be more valid and reliable, there is no reason to suggest the form of the retell should change in future editions. However, there is more reason to reevaluate the administration and scoring procedures of the instruments.

Although there was more consistency in the initial prompts than was apparent in the 12 retell measures employed in the 26 studies, there is still variation in the wording that could influence students' responses. In addition, nearly half (n = 5) the existing commercial measures provide only general follow-up prompting. The research data suggest that more specific follow-up prompting can potentially mitigate the influence of background knowledge and reading ability with expository text (Best et al., 2008; Zinar, 1990) and reflect practice effects with both narrative stories (Gambrell et al., 1991) and informational narratives (Gambrell et al., 1985). Therefore, those measures that include scripted follow-up questions or structured prompting may be more sensitive to students' true understanding of the text.

Of greatest concern is the diversity in scoring procedures within and across instruments, most of which rely upon subjective judgments or ratings of retell quality, coherence, and accuracy that make it difficult to achieve agreement among different raters. These issues have been noted by previous researchers (Klesius & Homan, 1985; Nilsson, 2008), but this review provides support from both studies of retells and data reported by existing measures that numerical counts of propositions or idea units would improve the reliability with which student responses can be scored.

What is the reliability and validity of the retell component in existing assessments?

Test developers report very little information on the technical adequacy of existing retell measures, and the data that are available are usually based on inadequate samples. Only one measure included a norming sample representative of the entire span of grade levels for which the instrument is intended (Applegate et al., 2008). The other instruments that described norming samples utilized limited numbers of students and/or limited grade levels. In addition, results were generally reported for studies of the reliability or validity of the instrument as a whole, and were not disaggregated by subtest or component. Therefore, it is not clear if the data are applicable to the retell protocol. Two technical manuals did not include any information on the reliability or validity of the instruments (Roe & Burns, 2007; Silvaroli & Wheelock, 2004), but no existing retell measure can be said to have a satisfactorily substantiated reliability and validity for students of the appropriate grade levels and from diverse backgrounds.

Inter-rater reliability was reported most frequently, but was still only provided for half the instruments. As mentioned above, these coefficients were consistent with what was reported in the research. Namely, holistic scoring procedures had lower inter-rater reliability than numerical counts of propositions or idea units. Little could be determined about how test developers controlled for measurement artifacts. Only 5 technical manuals included data on passage equivalency, and 2 different measures reported test-retest reliability (Beaver, 2003; Beaver, 2006; Cooter et al., 2007). The results were so inconsistent as to invite questions about the alternate form and test-retest reliability of those measures with no reported data (Newcomer, 1999; Nilsson, 2008).

Interestingly, there was not even a common understanding of whether correlations between the retell measure and another measures of reading comprehension established the instrument's reliability (Johns, 2008; Leslie & Caldwell, 2006) or validity (e.g., Beaver, 2003; Beaver, 2006; Good & Kaminski, 2002a; Good & Kaminski, 2002b; Roberts et al., 2005). This was because the exact statistical methods and procedures were often insufficiently described. Nevertheless, the correlations provided were somewhat consistent with what was found in the 26 studies reviewed in this synthesis: Retell tended to be moderately correlated with other measures or components for older students and had more variability at younger grade levels.

Implications for Creating Quality Retell Measures

Results indicate that retell measures hold promise as a means to assess the literal reading comprehension of students above grade 5, but commercially available retell measures probably need to be revised and validated before they can be used with confidence. From what can be concluded in the research, instruments for older students should include more clearly recognizable expository passages that resemble the type of reading an adolescent might be expected to do in school. Specific follow-up prompting or cuing should be included only after the initial free recall if the expectation is for students to produce lengthy or complete recalls of the information. Scoring procedures should be based on pre-determined story propositions or idea units, and a more lenient proportion should be considered for expository texts as opposed to narratives. At a minimum, the instruments should be validated across all grade levels (and other demographic variables) for which they are intended and have an established alternate form reliability.

Limitations and Directions for Future Research

Because retell and recalls were rarely the focus of the research or the primary component of existing measures, it was difficult to identify all relevant studies and instruments. An attempt was made to carryout the search in as systematic a way as possible and to carefully document the search procedures. However, most items included in this review were identified in ancestral or manual searches, which are more difficult to replicate and more prone to omissions.

In addition, the available data were often from a single study with that focus or a single measure that approached reliability or validity in that way. Hence, the generalizations made about retell measures are tenuous. Much more research is needed to provide a convergence of evidence on the reliability, validity, and utility of retell measures. The conclusions and recommendations provided in this review can only be considered preliminary. To advance the field, future studies should address the optimal wording of the initial prompt administered in a retell protocol. To the extent that variations in how and when a question is asked (Fuchs & Fuchs, 1992; Gagne et al., 1997; Seifert, 1994; van den Broek et al., 2001) or how instructions are provided (Cordon & Day, 1996; Gambrell & Jawitz, 1993) can substantively alter both the quantity and the quality of participants' responses, retell scores can confound students' comprehension with the influence of the prompt.

A more consistent, valid, and reliable means of eliciting a free recall must be determined before retells can be studied as a means to monitor the reading progress of students above grade 5. Furthermore, future research should attempt to determine the number or proportion of idea units that are associated with "better" or "weaker" comprehension in order to guide teachers in making instructional decisions. Finally, studies might be conducted to compare performance in oral and silent reading, to compare practice effects in narrative and expository passages, and to explore the influence that teacher or student characteristics might have on the assessment of retell performance.

A well-defined line of research on retell measures would explicate their role in assessing students' reading comprehension. If retells are less sensitive to decoding ability (Keenan et al., 2008) and can detect other instructional areas of need potentially missed by an ORF measure alone, retell protocols could become valuable tools in school-wide approaches to reading intervention that rely upon cost-effective and time-efficient data gathered at multiple times throughout the year.

Table A1

Study Characteristics

Study	**Purpose for and Form of Retell in Study**	**Prompt**[a]	**Scoring Procedure**	**Findings**
Validation Studies				
1. Fuchs & Fuchs (1992) Grade Level(s): 4 – 5 (n=38) All with LD or EBD Passages were read silently by student.	Determine the criterion validity of written retell and adequacy of the measure for monitoring students' reading progress. Form of retell: Written	Students were asked to write a summary of the passage.	Total number of words; Total number of words that matched original text	The instability in students' total number of words and total matched words made it difficult to interpret growth. Teachers felt the counts did not reflect quality of writing. Inter-rater reliability for total number of matched words was .93.
2. Fuchs, Fuchs, & Maxwell (1988) Grade Level(s): 4 – 8 (n=70) All boys with LD, EBD, or "mental retardation;" 31% minority Student read	Purpose: Determine the criterion and concurrent validity of informal reading measures, including narrative recall Form of retell: One session oral and one session	NR: Students given 10 minutes to freely recall. If they finished before the time limit, they were given 4 "controlled prompts" administered consecutively after 30 seconds of no response.	Number of words; percent of content words matching original text; percent of predetermined idea units	Inter-rater reliability ranged from .85 - .97 with higher agreement on most written retell elements (except percent idea units). Retells were moderately correlated with SAT-7 RC and WS (*r* ranged from .47 - .82). Retell correlations with SAT-7 RC were consistently and significantly higher than with WS. Correlations for written recall were consistently and significantly higher than for oral recall. There were no significant differences in recall scoring procedures. ORF had significantly higher correlations with SAT-7 RC than other measures, and had higher correlation with RC

Study	**Purpose for and Form of Retell in Study**	**Prompt**[a]	**Scoring Procedure**	**Findings**
some passages orally and some silently.	written (counterbalanced)			than with WS. ORF was moderately correlated with retell (mean *r* = .75).
3. Hansen (1978) Grade Level(s): 5-8 (n = 34) Half identified as LD Student read passages orally	Purpose: Determine correlation of retell to reading comprehension on instructional-level narrative text; Compare retell performance between students with and without LD Form of retell: Oral	Initial: Students asked to retell the passage in their own words. Follow-up: "Can you tell me more?" (repeated until no more information remembered) Open-ended comprehension questions drawing from factual information in passage.	Percent of pre-determined idea units	Inter-rater reliability = .94. Moderate to strong correlation between proportion of idea units recalled and performance on open-ended, factual comprehension questions (ρ = . 46 - .77). Students with LD included significantly fewer idea units than average readers. Both groups accurately retold just over one-third of total propositions when reading instructional-level material, had similar amounts of "other" information, and included few inaccuracies (mostly isolated, specific details). Students without LD had more partially-correct propositions and recalled significantly more super-ordinate propositions. However, both groups included similar amounts of subordinate details. On the comprehension questions, students without LD provided more correct answers than students with LD. Students with LD had significantly lower comprehension in instructional- versus independent-level text.
4. Loyd & Steele (1986)	Purpose: Determine whether	Students asked to paraphrase the passage in their	Sum of the weighted idea units in the written recall	Inter-rater reliability was .97. Reliability of idea unit score was .81 and of coherence scale was .72. Idea unit and coherence were

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
Grade Level(s): 11 – 12 (n=108) None in special education or LA in reading Cannot determine if student read passage orally or silently.	standardized reading comprehension measures are tapping the same constructs as free recall Form of retell: Written	own words	(weighted by the importance rating of idea units in the text); Coherence scale score of 1 (low) to 7 (high)	moderately correlated (r = .73). Correlations between idea unit scores and SRA reading comprehension and language arts mechanics scores were weak to moderate (range of *r* was .28 - .56). Correlations between coherence scores and SRA scores were all weak (range of *r* was .11 - .39).
5. Rasinski (1990) Grade Level(s): 3 and 5 (n=142) None with LD; 15% non-white; upper- and lower- middle class Passage read orally by student.	Purpose: Determine how well fluency measures predicted expository reading comprehension (free and cued recall) Form of retell: Oral	Free recall: Students asked to recall all they could remember from what they had read orally. Cued recall: Major headings from an outline of the passage were used to prompt recall.	NR	Retelling was not correlated with researcher-developed measures of phrasing and was weakly correlated with both miscue and reading rate (*r* = .38 - .52).
6. Riedel (2007) Grade Level(s): 1 (n=1,518)	Purpose: Compare the predictive validity of retell and other	"Please tell me all about what you just read. Try to tell me everything you can.	Number of words that illustrate an understanding of the passage	ORF was a better predictor of reading comprehension than retell or other subtests. Adding retell improved predictive accuracy by less than 1 percent. Retell was moderately

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
None in special education; 92% African-American; 85% free/reduced lunch; few ELL Passage read orally by student.	DIBELS subtests on first graders' end-of-year reading performance Form of retell: Oral with one-minute time limit	Begin."		correlated to GRADE (r =.41 - .69) and weakly correlated to TerraNova (r = .39 - .46)
7. Roberts, Good, & Corcoran (2005) Grade Level(s): 1 (n=86) 90% African-American; 100% free/reduced lunch Passage read orally by student.	Purpose: Determine relationship between retell and ORF on the VIP; Determine variance in reading performance accounted for by retell Form of retell: Oral	NR	Number of words used in correctly retelling the story	Some students' retell scores were inconsistent/unexpected given their ORF scores. Retell fluency explained 1% more of the variance on comprehension than ORF alone.
8. Shinn et al. (1992) Grade Level(s): 3 and 5 (n=238)	Examine the factor structure of reading and detect any developmental	Initial: Students were asked to retell the story (narrative folktale) in writing using their own	Total number of recognizable words written.	A one-factor model of reading was most parsimonious at grade 3 where factor loading for written retell on Reading Competence was .68. The highest factor loading was for ORF (.90) . Residual variance was highest for written

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
Predominately Caucasian; 6% in special education Passages read silently by students.	differences. Form of retell: Written	words. Follow-up: Students given four verbal prompts at 1-minute intervals (e.g., "Is there anything else you can remember about [folktale title]? Write it down.")		retell at 63%. A two-factor model of reading was most parsimonious at grade 5 where factor loading for written retell on Reading Comprehension was .61. The highest factor loading was for cloze exact matches (.86). Residual variance was highest for written retell at 74%.
Reliability Study				
9. Gambrell, Koskinen, & Kapinus (1991) Grade Level(s): 4 (n=48) No demographic data provided Passages read silently by student.	Purpose: Determine if practice affects retell performance of more and less proficient readers on narrative stories. Form of retell: Oral	Free recall: "Take a minute or two to think about how you will tell the story. Let me know when you are ready to tell the story into the tape recorder." Cued recall: Asked to respond to four text-explicit and four text-implicit questions specific to each story.	Free recall: Proportion of pre-determined story structure elements and propositions recalled. Cued recall: Number of correct responses.	Inter-rater reliability for free –recall of propositions was 94%, of story structure elements was 95%, and of cued recall was 92%. Both proficient and less-proficient readers recalled significantly more text-based propositions, themes, and plot episodes after four sessions. There were no differences in the inclusion of inconsistent or erroneous propositions, but proficient-readers included significantly more appropriate elaborations after four sessions. Proficient and less-proficient readers answered significantly more cued-recall questions after four sessions.

