12-16-2019

The Structural Relation between Oral Language and Reading Comprehension: A Secondary Data Analysis within A Latent Variable Framework

Congying Sun

THE STRUCTURAL RELATION BETWEEN ORAL LANGUAGE AND READING COMPREHENSION: A SECONDARY DATA ANALYSIS WITHIN A LATENT VARIABLE FRAMEWORK

by

CONGYING SUN

Under the Direction of Lee Branum-Martin, PhD

ABSTRACT

Oral language and reading comprehension are typically considered as two separate constructs in most studies, however, there is no strong evidence for this separation using a measurement model. Results from cognitive psychology suggest that reading or text-based skills represent facets of a more general, overall ability of language proficiency. A general language proficiency may make measures of oral language and reading comprehension appear less distinct than may be typically assumed. Such cognitive overlap can be empirically tested using confirmatory factor models. Using secondary data analyses, this study examined the extent to which oral language and reading comprehension measures represent two distinct constructs by reanalyzing 44 summary data sets reported in 25 published journal articles and three dissertations, representing a total of 12,367 participants. First, we fit and compared the unidimensional and two-dimensional models. The results show that the one-dimensional model fit well in 11 out of 44 data sets and the two-dimensional model fit better than the one-dimensional model in 33 data sets, however, the discriminant validity between the two latent

constructs was relatively low across most data sets. These results suggest that psychometrically, it is difficult to separate oral language from reading comprehension. Second, we fit a bi-factor model for each data set with a general language factor and a specific factor of oral language or reading comprehension. The results show a strong general language factor and much weaker specific factors among school-age students, implying that the language structure might be better represented as a bi-factor model among these students. However, the general language factor was weak in most adult samples. In addition, these results suggest that the relation between oral language and reading comprehension was weaker among participants with low reading ability, and weaker in English second language learners, especially for adults.

INDEX WORDS: English, Vocabulary, Syntax, Morphology, Listening Comprehension, Reading comprehension

THE STRUCTURAL RELATION BETWEEN ORAL LANGUAGE AND READING COMPREHENSION: A SECONDARY DATA ANALYSIS WITHIN A LATENT VARIABLE FRAMEWORK

by

CONGYING SUN

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2019

THE STRUCTURAL RELATION BETWEEN ORAL LANGUAGE AND READING COMPREHENSION: A SECONDARY DATA ANALYSIS WITHIN A LATENT VARIABLE FRAMEWORK

by

CONGYING SUN

Committee Chair: Lee Branum-Martin

Committee: Robin Morris

Daphne Greenberg

Elizabeth Tighe

Audrey Leroux

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2019

ACKNOWLEDGEMENTS

I would like to acknowledge several important people who provide support and great contribution to the current project. First and foremost, I thank Dr. Lee Branum-Martin for being an amazing advisor during my study in the last five years at Georgia State University. His expertise, great support, encouragement and wonderful feedback made my life easier and made this project more complete and valuable. I also thank all the other committee members, Dr. Robin Morris, Dr. Daphne Greenberg, Dr. Elizabeth Tighe, and Dr. Audrey Leroux, for their time and effort to review this project. Their insightful comments made this project more reasonable and valuable. Last but not least, I would like to thank Anita and Eleanor for reviewing this project and thank Dr. Amy Lederberg for offering me the fellowship and the flexible work schedule to complete this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Language is an essential skill for learning and development but it is unclear whether it is a single skill that changes over time, or develops more complexity—into multiple skills—as people develop. The language abilities for understanding oral and written texts are minimally differentiated among young children (Carroll, 1993), and might become differentiated but still highly related for developing readers (Foorman, Petscher, & Herrera, 2018; Kintsch, 1988). Thus, it is possible that oral and written language measures could represent a single factor of a general language ability (Carroll, 1993). Most empirical studies, however, distinguish oral language skills as the predictors of reading comprehension—forcing a division which has become a tradition, following the simple view of reading (SVR; Gough, Hoover, & Peterson, 1996; Gough & Tunmer, 1986). The results from these analyses are questionable on both theoretical as well as statistical grounds, and may be less informative for understanding the nature of oral and written language development, if oral language skills and reading comprehension are the indictors of a single latent construct or if they share a great deal of variance that is typically ignored. Therefore, the main purpose of the current study is to empirically examine the extent to which various oral language and reading comprehension measures represent two distinct latent constructs.

In addition, the content of what is assessed by reading comprehension measures changes with children's age and reading ability (Hua & Keenan, 2017; Keenan, Betjemann, & Olson, 2008). That is, for younger children or children with lower reading ability, the content of reading tests mostly reflects decoding skills, whereas for older children or children with high reading ability, the content of reading test items depends more on oral language skills (Hoover & Gough, 1990; Lonigan & Burgess, 2017; Keenan, Betjemann, & Olson, 2008; Kershaw &

Schatschneider, 2012). This shift in content of reading comprehension measures away from decoding toward higher order language processing might influence the relation between oral language and reading comprehension. Moreover, struggling adult readers might not have made this shift from reliance on decoding to oral language (Mellard, Fall, & Woods, 2010). The relation between oral language and reading comprehension is lower for struggling adult readers who are native or nonnative English speakers (Fritz, 2015; Nanda, Greenberg, & Morris, 2010). Therefore, the current study will also examine how the relation between oral language and reading comprehension differs across samples of different ages, different reading abilities, and whether or not they are native English speakers.

1.1 Theoretical Frameworks of Oral Language and Reading Comprehension

Based on a factor analytic study of over 460 data sets, Carroll (1993) revealed a three-stratum factor structure (akin to a bi-factor model) with a general factor (Stratum III), *g*, among 69 subtests (Stratum I) of mental ability test batteries, and eight specific factors (Stratum II) that are independent of *g* among the corresponding subtests aimed to measure the specific factor (Figure 1).

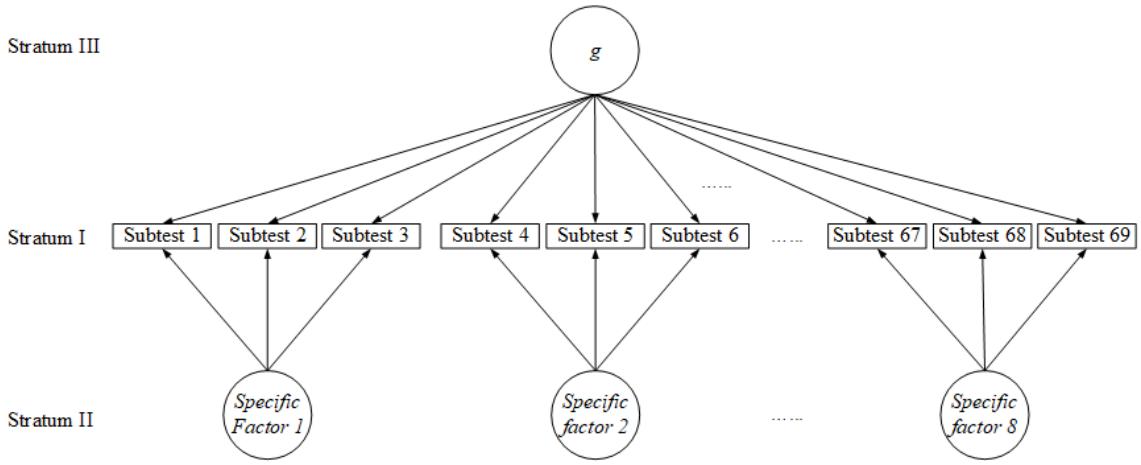

Figure 1 Diagram of the Three-Stratum Factor Structure, after Carroll (1993)

The eight specific factors are Fluid Intelligence, Crystallized Intelligence, General Memory and Learning, Broad Visual Perception, Broad Auditory Perception, Broad Retrieval Ability, Broad Cognitive Speediness, and Processing Speed. The listening and reading comprehension, lexical knowledge, grammatical sensitivity, decoding, spelling, reading fluency and writing are all Stratum I factors underlying the Crystallized Intelligence factor. This factor structure has been predominant in the research of cognitive abilities for several decades (Byle & Cucina, 2014; Cucina & Howardson, 2017). In terms of language ability, the three-stratum factor structure suggests that there might be a general language factor among all oral and written language measures. Moreover, Carroll (1993) also incorporated a developmental perspective in the language structure by proposing a kind of inverted umbrella or cone structure (Figure 2).

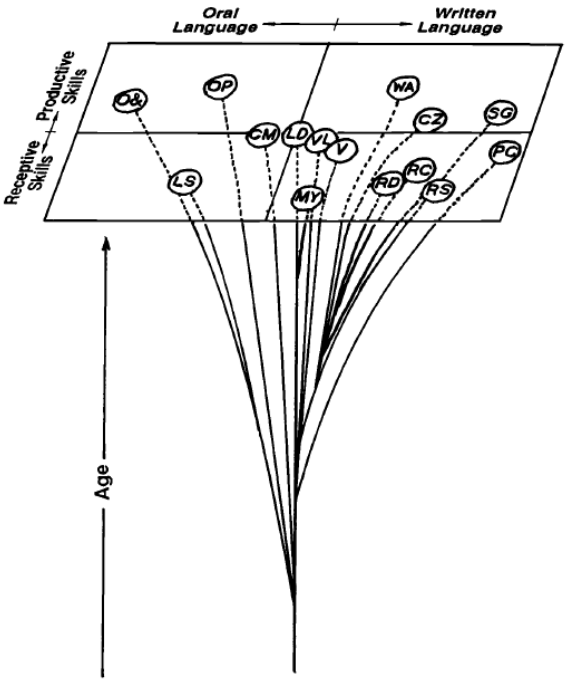

Figure 2 Conceptual Representation of Factors in the Language Domain (Carroll, 1993, p.147)