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
10. Gagne, Bing, & Bing (1977) Grade Level(s): 10 – 12 (n=24) Largely middle-income Passages read silently by student.	Purpose: Demonstrate that a stated expository reading goal can affect free recall organization of average- and advanced-ability students Form of retell: Written	Write down the first 10 facts that could be remembered from the passage	Number of topic and attribute facts	Students given topic (discrete facts about a single topic and given in same order as in paragraph) or attribute goals (2 to 3 facts or attributes of a topic given in different order than in paragraph) recalled the same amount of information but organized the information differently. Those given topic goals almost exclusively provided topic facts, and those given attribute goals almost exclusively provided attribute facts.
11. Gambrell & Jawitz (1993) Grade Level: 4 (n = 120) All on grade-level readers Students read passages silently.	Purpose: Compare the effect on retell when directions prior to reading an illustrated narrative were intended to induce mental imagery, draw attention to illustrations, emphasize both mental imagery and illustrations,	Initial: "Write the story you just read for a friend who has not read or heard the story before." Follow-up: 16 cued recall questions	Number of pre-determined propositions and 10-point analysis of recalled information about characters, setting, theme, plot episodes, resolution, sequence. Cued recall questions scored using a template of	Inter-rater reliability = .90 for the number of propositions; .85 for the 10-point analysis; and 1.00 for cued recall questions. Students in the mental imagery, illustrations, and mental imagery + illustrations treatment groups recalled a significantly greater number of propositions than the "read to remember" control group. The imagery + illustrations group recalled story structure elements significantly better than all other groups. Within the story structure elements, the imagery + illustrations group performed significantly better on setting, characters, and plot than the control group, and significantly better on characters than the illustration only group. In

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
	or “read to remember” Form of retell: Written		acceptable answers.	addition, the combined treatment group students were more likely to provide complete retells, performed better overall on cued recall questions, and better on text-implicit cued recall questions than students in the three other groups. For text-explicit questions, the imagery + illustrations group performed significantly better than the “read to remember” control.
Comprehension Outcome Measure				
12. Best, Floyd, & McNamara (2008) Grade Level(s): 3 (n=61) 57% African-American; 28% Caucasian; 7% bi-racial; 3% Asian/Pacific Islander; range of ability levels Passages read silently by student.	Purpose: Examine the effects of text genre, decoding skills, and world knowledge on text comprehension Form of retell: Oral	Free recall: “Tell me everything you can remember about what you have just read. Give me as many details as possible, like you were trying to tell a friend about what you just read.” Cued recalls were given in three parts, asking students to “Tell me everything” about each of the three major sections of the text.	Number of directly relevant idea units recalled divided by the number of pre-identified propositions in the text	Inter-rater reliability kappa weights were .85 for expository and narrative texts. Only 1% of free recalls and 3% of cued recalls contained inferences. Students recalled between 4 (free) and 7 (cued) percent of propositions in expository text and 10 (free) to 15 (cued) propositions in the narrative text. Recall on narrative text was significantly better than on expository. Narrative free recall, narrative cued recall, and expository free recall were weakly to moderately correlated with both the Woodcock Johnson (WJ) letter-word identification test (*r* = .36 - .58)and the WJ academic knowledge test (*r* = .48 - .64). Expository cued recall was moderately correlated with world knowledge (*r* = .55). Letter-word identification was the strongest predictor of narrative recall. Academic knowledge was a stronger predictor of expository recall than decoding skills.

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
13. Carlisle (1999) Grade Level(s): 6 and 8 (n=63) Some with LD; 40% Afican-American; 40% Caucasian Cannot determine if passages were read orally or silently by student.	Purpose: Determine whether comprehension of expository passages by students with LD is affected by the use of a recall task. Form of retell: Oral	Initial: Students were asked to tell as much of the passage as they could remember. Follow-up: "Is there anything more you can tell me?"	Number of predetermined ideas weighted by importance / centrality; number of words in recall	Inter-rater reliability was .97 - .98. Students without LD performed significantly better on overall recall, sentence verification, and vocabulary than those with LD at both grade levels. At both grade levels, recall was moderately correlated to scores on researcher-developed sentence verification (*r* = .50 - .74) and science vocabulary (*r* = .49 - .51) tests. However, after controlling for these scores, students with LD still performed more poorly on recall. Students with and without LD included similar numbers of ideas and total words, but a significantly greater proportion of the overall recall score was attributable to main ideas (as opposed to subordinate details) for students without LD. Better readers produced better constructed and elaborated recalls.
14. Doty, Popplewell, & Byers (2001) Grade Level: 2 (n = 39) All from a Title 1 elementary school Students read	Purpose: Compare comprehension differences with electronic versus print-based storybooks Form of retell: Oral	Initial: Students were asked to retell the story, what it was about, and what they remembered about the events. Follow-up: 3 literal and 3 inferential comprehension	10-point analysis of recalled information about characters, setting, theme, plot episodes, resolution, sequence	No significant differences in oral retelling of print versus electronically-based stories. A small, but significant (p< .05) difference was found in the performance on the comprehension questions favoring students reading from the electronic medium.

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
passages silently.		questions.		
15. Gambrell, Pfeiffer, & Wilson (1985) Grade Level: 4 (n = 93) All native English speakers; none LA in reading Students read passages silently.	Purpose: Investigate the effects of practice in retelling versus practice in illustrating important ideas upon the comprehension and recall of "expository" text information. (Based on prompt's reference to "story," passages are presumed to be informational narratives as opposed to true expository text.) Form of retell: Written outline, then oral response	Outline included one blank for "Important Idea" and tow blanks for "Supporting Detail." Initial prompt for oral retell: Students were instructed to retell "all the important ideas from the story." Follow-up: 10 literal-level cued questions and 10 inferential cued questions.	Outline and oral retell scored for number of predetermined idea units included from 6 categories: agent and action, modifier, where/when/how/why, belongs to, conjoining, and proposed action or event. (Categories were not defined.) Cued questions scored for accuracy.	Inter-rater reliability for scoring retell was .96, and for follow-up questions was 100%. The students who practiced retelling significantly outperformed the students who drew an illustration on the scoring categories of agent/action, modifier, where/how/when, and proposed action for both the outline and oral retell. Retell practice students also significantly outperformed the illustrating students on the cued literal and inferential questions. Performance of retell practice students on a 2-day delayed free recall task was significantly better than the immediate free recall of the illustrating students.
16. Horowitz &	Purpose: Indicate	Listening	Number of idea	Inter-rater reliability was .90. Poor readers had

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
Samuels (1985) Grade Level(s): 6 (n=38) Middle-class In listening condition, passage read to student (via audiotape). In reading condition, passage read orally by student.	"better" and "poor" readers' decoding and comprehension in expository text Form of retell: Oral	comprehension: "Tell me as much as you can remember about the passage that you just heard." Reading comprehension: "Tell me as much as you can remember about the passage that you just read."	units, number of ideas in the text hierarchy, and highest-order rhetorical predicates	greater recall when listening as opposed to reading text. Better readers had significantly higher recall than poor readers when reading text. There was no significant difference between good and poor readers' recall when listening to text. Both good and poor readers had significantly better recall when listening to easier texts, as opposed to reading them, but had no significant difference between listening and reading recall with more difficult texts.
17. Kouri & Telander (2008) Grade Level(s): K-1 (n = 30) All at-risk for reading problems due to speech/language delay and weak phonological awareness skills	Purpose: Determine effect of sung storybook readings on students' story retelling, reading comprehension, and story participation Form of retell: Oral	Initial: "Pretend that I have never heard the story and tell it back to me." Follow-up: "Just tell me anything you can remember about the story;" "Can you tell me more about the story?"	Holistic 4-point analysis on each of 4 text-based comprehension elements (explicit or inferred information), 4 reader response elements (connect, generalize, and relate story to real-world), and 4 language use	Inter-rater reliability = .98 on retelling; .93 on comprehension questions; .97 on a behavior rating. There were no significant differences in the number of prompts delivered, the retell performance, or the scale scores on the comprehension questions for sung versus spoken stories. On both formats, students' language use scores were significantly higher than their text-based and reader response scores. Text-based scores were significantly higher than reader response scores. Average MLU did not differ significantly in response to sung versus spoken stories; however, students

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
Teacher either read or sung passage aloud to student.		8 Comprehension questions about story content	elements (organize ideas and use appropriate communication). Mean length of utterance (MLU) and Type-token ratio (TTR). Comprehension questions 0-3 holistic scale score for completeness.	included a greater number of different words (TTR) when retelling sung stories as compared to spoken stories. Ratings for attention and on-task behavior were higher during spoken stories as compared to sung stories.
18. Mason, Snyder, Sukhram, & Kedem (2006) Grade Level(s): 4 (n=9) Most with LD, EBD, and/or SLI; 57% Caucasian; 16.3% African-American; 23.1% Hispanic;	Purpose: Examine the effects of multiple strategy instruction on comprehension (oral retell) of expository text and ability to summarize (written retell) of the text. Form of retell: Oral first, then	Oral retell (provided first): Tell orally everything that was read and learned in the passage, as if the assessor knew nothing about the passage topic Written retell (followed oral retell): Write an essay that told	Quality score of 0 to 6 points based on the number of pre-determined main ideas and number of details recalled. Number of idea units recalled. Total number of words in written retell.	Inter-rater reliability was 95% for number of main ideas, 82% for quality, 93% for number of idea units, and 100% for number of written words. Most students increase the number of main ideas included in oral retell. Some students increased the number of written main ideas. Variability in quality of oral and written recalls increased after instruction. Number of orally stated idea units was inconsistent.

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
3.7% Asian; all low income Student choice to read passage orally or silently.	written	everything that was read and learned in the passage, as if the assessor knew nothing about the passage topic.		
19. McGee (1982) Grade Level(s): 3 (n = 20 on grade-level readers) 5 (n = 40: 20 on grade-level readers; 20 below-level readers) Cannot determine if student read passage orally or silently	Purpose: Examine the differences between good and poor readers' awareness of text structure in expository passages written on the third-grade level and the influence of text structure awareness on recall Form of retell: Oral	Initial: Students asked to tell everything they could remember about the passage. Follow-up (one time deliver): Asked if they remember anything else	Proportion of pre-determined idea units recalled. Analysis of how similar the structure in the recall was to the text structure of the passage read, where "full" = at least 3 super-ordinate and at least 2 subordinate propositions; "partial" = 2 subordinate propositions; and "no structure" = any other patterns.	Inter-rater reliability = .97 - .98. Fifth-grade on-level readers recalled a greater proportion and more total ideas than their peers reading below grade-level. Below-level fifth-graders recalled a greater proportion and more total ideas than third-grade on-level readers. There were no significant differences among the groups on the number of subordinate ideas recalled. However, grade 3 average readers recalled proportionally more sub- than super-ordinate ideas; whereas, fifth-grade average readers recalled proportionally more super- than sub-ordinate ideas. Fifth-grade below-level readers had no significant differences between the proportion of super- and sub-ordinate ideas. Most fifth-grade average readers had full structure in recalls. Most fifth-grade below-level readers had partial structure in recalls. Most third-grade average readers had no text structure in their recalls – the recalled in a list-like fashion.
20. Moss (1997)	Purpose: Determine what	Initial: *Retell the book as	Holistic 5-point Scale for Judging	Most children could accurately and completely provide main ideas and details, retell

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
Grade Level(s): 1 (n=20) Below- and above- average ability levels; low- to upper-middle class Teacher read passage aloud to student.	retell reveals about understanding of informational trade books. Form of retell: Oral	if telling it to a friend who had never heard it before. Follow-up: *”What else do you remember?” *”If you were to tell a friend about this book in just a few words, what would you say?” *”What was the most important thing you learned from this book?” *”Did you like this book?”	the Richness of Retellings	information in sequence, and summarize what was most important. Answers varied widely, however.
21. Pearman (2008) Grade Level(s): 2 (n=54) 59% Caucasian; 2% African-American; 39% Hispanic; none	Purpose: Compare retell performance on electronically-based narrative passages with traditional print stories.	Initial: *“Tell me about the story” *“Can you tell me the story that you just read?” *“Pretend you are telling this story to your friend that has	10-point analysis of recalled information about characters, setting, theme, plot episodes, resolution, sequence	Inter-rater reliability = .84. For high- and medium-proficiency readers, there was no significant difference between text formats. For low reading proficiency students, mean retelling scores were significantly higher on electronic stories. There was no significant difference between mean scores on the first and second oral retellings (no practice effect).

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
ELL Student choice to read passage orally or silently.	Form of retell: Oral	never read it before. What will you tell them?" Follow-up: *"Can you tell me more?" *"What happened next?"		
22. Popplewell & Doty (2001) Grade Level: 2 (n = 71) Cannot determine if student read passage orally or silently	Purpose: Compare Four-Blocks framework instruction (Guided Reading, Self-Selected Reading, Writing, and Working with Words) and traditional basal instruction on students' ability to retell a narrative storybook and answer comprehension questions	Initial: Students asked to retell story, what that the story was about, and what they remembered about the events in the story. Follow-up: If student didn't start retelling immediately after reading, asked if remembered any part of the story or how the story began. If student stopped retelling,	10-point analysis of recalled information about characters, setting, theme, plot episodes, resolution, sequence.	Inter-rater reliability = .89. Students in the Four-Block group had significantly higher retell and comprehension question scores than students in the basal group. Students in Four-Block group remarked that the retell was "just like" what they usually did with their teacher during the Self-Selected Reading block individual conferences.

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
	Form of retell: Oral	asked "What happened next?" or "Anything else?" Asked 3 literal and 3 inferential comprehension questions.		
23. Richgels, McGee, Lomax, & Sheard (1987) Grade Level: 6 (n = 56) From a "wide range of abilities and backgrounds" Passages were read silently by student.	Purpose: Examine students' awareness of four expository text structures (collection, comparison-contrast, causation, problem-solution) and recall of texts written in those structures Form of retell: Written	Students were asked to write everything they could remember immediately after reading without looking back at the passage. Students were asked to write a summary immediately after engaging in a discussion with the examiner that followed one of the four structures.	Percent of predetermined idea units. Scale score of 0-7 on how well organization of recall resembled the text structure in the passage. Scale scores were converted into "full" = clearly the same structure as passage; "partial" = some lower-level information; "none" = ideas in random order or different structure from text.	Inter-rater reliability = .93 for percent of idea units; = .88 for scale of text structure match; = .90 for composition rating. Students recalled significantly more main ideas than details when passages were organized in comparison-contrast and problem-solution than when the passages were in scrambled order. Students recalled significantly more main ideas than details both when passages were organized in collection format and when the passages were in scrambled order. There were no significant differences in students' recalls of main ideas and details when passages were organized in causation. Significantly fewer students received "full knowledge" scores for causation recalls than for the other 3 structures. There were no significant differences among the numbers of students receiving "full knowledge scores" for collection, comparison-contrast, and problem-solution recalls. The more aware students were

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
				of text structures, the more likely their recalls reflected an understanding of the text. Following a structured discussion, significantly more students received "full knowledge" scores for compare-contrast compositions than the other 3 structures. Significantly more students received "full knowledge" scores for collection and problem-solution than causation compositions. The use of the passage structure in the reader's written recall was a less demanding task than writing a composition after engaging in a structured discussion.
24. van den Broek, Tzeng, Risden, & Trabasso (2001) Grade Level(s): 4, 7, 10, and undergraduate college (n = 240) No demographic information provided Passages were read silently by student.	Purpose: Determine the effect of question timing and student age/grade on recall of story propositions. Form of retell: written	Students were given a test booklet with space to record their recall in their own words.	Proportion of story propositions recalled that retained the general meaning from the original text.	Inter-rater reliability on free recall was .92. Students in grades 4 and 10 recalled significantly less information overall from the text when responding to probing questions during and after reading as compared to students who were not given any probing questions. Fourth-grade students' overall recall was most seriously impaired by the questioning conditions. When given questions during reading, students in grade 7 recalled similar amount of information as comparison no-question students, but recalled significantly less information if the questions were provided after the reading. College students recalled significantly more information than comparison students when provided questions during

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
				reading. Older students recalled more information overall than younger students. Students in grade 10 and college recalled similar amounts of information not specifically probed in the questions as did students in the no-question comparison. Students in grades 4 and 7 recalled significantly less information not specifically probed in the questions than students who were not given any questions. In contrast to the comparison no-question students, questioning condition students across the grade levels included more story propositions in their recalls that were also part of the answers to the probing questions.
25. Wright & Newhoff (2001) Grade Level(s): 3-7 (n = 30) 10 students with language-learning disabilities (LLD); 10 students without LLD matched by age; 10 students without LLD	Purpose: Compare the ability of students with and without LLD to retell and answer comprehension questions when reading or listening to narrative stories on the third-grade level.	Initial: Students asked to retell the story without referring back to it. Follow-up: 8 comprehension questions	Percent of full semantic content of story grammar parts (main setting, direct consequences, initiating events). Frequency count of complete and accurate answers to comprehension questions.	Inter-rater reliability = .95 - .98. Non-LLD students matched by chronological age produced more sentences, more verbatim information, and retold significantly more story grammar parts than the other two groups, which had no significant differences between them. Main settings were retold significantly more often than initiating events. There were no significant differences across the three groups in retelling performance when reading versus hearing the stories. However, students in all groups correctly and completely answered significantly more inferential comprehension questions when hearing the stories than when

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
matched by language ability Passage read orally by student in "story read" condition. Passage read to student by teacher in "story heard" condition.	Form of retell: Oral			reading them.
26. Zinar (1990) Grade Level(s): 5 (n=48) High- and low-comprehenders; high proportion of ethnic minorities in lower socioeconomic status Passage read orally by student.	Purpose: Determine the effect of explicit or implicit causal relationships in expository text upon recall of propositions Form of retell: Oral	Free recall: "Tell me everything you can remember from this story." Probed recall was individualized to elicit propositions not freely recalled.	Proportion of predetermined propositions recalled	Inter-rater reliability on free recall was .91 and on free + probed recall was .93. Students with higher comprehension ability recalled significantly more target information than students identified with low comprehension. There was no effect for passage type and no ability x passage type interaction. There were no significant differences in the amount of non-target information recalled. Both ability types recalled similar amounts of target information when probed recall was combined with free recall across passage types. High comprehenders were more likely to include causal information in free recalls when it was explicit rather than implicit,. Low comprehenders did not include causal relationships whether it was explicit or implicit.

Study	Purpose for and Form of Retell in Study	Prompt[a]	Scoring Procedure	Findings
				Both abilities included similar amounts of causal relationships when they were probed.

Abbreviations: LD = learning disabilities; EBD = emotional and/or behavioral disorders; LA = low achieving; ELL = English language learners; SLI = speech language impairment; NR = not reported; NARA = Neale Analysis of Reading Ability; ORF = oral reading fluency; SAT-7 RC = Stanford Achievement Test 7th edition reading comprehension subtest; SAT-7 WS = Stanford Achievement Test 7th edition word study subtest; VIP = Vital Indicators of Progress

[a]Notes: Prompts enclosed in quotations are the exact wording as reported in the study; prompts not in quotations are based on the description provided in the study

Table A2

Retell Measures

Assessment	**Prompt**[a]	**Scoring Procedure**	**Reliability**	**Validity**
1. *The Critical Reading Inventory, 2nd ed.* Applegate, Quinn, & Applegate (2008) Grade Level(s): PreK – 12 Form of Retell: Oral Text type(s): Narrative and expository	(Passage may be read orally or silently) Initial: "Tell me about what you just read and what you thought about it." Follow-up: "Tell me what you thought about the passage."	Score from 0 to 1 on eight pre-determined story structure elements (key characters and setting, character's problem/goal, problem-solving or goal-meeting process, and personal response) for narratives or eight pre-determined macro- and micro-concepts for expository passages. Items are weighted in the calculation of the Final Score with a scale of 0 (vague idea of the topic) to 4 (perfect, well supported retelling)	Trained test administrator and expert scorer agreed on the scoring of 92.5% of the retellings. Norming sample: • 215 students (93 in grades 1-3, 95 in grades 4-8, 27 in grades 9-12) • 105 males and 110 females • 150 Caucasian, 38 Black, 15 Hispanic, 6 Asian, 6 other • 157 public school,	The retell score and the total comprehension item percentage for each narrative passage had a correlation coefficient of .51 (*p* < .001). The retell score and the total comprehension item percentage for each informational passage had a correlation coefficient of .43 (*p* < .001). Norming sample: (see information in reliability column)

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
			21 private, 38 parochial • 56 high achieving in reading, 68 average, 89 low, 2 information n/a • 30 in special education, 5 gifted, 14 ELL	
2. *BADER Reading and Language Inventory* Bader & Pearce (2009) Grade Level(s): K-12 + adult Form of Retell: Oral	(Retell is asked after the silent reading and listening comprehension portions.) Initial: For pre-primer through grade 4 passages, students are prompted to "Please retell the story." In grades 5	Number of unprompted plus prompted ideas (prompting is to continue until students meets a minimum number of ideas for each passage). Judgment of whether the retell was organized (yes/no).	Correlations of passage equivalents ranged from .82 to .85 for silent reading passages. Inter-rate reliability for the silent reading passage scoring was 90%. <u>Norming samples</u>: • 30 elementary students, 30 middle	1) Correlation of scores to school reading specialists' judgments of students' reading levels was .93. 2) Correlation of scores to classroom teachers' judgments of students' reading levels was .89. <u>Norming samples</u>: 1) 27 students in grades

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
Text type(s): PreK – 4: narrative Grades 5-12: mixture of narrative and expository (not clearly separated)	through 12 passages, the directions indicate students should be asked to retell all the information remembered about the passage. Follow-up: "Can you think of something else?"; "What else happened?"; "Tell me more." Ask scripted comprehension questions for any information not provided in free recall.		school students, 30 secondary students, and 30 adults in passage equivalents study. • Inter-rater reliability established with one elementary reader	1 – 5 "with diverse ethnic and racial backgrounds fom low-income families" (p. 167) . 2) 30 students in grades 4-8

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
3. *Developmental Reading Assessment* Beaver (2003; 2006) Grade Level(s): K-3; 4-8 Form of Retell: Oral Text type(s): Narrative with alternative materials containing narrative biography and expository passages	(At lower levels, the examiner reads a portion of the passage and then has the student read orally. At higher levels, the student reads silently and then a portion is read orally.) Initial: Student asked to retell the story. Follow-up: Ask scripted comprehension questions.	Rubric score of 6 to 24 based on inclusion of main ideas, key facts, sequencing of information, characters or topics, specificity, level of interpretation, and relation of free recall to prompted recall.	1) Test re-test reliability in the .90 range. 2) Inter-rater reliability of overall scoring on assessment ranged from .74 to .80. Norming samples: 1) 306 students in grades 1- 3 at four elementary schools; 356 bilingual students in grades 1 – 3 at eight elementary schools 2) 87 teachers from 10 states scored three or more students from their individual	1) Content validity determined by percentage of teachers agreeing that assessment is helpful in evaluating students' reading progress (89%) and determining instructional goals (82%). 2) DRA independent reading levels were moderately correlated with: A) ITBS comprehension grade-equivalent scores (r = .83), B) ITBS national curve equivalent scores for comprehension (r = .675), and C) Lexile

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
			classes; two additional blind raters also scored the students	measures (*r* = .69). Norming samples: 1) 84 teachers from test development site in Ohio 2A) 284 students in grades 1 - 3 in four elementary schools 2B) 2,470 second-grade students from a large urban/suburban district 2C) 1,140 second- and third-grade students from a large, suburban district in Florida
4. *Comprehensive Reading Inventory* Cooter, Flynt, & Cooter (2007)	(Retell is only asked after the silent reading portion.) Initial/Unaided:	Number of pre-determined story elements or ideas related to the text structure correctly retold.	Overall test-retest Pearson product-moment correlations ranged from .67 to .93. The retell	NR

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
Grade Level(s): PreK – 9 Form of Retell: Oral Text type(s): Narrative and expository	"Tell me about the story you just read." Follow-up/Aided: Ask scripted comprehension questions related to narrative story elements (character/ characterizaton, setting, story problem, problem resolution, theme) or expository text structure.		measure was not separated from the overall analysis of all assessment components. Norming sample: • 714 students in grades K – 8 at 30 different schools • 51.2% male and 48.8 % female • 98% eligible for free or reduced lunch • 98.7% minority • 12.9% in special education, 1.3% gifted, 16% ELL	
5. *Dynamic Indicators of Basic Early Literacy*	(Passages are read orally.)	Total number of words produced, except: exclamatory sounds, singing, recitations of the ABCs, repetitions	Alternate form reliability ranged from .68 - .72.	Concurrent validity of retell component established with ORF

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
Skills (*DIBELS*) Good & Kaminski (2002a) Grade Level(s): 1-6 Form of Retell: Oral Text type(s): Narrative and expository	Initial: "Please tell me all about what you just read. Try to tell me everything you can. Begin." Follow-up: "Try to tell me everything you can."	of the same statement, irrelevant sidebars/stories.	Norming sample: NR	(.59). Predictive validity established with the Oregon State Assessment Test (.50). Norming sample: NR
6. *Vital Indicators of Progress (VIP)* Good & Kaminski (2002b) Grade Level(s): 1-6 Form of Retell: Oral	(Passages are read orally.) *Initial: Students are asked to retell as many details as they can recall from the passage they just read.	*Number of words used to accurately retell the story within 1 minute	*Alternate-form reliability was .57. Using the Spearman-Brown prophecy formula, the estimated reliability of 3 passages for retell fluency was calculated to be .80.	*Average retell fluency passage scores correlated .51 with the Broad Reading Cluster (26% of variance explained) and .61 with the VIP oral reading fluency average. Adding retell fluency to

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
Text type(s): Narrative and expository	Follow-up: Students are encouraged to tell everything they can recall.		Norming sample: 86 first grade students from predominately low-income, Title I populations. Ninety-percent of students were African American and 100% received free or reduced-priced lunches.	the prediction of Broad Reading Cluster standard scores added a very small amount of additional explained variance (about 1%) to that explained by VIP oral reading fluency alone. Norming sample: (see Reliability Norming Sample)
7. *Basic Reading Inventory, 10th ed.* Johns (2008) Grade Level(s): PreK – 12 Form of Retell:	(Some passage are read orally and others silently) Initial: “Tell me about (name of passage) as if you were telling it to	Option 1: Scale score of “none” to “high degree” for 12 items reflecting textual information, metacognitive awareness, strategy use/involvement with text, and language development. Option 2: Points for inclusion of	1) Basic Reading Inventory was moderately correlated to two other commercially prepared reading inventories (r = .72 for instructional	NR

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
Oral Text type(s): Narrative and expository	someone who has never heard it before.” Follow-up: “What comes next?”; “Then what happened?” Step-by-step prompting (when needed): “Who was the passage about?”; “When did the story happen?”; “Where did the story happen?”; “What was the main character’s problem?”; “How did he/she try to solve the problem? What was done	narrative story structure items (characters, setting, theme, plot episodes, resolution, and sequence). Option 3: Classification of independent, instructional, or frustration level in expository passage based on inclusion of text structure information, organization, accuracy, completeness of main ideas and details. Option 4: Scale score of 1 to 5 based on generalizations beyond text, thesis statement, major points, supporting details, relevant supplementations, coherence, completeness, and comprehensibility.	level; *r* = .73 for frustration level; *r* = .64 for independent level) 2) Based on results from the Basic Reading Inventory and a clinician-constructed inventory, 66% of students placed in the same instructional level and 33% of the students were placed within one grade level of each other. 3) Inter-rater agreement on four comprehension scoring tasks was 79 – 81%	

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
	first/next?"; "How was the problem solved?"; "How did the story end?"		Norming samples: 1) 75 students in grade 4 2) 33 students ages 7 to 15 at reading levels pre-primer to sixth-grade 3) 49 pre-service teachers in second undergraduate reading course (research conducted by test developer)	
8. *Stanford Reading Diagnostic Test (SDRT, 4th ed)..* Karlsen & Gardner	(Retell is a separate, informal subtest using one narrative passage that can be read either orally or	Rubric rating of "none" to "thorough" for story structure elements (introduction, setting, characters, problem, plot/events, resolution, theme, sequence) and	*[Information provided was for SDRT in general, not specific to retell subtest. Retell subtest*	*[Information provided was for SDRT in general, not specific to retell subtest. Retell subtest was not*

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
(1996) Grade Level(s): 1-12 + adult Form of Retell: Oral or Written Text type(s): Narrative	silently.) Initial written: "…retell the story in writing as if you are telling it to a friend who has never heard it before. You should write as much of the story as you can remember and tell the story in the right order." Initial oral: "…retell the story to me as if you are telling it to a friend who has never heard it before. You will tell me as much of the story as you	reading process (inclusion of literal and inferential information, critical analysis of story, summarization, generalization, relevant prior knowledge, and expressiveness). Suggested responses are offered for story elements only.	*was not disaggregated in the data.]* Internal consistency determined with Kuder-Richardson Reliability coefficients (*r* = .79 - .97; SEM = 1.6 - 4.3). Alternate forms reliability coefficients for the comprehension portion of SDRT reported as *r* = .71 - .82 (SEM = 3.8 – 4.4) Norming sample: 33,000 students from 400 school districts across the nation, and	*disaggregated in the data.]* Construct validity determined by correlation of SDRT to Otis-Lennon School Ability Test (grades 2-12, *r* = .43 - .95). Continuity of SDRT across grade levels established with correlations between corresponding subtests (*r* = .59 - .87) Norming sample: 33,000 students from 400 school districts across the nation, and another 7,000 students in three equating programs. Sample was

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
	can remember and tell the story in the right order." Follow-up: "You have remembered many things. Can you think of anything else about the story?"		another 7,000 students in three equating programs. Sample was chosen to be representative of the national school population and was stratified on the basis of geographic region, socioeconomic status, urbanicity, ethnicity, and public versus private school.	chosen to be representative of the national school population and was stratified on the basis of geographic region, socioeconomic status, urbanicity, ethnicity, and public versus private school.
9. *Qualitative Reading Inventory (QRI) – 4* Leslie & Caldwell (2006) Grade Level(s): preK-6; upper	(Passage may be read orally or silently) Initial: Retell the passage as if it were being told to someone who had never read or heard it	Number and order of pre-determined idea units recalled. In narrative passages, idea units are based on goals, setting, events, and resolution. In informational passages, idea units are based on main ideas and details. Overall interpretation of retell is based on accuracy, completeness,	Inter-rater reliability of propositions identified was .98 or higher. Alternate form reliability was in the .90 range. Important propositions were	1) Criterion validity assessed with California Achievement Test (grads 1-3) or the Iowa Test of Basic Skills (grades 3-8). Correlations ranged

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
middle school; high school Form of Retell: Oral Text type(s): Narrative and expository	before. Follow-up: Ask if there is anything else the student would like to say; Draw students' attention to the title and ask whether student can remember what the author wrote about it.	organization, and specificity of important information.	recalled by 20% of students in field test or were identified by 50% of teachers in field test. Percent of idea units recalled on narrative passages ranged from 17 – 41%. Percent of idea units recalled on expository passages ranged from 13 – 31%. Correlations between retell scores and comprehension question scores were provided for individual passages (*r* provided ranged from .34 to .60). All correlations were low	from .27 (grade 6) to .85 (grade 1) in narrative text with no significant correlation at grades 6 or 7. Correlations for expository text at grades 5 through 8 ranged from .28 (grade 7) to .55 (grade 9) with no significant correlation at grade 7. 2) QRI was moderately correlated with Woodcock Reading Mastery Test passage comprehension (*r* = .75). 3) Word identification and rate were correlated with reading

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
			to moderate, but authors indicated high variability in correlations, particularly at lower grade levels. No overall correlation coefficient provided. <u>Norming sample(s)</u>: NR	comprehension at preK, second-, third-, and fourth-grades (no coefficients provided). 4) Prior knowledge was correlated with retelling comprehension at kinder – upper middle school (no coefficients provided). <u>Norming sample(s):</u> 1) 266 students in grades 1-8 2) NR 3) NR 4) NR
10. *Informal Reading Inventory* Roe & Burns	(Passage may be read orally or silently)	No formal procedure. Suggestion for teachers to use a rating of 1 (poorly) to 5 (very well) on scorer's	NR	NR

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
(2007) Grade Level(s): PreK – 12 Form of Retell: Oral Text type(s): Mixture of narrative and expository (not clearly separated)	Initial: "Retell this selection for someone who has not read it, so that the person would understand it as well as you do." Follow-up: "Can you tell anything else that it said?" Ask scripted comprehension questions for any information not provided in free recall.	guiding questions about completeness, accuracy, main ideas, details, summarizing statements, organization, implicit and explicit information. Suggested rubric for narratives with 1 to 3 rating on characterization, setting, plot, and conflict.		
11. *Classroom Reading Inventory, 10th ed.* Silvaroli &	(Passage is read aloud by teacher and then read silently by student.)	Scale score of 1 to 3 for each category of information (characters, problems, outcomes). Overall score ranges indicate excellent	NR	NR

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
Wheelock (2004) Grade Level(s): preP – 8 Form of Retell: Oral Text type(s): Narrative	Initial: Allow free recall based on teacher modeling with a practice passage. Follow-up: Ask up to three scripted questions on characters, problem(s), and outcome(s) / solution(s).	comprehension, needs assistance, or inadequate comprehension.		
12. *Analytical Reading Inventory, 8th ed.* Woods & Moe (2007)	(Passage may be read orally or silently) Initial: "Retell everything you can remember from the	Score of "all," "some," or "none" for inclusion of narrative story structure elements (main character, time and place, problem, plot details in sequence, turning point, and resolution) or expository text	Teachers in pilot study had some variation in comprehension scoring but great consistency in	NR

Assessment	Prompt[a]	Scoring Procedure	Reliability	Validity
Grade Level(s): preK – 9 Form of Retell: Oral Text type(s): Narrative and expository	passage, and I will write down what you say." Follow-up: "Can you tell me more?"; "And?"; "More?" Final: "In one or two short sentences, tell what this passage is about."	elements (description, collection, causation, problem/solution, and comparison). Retells judged for completeness, organization, and sentence structure, style, word choice. Summary statement judged for adequacy.	determination of overall reading level (percent agreements not provided). Norming sample: 9 reading teachers from one district in Indiana listened to training materials featuring readers at independent, instructional, and frustration levels.	

Abbreviations: NR = not reported; ELL = English language learner

[a]Note: Prompts enclosed in quotations are the exact wording as reported in the study; prompts not in quotations are based on the description provided in the study

*Note: VIP reportedly was developed by the same researchers as DIBELS and, therefore, parallels DIBELS in its administration and scoring procedures. Reliability and validity of the VIP retell fluency was determined in the Roberts et al., 2005, study included in Table 1. Reliability and validity were not reported in the technical information provided with the VIP.

Table A3

Matrix of Text Types, Reading Conditions, and Retell Conditions

		Oral Reading	**Silent Reading**	**Oral and Silent Reading**	**Listening**	**Listening and Oral Reading**	**Unknown**	**Oral or Silent Reading**
Oral Retell	Narrative Text	1	2		1	1	1	1
	Expository Text	2			1	1	2	
	Both		1					
	Unknown	2						
Written Retell	Narrative Text		3					
	Expository Text		2					
	Both							
	Unknown		1				1	
Both	Narrative Text			1				
	Expository Text		1					1
	Both							
	Unknown							

Number represents the number of studies employing that combination of text type and reading format.

Table A4
Matrix of Initial and Follow-up Prompts

		Follow-up Prompt			
		Scripted follow-up questions	General prompting to continue giving information	General prompting increasing to more structured prompting	General prompting plus comprehension questions
Initial Prompt	Tell me about what/ the story you just read.	1	1 (what you thought about the passage)		
	Tell me/write about the story as if telling a friend/someone who has never read it.	1	1	3	1
	Retell the story/passage	2		1	1
	Retell as many details as/ everything you can recall/remember.		4		
	[Free recall based on teacher modeling.]	1			
	Retell the story, what it was about, and what you remember about the events.	1			1

Number represents the number of existing retell measures using that combination of initial and follow-up prompt.

References for Appendix A

*Applegate, M. D., Quinn, K. B., & Applegate, A. J. (2008). *The Critical Reading Inventory: Assessing students' reading and thinking* (2nd ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

*Bader, L. A., & Pearce, D. L. (2009). *BADER Reading and Language Inventory*. Boston: Allyn & Bacon.

*Beaver, J. M. (2006). *Developmental Reading Assessment, Grades K-3 (2nd ed.)*. Parsippany, NJ: Pearson Education, Inc.

*Beaver, J. M. (2003). *Developmental Reading Assessment, Grades 4-8 (2nd ed.)*. Parsippany, NJ: Pearson Education, Inc.

*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*(2), 137-164.

*Carlisle, J. F. (1999). Free recall as a test of reading comprehension for students with learning disabilities. *Learning Disability Quarterly, 22*(1), 11-22.

Catts, H. W., Adlof, S. M., & Weisner, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech, Language, and Hearing Research, 49*(2), 278–294.

*Cooter, R. B., Flynt, E. S., & Cooter, K. S. (2007). *Comprehensive Reading Inventory: Measuring reading development in regular and special education classrooms.* Upper Saddle River, NJ: Pearson Prentice Hall.

Cordon, L. A., & Day, J. D. (1996). Strategy use on standardized reading comprehension tests. *Journal of Educational Psychology, 88*(2), 288-295.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contribution of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10*(3), 277-299.

*Doty, D. E., Popplewell, S. R., & Byers, G. O. (2001). Interactive CD-ROMS storybooks and young readers' reading comprehension. *Journal of Research on Computing in Education, 33*(4), 374-384.

Doyle, P. J., McNeil, M. R., Park, G., Goda, A., Rubenstein, E., Spencer, K., Carroll, B., Lustig, A., & Szwarc, L. (2000). Linguistic validation of four parallel forms of a story retelling procedure. *Aphasiology, 14*(5/6), 537-549.

Duffelmeyer, F. A., & Duffelmeyer, B. B. (1987). Main idea questions on informal reading inventories. *The Reading Teacher, 41*(2), 162-166.

Ferstl, E. C., Walther, K., Guthke, T., & Yves von Cramon, D. (2005). Assessment of story comprehension deficits after brain damage. *Journal of Clinical and Experimental Neuropsychology, 27*, 367-384.

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2007). *Learning disabilities: From identification to intervention*. New York: Guilford.

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

*Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*(1), 45-58.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239-256.

*Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20-28.

*Gagne, E. D., Bing, S. B., & Bing, J. R. (1977). Combined effect of goal organization and test expectations on organization in free recall following learning from text. *Journal of Educational Psychology, 69*(4), 428-431.

*Gambrell, L. B., & Jawitz, P. B. (1993). Mental imagery, text illustrations, and children's story comprehension and recall. *Reading Research Quarterly, 28*(3), 265-276.

*Gambrell, L. B., Pfeiffer, W. R., & Wilson, R. M. (1985). The effects of retelling upon reading comprehension and recall of text information. Journal of Educational Research, 78(4), 216-220.

*Gambrell, L. B., Koskinen, P. S., & Kapinus, B. A. (1991). Retelling and the reading comprehension o proficient and less-proficient readers. *Journal of Educational Research, 84*(6), 356-362.

*Good, R. H., & Kaminski, R. A. (2002a). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoreon.edu.

*Good, R. H., & Kaminski, R. A. (2002b). *Vital Indicators of Progress (VIP)*. Dallas, TX: Voyager Expanded Learning, Inc.

Goodman, K. S. (2006). *The truth about DIBELS*. Portsmouth, NH: Heinamann.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading and reading disability. *Remedial and Special Education, 7*(1), 6-10.

*Hansen, C. L. (1978). Story retelling used with average and learning disabled readers as a measure of reading comprehension. *Learning Disability Quarterly, 1*(3), 62-69.

Hintze, J. M., Shapiro, E. S., Conte, K. L., & Basile, I. M. (1997). Oral reading fluency and authentic reading material: Criterion validity of the technical features of CBM survey-level assessment. *School Psychology Review, 26*(4), 535-553.

*Horowitz, R., & Samuels, S. J. (1985). Reading and listening to expository text. *Journal of Reading Behavior, 17*(3), 185-198.

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children, 59*(421-432).

*Johns, J. L. (2008). *Basic Reading Inventory: Pre-primer through grade twelve and early literacy assessments.* (10th ed.). Dubuque, IA: Kendall/Hunt Publishing Co.

Johnston, P. (1981). Implications of basic research for the assessment of reading comprehension. (Technical Report No. 206, National Institute of Education). Washington, DC Ed-201987.

*Karlsen, B., & Gardner, E. F. (1996). *Stanford Reading Diagnostic Test* (*SDRT*, 4th ed.). San Antonio, TX: Harcourt Brace & Co.

Keenan, J. M, Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*(3), 281-300.

Kintsch, W., & van Dijk, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review, 85*, 363-349.

Klesius, J. P., & Homan, S. P. (1985). A validity and reliability update on the informal reading inventory with suggestions for improvement. *Journal of Learning Disabilities, 18*(2), 71-76.

Klingner, J. K. (2004). Assessing reading comprehension. *Assessment for Effective Intervention, 29*(4), 59-70.

*Kouri, T., & Telander, K. (2008). Children's reading comprehension and narrative recall in sung and spoken story contexts. *Child Language Teaching and Therapy, 24*(3), 329-349.

*Leslie, L., & Caldwell, J. (2006). *Qualitative Reading Inventory* (4th ed.). Boston: Pearson Education, Inc.

Lindo, E.J. (2006). The African American presence in reading intervention experiments. *Remedial and Special Education, 27*, 3, 148-153.

*Loyd, B. H., & Steele, J. L. (1986). Assessment of reading comprehension: A comparison of constructs. *Reading Psychology: An International Quarterly, 7*, 1-10.

Madelaine, A., & Wheldall, K. (2005). Identifying low-progress readers: Comparing teacher judgment with a curriculum-based measurement procedure. *International Journal of Disability, Development, and Education, 52*(1), 33-42.

*Mason, L. H., Snyder, K. H., Sukhram, D. P., & Kedem, Y. (2006). TWA + PLANS strategies for expository reading and writing: Effects for nine fourth-grade students. *Exceptional Children, 73*(1), 69-89.

*McGee, L. M. (1982). Awareness of text structure: Effects on children's recall of expository text. *Reading Research Quarterly, 17*, 581-590.

McNeil, M. R., Doyle, P. J., Fossett, T. R. D., Park, G. H., & Goda, A. J. (2001). Reliability and concurrent validity of the information unit scoring metric for the story retelling procedure. *Aphasiology, 15*(10/11), 991-1006.

*Moss, B. (1997). A qualitative assessment of first graders' retelling of expository text. *Reading Research and Instruction, 37*, 1-13.

Newcomer, P. L. (1999). *Standardized Reading Inventory* (2nd ed.). Austin, TX: Pro-Ed, Inc.

Nicholas, L. E., & Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech and Hearing Research, 36*(2), 338-350.

Nilsson, N. L. (2008). A critical analysis of eight informal reading inventories. *The Reading Teacher, 61*(7), 526-536.

Olive, T., & Kellogg, R. T. (2002). Concurrent activation of high- and low-level production processes in written composition. *Memory & Cognition, 30*, 594-600.

Otto, W., Barrett, T. C., & Koenke, K. (1968, April). *The assessment of children's statement of the main idea in reading*. Paper presented at the International Reading Association Conference, Boston, MA.

*Pearman, C. J. (2008). Independent reading of CD-Rom storybooks: Measuring comprehension with oral retellings. *The Reading Teacher, 61*(8), 594-602.

Peverly, S. T., Ramaswamy, V., Brown, C., Sumowski, J., Alidoost, M., & Garner, J. (2007). What predicts skill in lecture note taking? *Journal of Educational Psychology, 99*(1), 167-180.

*Popplewell, S. R., & Doty, D. E. (2001). Classroom instruction and reading comprehension: A comparison of one basal reader approach and the four-blocks framework. *Reading Psychology, 22*, 83-94.

Pressley, M., Ghatala, E. S., Woloshyn, V., & Pirie, J. (1990). Sometimes adults miss the main ideas and do not realize it: Confidence in responses to short-answer and multiple-choice comprehension questions. *Reading Research Quarterly, 25*(3), 232-249.

*Rasinski, T. V. (1990). Investigating measures of reading fluency. *Educational Research Quarterly, 14*(3), 36-44.

*Richgels, D. J., McGee, L. M., Lomax, R. G., & Sheard, C. (1982). Awareness of text structures: Effects on recall of expository text. *Reading Research Quarterly, 22*(2), 177-196.

*Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42*(4), 546-567.

Roberts, G., Good, R., & Corcoran, S. (2005). Story retell: A fluency-based indicator of reading comprehension. *School Psychology Quarterly, 20*(3), 304-317.

*Roe, B. D., & Burns, P. C. (2007). *Informal Reading Inventory*. Boston: Houghton Mifflin Co.

*Shinn, M. R., Good, R., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*(3), 459-479.

Savage, R. (2006). Reading comprehension is not always the product of nonsense word decoding and linguistic comprehension: Evidence from teenagers who are extremely poor readers. *Scientific Studies of Reading, 10*(2), 143-164.

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.), Advances in applied psycholinguistics, Vol. 1: Disorders of first-language development (pp. 142-175). New York: Cambridge University Press.

Schatschneider, C., Buck, J., Wagner, R., Hassler, L., Hecht, S., & Powell-Smith, K. (2004). A multivariate study of individual differences in performance on the reading portion of the Florida Comprehensive Assessment Test: A brief report. (Technical Report No. 5, Florida Center for Reading Research), Tallahassee, FL.

Seifert, T. L. (1994). Brief research report: Enhancing memory for main ideas using elaborative interrogation. *Contemporary Educational Psychology, 19*, 360-366.

Shinn, M., Good, R., Knutson, N., Tilly, W., & Collins, V. (1992). Curriculum-based measurement reading fluency: A confirmatory factor analysis of its relation to reading. *School Psychology Review, 21*, 459-479.

*Silvaroli, N. J., & Wheelock, W. H. (2004). *Classroom Reading Inventory* (10th ed.). Boston: McGraw Hill.

Snyder, L., Caccamise, D., & Wise, B. (2005). The assessment of reading comprehension: Considerations and cautions. *Topics in Language Disorders, 25*(1), 33-50.

Spooner, A. L. R., Baddeley, A. D., & Gathercole, S. E. (2004). Can reading accuracy and comprehension be separated in the Neale Analysis or Reading Ability? *British Journal of Educational Psychology, 74*, 187-204.

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*(3), 407-419.

Talbott, E., Lloyd, J. W., & Tankersley, M. (1994). Effects of reading comprehension interventions for students with learning disabilities. *Learning Disability Quarterly, 17*, 223-232.

Valencia, S. W., & Buly, M. R. (2004). Behind test scores: What struggling readers really need. *Reading Teacher, 57*(6), 520-531.

*van den Broek, P., Tzeng, Y., Risden, K., Trabasso, T., & Basche, P. (2001). Inferential questioning: Effects on comprehension of narrative texts as a function of grade and timing. *Journal of Educational Psychology, 93*(3), 521-592.

Wheldall, K., & Madelaine, A. (2000). A curriculum-based passage reading test for monitoring the performance of low-progress readers: The development of the WARP. *International Journal of Disability, Development, and Education, 47*(4), 371-382.

Wiley, H. I., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education, 26*(4), 207-214.

*Woods, M. L., & Moe, A. J. (2007). *Analytical Reading Inventory: Comprehensive standards-based assessment for all students, including gifted and remedial* (8th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

*Wright, H. H., & Newhoff, M. (2001). Narration abilities of children with language-learning disabilities in response to oral and written stimuli. *American Journal of Speech-Language Pathology, 10*(3), 308-319.

*Zinar, S. (1990). Fifth-graders' recall of propositional content and causal relationships from expository prose. *Journal of Reading Behavior, 22*(2), 181-199.

The asterisk (*) indicates articles or assessments synthesized for this paper.

TMSFA - Word Reading Fluency (Word Lists)

MATERIALS: Stopwatch, stimulus

DESCRIPTION: Student will be administered three word lists that vary in difficulty. This measure assesses the number of real words that can be accurately identified within 60 seconds.

TIME LIMIT: 60 seconds

SCORING:

- Slash across all incorrect words
- Circle the last word read at 60 seconds
- Note the time the last word was read if the student finished in less than 60 seconds.
- If the student skips a word, count it as an error.
- If the student hesitates for more than 3 seconds on a word, mark it as incorrect and instruct him/her to go to the next word.
- If a student self-corrects a word, write "SC" above the word and count as correct.

TELEFORM: *Record the following*

- Last Word Read
- Number of Words Read Incorrectly
- Number of Words Read Correctly
- Time in Seconds –enter actual time taken to administer word lists

---

DIRECTIONS:

First Word List:

Say, **I want you to read some lists of words as fast as you can. Begin at the top, and read down the list as fast as you can until I tell you to stop. If you come to a word you cannot read, just skip it and go to the next word. If you skip more than one word, point to the word you are reading next. Do you understand? Okay, you will begin as soon as I turn the page.**

- TIMER - Start timing when the student says the **first word**.
- ERRORS - Slash through any words that are misread, skipped, or not read within 3 seconds.

- If the student hesitates for more than 3 seconds on a word, mark it incorrect and say, **Go on.**
- After 60 seconds, say, **Stop.** Circle the last word read.
- If the student finishes all the words before the time is up, note the time it took them.
- If, before the time is up, the student indicates that he or she cannot read any more words, say, **Look over the whole list to see if there are any more words you can read.** If the student indicates he or she can read no more words, circle the last word, record the time, and stop testing.

For Second and Third Word Lists:

Say, **Now try this list. Ready? Begin.** Follow guidelines listed above.

TMSFA - Passage Reading Fluency

MATERIALS:	Timer, stimulus
DESCRIPTION:	Assesses the number of real words the student can accurately and quickly read within 60 seconds and how well he/she comprehends the text.
TIME LIMIT:	60 seconds

DIRECTIONS:

1. **Please read this** *(point to the passage and read the title)* **out loud. If you get stuck, I will tell you the word so you can keep reading. When I say 'stop', I will ask you some questions about what you read, so do your best reading. Start here.** *(Point to the 1st word)*. **Begin.**
2. Start timer when the student says the 1st word of the passage. The *title is not counted*. If the student fails to say the first word after 3 seconds, tell them the word, mark it as incorrect, then start the timer.
3. If the student does not provide a word within 3 seconds, say the word and mark it as incorrect.
4. Follow along on the examiner protocol. Put a slash **(/)** over words read incorrectly or skipped. Put **(SC)** over words that the student self corrects. Write all words the student inserts.
5. At the end of **60 seconds**, say, **Stop**. Circle the last word read, stop and reset the timer.
6. **MAIN IDEA:** After each passage has been administered, **remove/cover the stimulus** and ask, **Tell me in your own words what this passage is mostly about.** Record the student's response. If the student gives a one-word response and/or repeats just the title, you may give the prompt of **Tell me more** one time only. Use the scoring guidelines below to give a score from 0-3 to indicate if the student response suggests that he/she comprehended what was read.
7. Place the next passage in front of the student. Say, **Let's try another passage. Please read this (point to passage and read title). Ready? Begin.** Follow testing procedures as outlined in #4-8, and do the same for the last passage.

Underground Town
Expository; Lexile: 700
Source: SDAA 2004

How would you like to live underground? Many families in the town of
14 Coober Pedy, Australia, do just that. Their houses, called dugouts, are carved out
29 of the earth. They are similar to regular houses. They have kitchens, bedrooms,
43 and living rooms, but they have few windows. Most of the light in the houses is
60 artificial. It comes from lamps and overhead lights instead of direct sunlight. The
76 roofs of the homes are made of layers and layers of dirt.

80 People in the town build dugouts because of where they live. Coober Pedy is
96 in a desert in southern Australia. Temperatures can climb as high as 120 degrees
110 in the summer. They can plunge as low as 32 degrees in the winter. Dust storms
127 and swarms of flies can make life miserable. Underground, though, it is always a
140 comfortable 75 degrees. People don't even need fans.

146 Why would people want to live in such a place? Coober Pedy is an opal-
162 mining town. Opals are colorful stones used for jewelry. The mines in Coober
176 Pedy produce most of the world's opals.

181 settlers in Coober Pedy realized that they could avoid the harsh temperatures by
194 building their homes underground. Today almost half of the 3,500 people in the

209 town live in dugouts. Restaurants, schools, and other buildings are also

222 underground. People in Coober Pedy enjoy their lives “down under.”

229

Let's Do it Again
Narrative; Lexile: 840
Source: SDAA 2005

My heart was beating so loudly that I was sure everyone could hear it over
15 the slow rumbling of the motor. I jumped into the water and put on my skis.
31 Slowly the boat crept forward, tightening up the ski rope. I held on for dear life to
48 the handle on the end of the rope while Mom smiled encouragingly at me from
63 the back of the boat.

68 I was trying very hard to recover my earlier feelings of excitement about
81 learning to water-ski. “Whose bright idea was this anyway?” I asked myself
94 anxiously. I sat in the cool water bobbing gently in my bright orange life jacket. I
110 tried to keep the tips of my water skis pointing up out of the water as I had been
129 shown. A wave of fear washed over me. There were just too many instructions to
144 remember. My little sister Nikki cheered as she jumped up and down in the back
159 of the boat next to Mom.

165 Nikki had learned to water-ski at a very young age. I, on the other hand,
180 always liked underwater sports such as scuba diving. Moving on top of the water
194 was going to be very different for me. But once I mastered this, we would have
210 another activity that the whole family could enjoy together.

219 “Deep breath,” I reminded myself. Dad pulled back the lever to open up the
233 throttle. The motor roared to life. “Here we go,” I thought wildly.

245 Mom gave me a big thumbs-up, and the boat lurched forward and gave a
262 mighty pull. I pushed up on my legs as hard as I could and let out a yell. I was
279 actually standing on my skis, skimming across the water, but not for long. I fell
294 forward and landed facedown in the water. Thank goodness I remembered to let
307 go of the rope. My skis came off, and my life jacket kept me floating on the
324 surface of the lake.

328 “I don’t believe it,” I thought, flipping over to my back with a grin. “I almost
344 felt like I was flying.”

349 “Let’s do it again,” I called to Dad as he circled the boat around to pick me
368 up.

369

Spreading Wildflowers

Expository; Lexile: 910
Source: TAAS 2001

Claudia Taylor was born in Karnack, Texas, in 1912. As a young child she
14 was given the nickname Lady Bird. She grew up in the country, and it was there
30 that her lifelong love of nature began. Throughout her childhood and adult years,
43 she has enjoyed being outdoors, looking for the beautiful flowers that grow
55 naturally in open fields.

59 In 1929 the state of Texas started a wildflower program. The highway
71 department waited for the flowers to go to seed before they were mowed. Then
85 the seeds would spread and grow into plants the next year. Lady Bird enjoyed
99 exploring the countryside in search of different wildflowers. She continued to do
111 so after moving to Austin in 1930 to attend the University of Texas. Four years
126 later Lady Bird married Lyndon B. Johnson.

133 In 1964 Lyndon Johnson was running for President of the United States. As
146 he and his wife traveled around the country, Lady Bird saw beauty as well as
161 blight. Some areas suffered from neglect and ugliness. When Lady Bird's
172 husband won the election, she wanted to do something to make the nation's
185 capital look more beautiful. The following year she found a way to do that.

199 Lady Bird helped set up the Committee for a More Beautiful Capital. She
212 was chosen to head the group of volunteers. They met once a month at the White
228 House to discuss ideas and make plans. They decided their program could be
241 successful only if people in the community were willing to get involved. To
254 attract attention, volunteers planted flowers around the city in hundreds of places
266 that many people passed each day. They encouraged businesses to plant grass,
278 shrubs, and flowers. They organized cleanups and fix-up projects in
288 neighborhoods. They also tried to improve school yards and playgrounds. The
299 committee gave awards each year to neighborhoods, businesses, and public
309 spaces.

310 The ideas of the committee quickly spread across the country. Some states
322 began setting up their own programs to preserve flowers and to plant new ones.
336 Thanks to Lady Bird, many of these programs included wildflowers. In the state
349 of Texas, people continued to strengthen the program that had been adopted
361 almost 40 years before the committee began its work.

370 The Johnsons returned to Texas in 1969. Lady Bird wanted to do something
383 to encourage more people to plant wildflowers. She knew that little was known
396 about growing these flowers in gardens and that more research needed to be done.

410 In 1970 Lady Bird began a project to make the city of Austin more lovely. A
426 variety of colorful flowers and trees were planted along the banks of a major
440 river. Trails for hiking and biking were also added. This project helped inspire the
454 idea for building a center for studying native plants. In 1982 Lady Bird gave a
469 large sum of money and 60 acres of land near Austin to build the National
484 Wildflower Research Center. The purpose of the center was to learn about
496 wildflowers and share new information with interested people everywhere. In
504 1998 Lady Bird was honored for her tireless efforts to make our nation more
519 beautiful. The name of the center was changed to the Lady Bird Johnson
534 Wildflower Center.
536

Appendix C: TMSFA Retell Component

Main Idea Statement Rubric

<u>Score</u>

0: No response.

1: Weak comprehension

- Contains information that was not in the passage or that misinterprets information in the passage.
- Consists of an isolated fact or name.
- Is rambling or incoherent.
- No apparent understanding of the main idea.

2: Partial comprehension

- Contains some minor inaccuracies.
- Contains basic information from the passage, but not the most important point.
- Is not concisely stated and/or does not reflect the relationship among the ideas.
- Partial understanding of the main idea.

3: Strong comprehension

- Accurately reports information in the passage.
- Contains the most important point from the passage.
- Is coherently stated and reflects the proper relationship among the ideas.
- Response reflects a clear understanding of the main idea.

Main Idea Exemplars

Passage: Underground Town (Lexile 700)

Score 3
136 equated score
"It's about how underground homes are carved in the Earth because it's in a desert in Southern Australia. There are few windows in the homes. The temperatures get as high as 120 degrees in the summer. But it's only 32 degrees in winter."

Explanation:
The student's response accurately reports the most important information about the town in Australia and why the homes are built underground. He knows details about the geographic location, the homes, and the temperatures during the different seasons. Although the response is not as concisely stated as might be preferred, it does reflect the proper relationship among the ideas. It seems this student had a strong understanding of what he read.

Score 2
156 equated score
"The Australians. The way they made their homes and why they made their homes."

Explanation:
This student's response is accurate and contains more than an isolated fact. However, she has not made an attempt to provide the important information about the homes being underground or to relate this to the desert conditions. It is not clear in the last statement whether the student is referring to the temperatures and dust or to the opal mine. The response is, perhaps, too concise to determine whether the student had adequate understanding.

Score 1
149 equated score
"People living underground in the desert."

Explanation:
This student provides only an isolated fact from the passage. There is no indication that she understood the significance of people living underground or how the conditions necessitated certain building features.

Passage: Let's Do It Again (Lexile 840)

Score 3
124 equated score
"About learning how to water ski. It's his first time and he's really scared. His mom is encouraging him."

Explanation:
What this student says is very accurate and is provided in a logical order. He understands the significance of the water skiing and even why the mother played an important role. It is not clear from what was read whether the main character is a boy or a girl because only the younger sister's gender is identified. Therefore, it is acceptable for the student to refer to the character as either "him" or "her."

Score 2
122 equated score
"About this girl learning how to water ski; Nikki."

Explanation:
This student's response contains the point that the main character was learning to ski . However, he does not convey the important information about the character's feelings or experience during the attempt. Moreover, the student seems to have confused the younger sister in the story with the main character. The response lacks coherence.

Score 1
114 equated score
"I think it's about swimming or skiing."

Explanation:
This student does not offer any important information about the story. She only attempts to recall an isolated fact but is unsure of even that.

Passage: Spreading Wildflowers (Lexile 910)

Score 3
99 WCPM
"I think that one was about a girl named Claudia. She liked the wildflowers and she liked nature. And they opened a new program about the wildflowers and they left the seeds there so they would grow next year."

Explanation:
The student recalls a lot of information from the passage and does so accurately. She includes the most important points and connects them coherently. It is clear that she understood the passage and is even able to provide some details such as the given name of Lady Bird as well as the way the program ensured the wildflower seeds would be left to grow in subsequent years.

Score 2
97 equated score
"Lady Bird and her flowers. How they started protecting the wildflowers"

Explanation:
This response is very concise, but not very coherent. It is not clear if the student thinks Lady Bird owned the flowers and wanted to protect them, or if he understands that Lady Bird's love of flowers led to the wildflower program. The student does not attempt to show the relationships among those ideas.

Score 1
97 equated score
"She came to Austin to attend The University of Texas."

Explanation:
This student's response reflects a significant misunderstanding of the main idea. She focuses on the isolated fact of attending The University of Texas rather than on the love and protection of wildflowers mentioned in every paragraph.

References

ACT. (2008). *The forgotten middle: Ensuring that all students are on target for college and career readiness before high school*. Iowa City, IA: Author.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology, 112*(4), 545-557.

Akeike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317-322.

Applegate, M. D., Applegate, A. J., & Modla, V. B. (2009). "She's my best reader; she just can't comprehend": Studying the relationship between fluency and comprehension. *The Reading Teacher, 62*(6), 512-521.

Applegate, M. D., Quinn, K. B., & Applegate, A. J. (2008). *The Critical Reading Inventory: Assessing students' reading and thinking* (2nd ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Bader, L. A., & Pearce, D. L. (2009). *BADER Reading and Language Inventory*. Boston: Allyn & Bacon.

Balfanz, R., & Herzog, L. (March, 2005). *Keeping middle grades students on-track to graduation: Initial analysis and implications.* Paper presented at the second Regional Middle Grades Symposium, Philadelphia, PA.

Beaver, J. M. (2006). *Developmental Reading Assessment, Grades K-3* (2nd ed.). Parsippany, NJ: Pearson Education, Inc.

Beaver, J. M. (2003). *Developmental Reading Assessment, Grades 4-8* (2nd ed.). Parsippany, NJ: Pearson Education, Inc.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M., & Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika, 45*, 289-308.

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*(2), 137-164.

Blachman, B. A., Tangel, D. M., Ball, E. W., Black, R., & McGraw, C. K. (1999). Developing phonological awareness and word recognition sills: A two-year intervention with low-income, inner-city children. *Reading and Writing: An Interdisciplinary Journal, 11*, 239-273.

Bovaird, J. A. (2007). Multilevel structural equation models for contextual factors. In T. A. Little & J. A. Bovaird (Eds.), *Modeling contextual effects in longitudinal studies*. Mahwah, NJ: Lawrence Erlbaum Associates.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research: Methodology in the social sciences*. New York: Guilford.

Browne, M. W., & Cudeck, R. (1993). Alternate ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 136-162). Newbury Park, CA: Sage.

Buhi, E. R., Goodson, P., & Neilands, T. B. (2008). Out of sight, not out of mind: Strategies for handling missing data. *American Journal of Health Behavior, 32*(1), 83-92.

Buly, M. R., & Valencia, S. (2003). *Meeting the needs of ailing readers: Cautions and considerations for state policy*. Seattle: University of Washington, Center for the Study of Teaching and Policy.

Burke, M. D., & Hagan-Burke, S. (2007). Concurrent criterion-related validity of early literacy indicators for middle of first grade. *Assessment for effective Intervention, 32*(2), 66-77.

Burke, M. D., Hagan-Burke, S., Kwok, O., & Parker, R. (2009). Predictive validity of early literacy indicators from the middle of kindergarten to second grade*. The Journal of Special Education, 42*(4), 209-226.

Byrne, B. M. (1988). Testing the factorial validity and invariance of a measuring instrument using LISREL confirmatory factor analyses: A reexamination and application. *Multiple Linear Regression Viewpoints, 16*, 33-80.

Carlisle, J. F. (1999). Free recall as a test of reading comprehension for students with learning disabilities. *Learning Disability Quarterly, 22*(1), 11-22.

Carlo, M. S., August, D., McLaughlin, B., Snow, C. E., Dressler, C., Lippman, D. N., Lively, T. J., & White, C. E. (2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39*(2), 188-215.

Cassar, M., Treiman, R., Moats, L. C., Pollo, T. C., & Kessler, B. (2005). How do the spellings of children with dyslexia compare with those of nondyslexic children? *Reading and Writing: An Interdisciplinary Journal, 18*, 27-49.

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech, Language, and Hearing Research, 49*(2), 278–294.

Catts, H. W., Gillispie, M., Leonard, L. B., Kail, R. V., & Miller, C. A. (2002). The role of speed processing, rapid naming, and phonological awareness in reading achievement. *Journal of Learning Disabilities, 35*, 510-525.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cooter, R. B., Flynt, E. S., & Cooter, K. S. (2007). *Comprehensive Reading Inventory: Measuring reading development in regular and special education classrooms.* Upper Saddle River, NJ: Pearson Prentice Hall.

Coyne, M. K., Kame'enui, E. J., Simmons, D. C., & Harn, B. (2004). Beginning reading intervention as inoculation or insulin: First-grade reading performance of strong responders to kindergarten intervention. *Journal of Learning Disabilities, 37*, 90-105.

Cromley, J. G., & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading Comprehension. *Journal of Educational Psychology, 99*(2), 311-325.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contribution of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10*(3), 277-299.

Deno, S. L. (1986). Formative evaluation of individual student programs: A new role for school psychologists. *School Psychology Review, 15*, 358-374.

Doty, D. E., Popplewell, S. R., & Byers, G. O. (2001). Interactive CD-ROMS storybooks and young readers' reading comprehension. *Journal of Research on Computing in Education, 33*(4), 374-384.

Duffelmeyer, F. A., & Duffelmeyer, B. B. (1987). Main idea questions on informal reading inventories. *The Reading Teacher, 41*(2), 162-166.

Dynarski, M., Clarke, L., Cobb, B., Finn, J., Rumberger, R., Smink, J., Hallgren, K., & Gill, B. (2008). *IES Practice Guide: Dropout Prevention (NCEE 2008-4025).* National Center for Education Evaluation and Regional Assistance, Institute for Education Sciences, U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Ehri, L. C. (2000). Learning to read and learning to spell: Two sides of a coin. *Topics in Language Disorders, 20*, 19-49.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full ifomation maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*(3), 430-457.

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2007). *Learning disabilities: From identification to intervention*. New York: Guilford.

Fuchs, L. S., & Fuchs, D. (1984). Criterion-referenced assessment without measurement: How accurate for special education. *Remedial and Special Education, 5*, 29-32.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review, 21*(1), 45-58.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239-256.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20-29.

Gagne, E. D., Bing, S. B., & Bing, J. R. (1977). Combined effect of goal organization and test expectations on organization in free recall following learning from text. *Journal of Educational Psychology, 69*(4), 428-431.

Gambrell, L. B., & Jawitz, P. B. (1993). Mental imagery, text illustrations, and children's story comprehension and recall. *Reading Research Quarterly, 28*(3), 265-276.

Gambrell, L. B., Pfeiffer, W. R., & Wilson, R. M. (1985). The effects of retelling upon reading comprehension and recall of text information. Journal of Educational Research, 78(4), 216-220.

Gambrell, L. B., Koskinen, P. S., & Kapinus, B. A. (1991). Retelling and the reading comprehension o proficient and less-proficient readers. *Journal of Educational Research, 84*(6), 356-362.

Garson, D. G. (n.d.). Testing of assumptions. Retrieved November 12, 2009, from http://www2.chass.ncsu.edu/garson/pa765/assumpt.htm.

Good, R. H., & Kaminski, R. A. (2002a). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement. Available: http://dibels.uoreon.edu.

Good, R. H., & Kaminski, R. A. (2002b). *Vital Indicators of Progress (VIP)*. Dallas, TX: Voyager Expanded Learning, Inc.

Goodman, K. S. (2006). *The truth about DIBELS*. Portsmouth, NH: Heinamann.

Gough, P. B., Hoover, W. A., & Peterson, C. L. (1996). Some observations on a simple view of reading. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 1-13). Mahwah, NJ: Erlbaum.

Hansen, C. L. (1978). Story retelling used with average and learning disabled readers as a measure of reading comprehension. *Learning Disability Quarterly, 1*(3), 62-69.

Harcourt Assessment. (2007). *AIMSWEB maze curriculum-based measure.* San Antonio, TX: Author.

Hart, B. & Risley, T.R. (1995). *Meaningful differences in the everyday experiences of young American children.* Baltimore, MD: Brookes.

Hock, M. F., Brasseur, I. F., Deshler, D. D., Catts, H. W., Marquis, J. G., Mark, C. A., & Stribling, J. W. (2009). What is the reading component skill profile of adolescent struggling readers in urban schools? *Learning Disability Quarterly, 32*, 21-38.

Horowitz, R., & Samuels, S. J. (1985). Reading and listening to expository text. *Journal of Reading Behavior, 17*(3), 185-198.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structural analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Jackson, N. E. (2005). Are university students' component reading skills related to their text comprehension and academic achievement? *Learning and Individual Differences, 15*, 113-139.

Jenkins, J. R., & Jewell, M. (1993). Examining the validity of two measures for formative teaching: Reading aloud and maze. *Exceptional Children, 59*(421-432).

Johns, J. L. (2008). *Basic Reading Inventory: Pre-primer through grade twelve and early literacy assessments*. (10th ed.). Dubuque, IA: Kendall/Hunt Publishing Co.

Joreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 70*, 631-639.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*, 437-447.

Kamps, D. M., & Greenwood, C. R. (2005). Formulating secondary-level reading interventions. *Journal of Learning Disabilities, 38*(6), 500-509.

Karlsen, B., & Gardner, E. F. (1996). *Stanford Reading Diagnostic Test* (*SDRT*, 4th ed.). San Antonio, TX: Harcourt Brace & Co.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test* (2nd ed.). Minneapolis, MN: Pearson Assessment.

Kazdin, A. E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification, 1*, 427-451.

Keenan, J. M, Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*(3), 281-300.

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

Kintsch, W., & van Dijk, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review, 85*, 363-349.

Kirby, J. R., Parrila, R. K., & Pfeiffer, S. L. (2003). Naming speed and phonological awareness as predictors of reading development. *Journal of Educational Psychology, 95*(3), 453-464.

Klesius, J. P., & Homan, S. P. (1985). A validity and reliability update on the informal reading inventory with suggestions for improvement. *Journal of Learning Disabilities, 18*(2), 71-76.

Kouri, T., & Telander, K. (2008). Children's reading comprehension and narrative recall in sung and spoken story contexts. *Child Language Teaching and Therapy, 24*(3), 329-349.

Kucer, S. B. (2009). Examining the relationship between text processing and text comprehension in fourth-grade readers. *Reading Psychology, 30*(4), 340-358.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Leach, J. M., Scarborough, H. S., & Rescorla, L. (2003). Late-emerging reading disabilities. *Journal of Educational Psychology, 95*(2), 211-224.

Lee, J., Grigg, W., & Donahue, P. (2007). *The Nation's Report Card: Reading 2007 (NCES 2007-496).* National Center for Education Statistics, Institute of Education Science, U.S. Department of Education. Washington, DC: U.S. Government Printing Office.

Leslie, L., & Caldwell, J. (2006). *Qualitative Reading Inventory* (4th ed.). Boston: Pearson Education, Inc.

The Lexile Framework. (2007). *The lexile calculator*. Retrieved July 1, 2009, from www.lexile.com.

Loyd, B. H., & Steele, J. L. (1986). Assessment of reading comprehension: A comparison of constructs. *Reading Psychology: An International Quarterly, 7*, 1-10.

Lyon, G. R., Fletcher, J. M., Shaywitz, S. E., Shaywitz, B. A., Torgesen, J. K., Wood, F. B., Schulte, A., & Olson, R. (2001). Rethinking learning disabilities. In C. E. Finn, Jr., R. A. J. Rotherham, & C. R. Hokanson, Jr. (Eds.), *Rethinking special education for a new century* (pp. 259-287). Washington, DC: Thomas B. Fordham Foundation and Progressive Policy Institute.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Method, 1*, 130-149.

MacIntosh, R., & Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Applied Psychological Measurement, 27*, 372-379.

Madelaine, A., & Wheldall, K. (2005). Identifying low-progress readers: Comparing teacher judgment with a curriculum-based measurement procedure. *International Journal of Disability, Development, and Education, 52*(1), 33-42.

Marcotte, A. M., & Hintze, J. M. (2009). Incremental and predictive utility of formative assessment methods of reading comprehension. *Journal of School Psychology, 47*, 315-335.

Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement, 15*, 47-70.

Mason, L. H., Snyder, K. H., Sukhram, D. P., & Kedem, Y. (2006). TWA + PLANS strategies for expository reading and writing: Effects for nine fourth-grade students. *Exceptional Children, 73*(1), 69-89.

McGee, L. M. (1982). Awareness of text structure: Effects on children's recall of expository text. *Reading Research Quarterly, 17*, 581-590.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III technical manual*. Itasca, IL: Riverside.

Mehta, P., & Neale, M. (2005). People are variables too: Multilevel structural equation models. *Psychological Methods, 10*, 259-284.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-104). New York: Macmillan.

Moss, B. (1997). A qualitative assessment of first graders' retelling of expository text. *Reading Research and Instruction, 37*, 1-13.

Muthen, B. O., Kao, C., & Burstein, L. (1991). Instructional sensitivity in mathematics achievement test items: Applications of a new IRT-based detection technique. *Journal of Educational Measurement, 28*, 1-22.

Muthen, L., & Muthen, B. (2009). *Mplus version 5.21*. Los Angeles: Author.

Nilsson, N. L. (2008). A critical analysis of eight informal reading inventories. *The Reading Teacher, 61*(7), 526-536.

Pearman, C. J. (2008). Independent reading of CD-Rom storybooks: Measuring comprehension with oral retellings. *The Reading Teacher, 61*(8), 594-602.

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

Popplewell, S. R., & Doty, D. E. (2001). Classroom instruction and reading comprehension: A comparison of one basal reader approach and the four-blocks framework. *Reading Psychology, 22*, 83-94.

Psychological Corporation. (1992). *Wechsler Individual Achievement Test: Basic reading scale*. San Antonio, TX: The Psychological Corporation.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Retrieved December 22, 2009, from http://www.rand.org/pubs/research_briefs/RB8024/index1.html.

Rasinski, T. V. (1990). Investigating measures of reading fluency. *Educational Research Quarterly, 14*(3), 36-44.

Reed, D. K., & Vaughn, S. (manuscript under review). Retell as an indicator of reading comprehension. *Scientific Studies of Reading*.

Richgels, D. J., McGee, L. M., Lomax, R. G., & Sheard, C. (1987). Awareness of text structures: Effects on recall of expository text. *Reading Research Quarterly, 22*(2), 177-196.

Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42*(4), 546-567.

Roberts, G., Good, R., & Corcoran, S. (2005). Story retell: A fluency-based indicator of reading comprehension. *School Psychology Quarterly, 20*(3), 304-317.

Roe, B. D., & Burns, P. C. (2007). *Informal Reading Inventory*. Boston: Houghton Mifflin Co.

Savage, R. (2006). Reading comprehension is not always the product of nonsense word decoding and linguistic comprehension: Evidence from teenagers who are extremely poor readers. *Scientific Studies of Reading, 10*(2), 143-164.

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics, Vol. 1: Disorders of first-language development* (pp. 142-175). New York: Cambridge University Press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Schatschneider, C., Buck, J., Wagner, R., Hassler, L., Hecht, S., & Powell-Smith, K. (2004). A multivariate study of individual differences in performance on the reading portion of the Florida Comprehensive Assessment Test: A brief report. (Technical Report No. 5, Florida Center for Reading Research), Tallahassee, FL.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210-222.

Shinn, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education, 34*(3), 164-172.

Shinn, M. R., Good, R., Knutson, N., Tilly, W. D., & Collins, V. L. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*(3), 459-479.

Shinn, M. R., & Shinn, M. M. (2002). *AIMSweb training workbook: Administration and scoring of reading maze for use in general outcome measurement*. Eden Prairie, MN: Edformation, Inc.

Silvaroli, N. J., & Wheelock, W. H. (2004). *Classroom Reading Inventory* (10th ed.). Boston: McGraw Hill.

Simmons, D. C., Kame'enui, E. J., Harn, B., Coyne, M. D., Stoolmiller, M., Santoro, L. E., Smith, S. B., Beck, C. T., & Kaufman, N. K. (2007). Attributes of effective and efficient kindergarten reading intervention: An examination of instructional time and design specificity. *Journal of Learning Disabilities, 40*(4), 331-347.

Smith, J.R., Brooks-Gunn, J., and Klebanov, P.K. (1997). "The consequences of living in poverty for young children's cognitive and verbal ability and early school achievement." In G. J. Duncan & J. Brooks-Gunn (Eds.), *Consequences of* *growing up poor* (pp. 132-189). New York: Russell Sage Foundation Press.

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

Snyder, L., Caccamise, D., & Wise, B. (2005). The assessment of reading comprehension: Considerations and cautions. *Topics in Language Disorders, 25*(1), 33-50.

Spearritt, D. (1972). Identification of subskills of reading competence by maximum likelihood factor analysis. *Reading Research Quarterly, 8*(1), 92-111.

Spear-Swerling, L. (2006). Children's reading comprehension and oral reading fluency in easy text. *Reading & Writing: An Interdisciplinary Journal, 19*, 199-220.

Spooner, A. L. R., Baddeley, A. D., & Gathercole, S. E. (2004). Can reading accuracy and comprehension be separated in the Neale Analysis or Reading Ability? *British Journal of Educational Psychology, 74*, 187-204.

SPSS Inc. (2009). *SPSS PASW Statistics version 17.0*. Chicago, IL: SPSS, Inc.

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*(3), 407-419.

Stecker, P.M., & Fuchs, L.S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disability Research and Practice, 15*, 128-134.

Steiger, J. H., & Lind, J. M. (1980). *Statistically based tests for the number of common factors*. Paper presented at the meeting of the Psychometric Society, Iowa City, IA.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NJ: Erlbaum.

Talbott, E., Lloyd, J. W., & Tankersley, M. (1994). Effects of reading comprehension interventions for students with learning disabilities. *Learning Disability Quarterly, 17*, 223-232.

Texas Education Agency. (2004). *Texas Assessment of Knowledge and Skills (TAKS) information booklet: Reading, grade 6, revised*. Austin, TX: Author. Retrieved July 1, 2009, from http://www.tea.state.tx.us/index3.aspx?id=3693&menu_id3=793.

Texas Education Agency, & Pearson Educational Measurement. (2007). *Texas Assessment of Knowledge and Skills (TAKS): Technical Digest 2006-2007*. Austin, TX: Author.

Texas Education Agency, University of Houston, & The University of Texas System. (2008a). *Texas middle school fluency assessment.* Austin, TX: Author.

Texas Education Agency, University of Houston, & The University of Texas System. (2008b). *Texas middle school fluency assessment teacher's guide.* Austin, TX: Author.

Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications. Washington, DC: American Psychological Association.

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34,* 33-58, 78.

Torgesen, J., Wagner, R., & Rashotte, C. (1999). *Test of Word Reading Efficiency (TOWRE).* Austin, TX: Pro-Ed.

Trochim, W., & Donnelly, J. P. (2006). *Research methods knowledge base* (4th ed.). Mason, OH: Atomic Dog Publishing.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

U.S. Department of Education, National Assessment Governing Board. (2006). *Reading framework for the 2007 National Assessment of Educational Progress.* Washington, DC: Author. Retrieved June 20, 2009, from http://nagb.org/publications/frameworks/reading_07.pdf.

van den Broek, P., Tzeng, Y., Risden, K., Trabasso, T., & Basche, P. (2001). Inferential questioning: Effects on comprehension of narrative texts as a function of grade and timing. *Journal of Educational Psychology, 93*(3), 521-592.

Valencia, S. W., & Buly, M. R. (2004). Behind test scores: What struggling readers really need. *Reading Teacher, 57*(6), 520-531.

Velluntino, F., Scanlon, D., Sipay, E., Small, S., Pratt, A., Chen, R., Y.Y., & Denckla, M. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology*, *88*(4), 601-638.

Vukovic, R. K., Wilson, A. M., & Nash, K. K. (2004). Naming speed deficits in adults with reading disabilities: A test of the double-deficit hypothesis. *Journal of Learning Disabilities, 37*(5), 440-450.

Wagner, R. (in press). *Test of sentence reading efficiency.* Austin, TX: Pro-Ed.

Wagner, R. K., Muse, A. E., & Tannenbaum, K. R. (2007). Promising avenues for better understanding implications of vocabulary development for reading comprehension. In R. K. Wagner, A. E. Muse, & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 276-291). New York: The Guilford Press.

Wexler, J., Vaughn, S., Edmonds, M., & Reutebuch, C. K. (2008). A synthesis of fluency interventions for secondary struggling readers. *Reading and Writing: An Interdisciplinary Journal, 21*(4), 317–347

Wiley, H. I., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education, 26*(4), 207-214.

Williams, F. (1968). *Reasoning with statistics*. New York: Holt, Rinehart, and Winston.

Williams, K. T. (2001). *The Group Reading Assessment Diagnostic Evaluation (GRADE)*. Teacher's scoring and interpretive manual. Circle Pines, MN: American Guidance Service, Inc.

Williamson, G. L. (April, 2006). *Student readiness for postsecondary endeavors*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), San Francisco, CA.

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11*, 203-214.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of achievement*. Itasca, IL: Riverside Publishing.

Woods, M. L., & Moe, A. J. (2007). *Analytical Reading Inventory: Comprehensive standards-based assessment for all students, including gifted and remedial* (8th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Wright, H. H., & Newhoff, M. (2001). Narration abilities of children with language-learning disabilities in response to oral and written stimuli. *American Journal of Speech-Language Pathology, 10*(3), 308-319.

Zeno, S. (Ed.). (1995). *The educator's word frequency guide*. Brewster, NJ: Touchstone Applied Science Associates.

Zinar, S. (1990). Fifth-graders' recall of propositional content and causal relationships from expository prose. *Journal of Reading Behavior, 22*(2), 181-199.

VITA

Deborah Kay Reed was born in Nebraska. She completed her Bachelor of Arts in government at Claremont McKenna College (Claremont, CA) in 1993, and her initial teaching certificate at Chapman University (Orange, CA) in 1995. After relocating to Texas, she completed her Master of Arts in Curriculum and Instruction with a Reading Specialist certificate at The University of Texas at San Antonio in 1999. The author spent 10 years as a school-based practitioner before entering the field of educational research. In 2006, she enrolled in the Special Education doctoral program at The University of Texas at Austin. For the past seven years, she has participated in research and technical assistance efforts focused on 3-tier approaches to reading instruction and intervention for adolescents. Her publications have appeared in *Learning Disability Research and Practice*, *Middle School Journal*, *Research in Middle Level Education, Journal of Adolescent and Adult Literacy*, *Journal of Staff Development,* and the recently released book *The Promise of Response to Intervention*.

Permanent Address: 3422 Silver Spur Dr., San Angelo, TX 76904

This manuscript was typed by the author.