Specifically, language abilities are minimal, and minimally differentiated (like a single, vertical pole), in the earliest years of development. During these years, individuals learn to speak and understand the spoken form of the language. A normal child usually develops this

competence by the age of about five years. Then, as the child gets older, more and more specialized abilities become differentiated, represented by the various "spokes" that depart from the central pole or core of language ability. Individuals tend to become differentiated in levels of these specialized abilities beyond the age of five or so. By the time of adulthood, the individual differences in various specialized language abilities can become more pronounced and substantially less related to each other. According to this structure, multiple measures of oral language and reading comprehension might represent a single construct for young children but become differentiated into multiple constructs as children get older.

While the three-stratum factor structure is popular in cognitive research, it is not widely recognized in reading and language research, where the SVR is the predominant framework. The SVR divides reading into two parts: decoding, which is unique to reading, versus comprehension, which is shared between reading and listening. Once a text has been decoded, reading and listening appear to require essentially the same processes (Gough, Hoover, & Peterson, 1996; Gough & Tunmer, 1986). Thus, the relation between reading and listening comprehension is restricted by the child's decoding skills. If a child has high decoding skill that can support recognizing all the words in passages, the reading comprehension performance reflects the child's listening comprehension. Alternatively, if a child has low decoding skill, reading comprehension performance is more dependent on how many words that the child can recognize. Following the SVR, reading comprehension can be distinct from listening comprehension before an individual fully develops decoding skill.

Although reading comprehension for developing readers is constrained by word reading (Perfetti, Landi, & Oakhill, 2005), the construction-integration model of text comprehension (Kintsch, 1988) argues that the processes of text comprehension are largely the same for oral and

written texts (Graesser, Singer, & Trabasso, 1994). Text comprehension ultimately requires constructing a coherent mental representation of meaning as it is actually expressed by the text (oral or written text; Kintsch, 1988; Kintsch & Rawson, 2007; Perfetti & Stafura, 2014). According to the construction-integration model, multiple processes are involved in establishing a coherent and integrated mental representation of the ideas from text. The lowest level is established by parsing sentences and phrases and holding them briefly in memory. Then, initial propositions are constructed through semantic analysis, called "textbase representation" (Kim, 2016; *p*. 103). Finally, the mental representation of meaning is established by integrating initial propositions across the texts and with background knowledge for deeper understanding of the text. The component skills necessary at different levels for successful listening comprehension converge with those for reading comprehension, such as working memory, vocabulary, grammatical knowledge, inference and comprehension monitoring (Cain, Oakhill, & Bryant, 2004; Cromley & Azevedo, 2007; Kim, 2017).

To conclude, theoretical frameworks differ in the way oral language and reading comprehension relate to each other. Carroll (1993) suggests that oral language and reading comprehension measures might represent a general language ability, especially among young children, and as children get older, oral language and reading comprehension might be less associated with each other. SVR and the construction-integration model of text comprehension both assume that listening and reading comprehension share the same cognitive processes, but the SVR emphasizes that the relation between listening and reading comprehension depends on the child's decoding skill. That is, for young children, reading comprehension is more restricted by a child's decoding skill, and less associated with listening comprehension. For older children who have mastered decoding skills, reading comprehension mostly relies on language skills,

largely represented by listening comprehension (Adlof, Catts, & Little, 2006; Catts, Hogan, & Adlof, 2005; Hoover & Gough, 1990; Kendeou, Broek, White, & Lynch, 2009; Quinn, 2016; Tilstra, McMaster, Van den Broek, Kendeou, & Rapp, 2009). Other lower-level language skills, such as vocabulary, syntax, and grammatical knowledge, are viewed as necessary components for successful listening and reading comprehension, but are not included in the SVR.

1.2 Dimensionality of Oral Language and Relation to Reading Comprehension

Oral language is the ability to express knowledge, thought, and feelings using spoken words. It is a broad construct encompassing lexical- (i.e., vocabulary), sentence- (i.e., syntax, grammatical knowledge), and discourse-level (i.e., listening comprehension) skills. Several studies have explored the dimensionality of oral language with multiple measures of various oral language skills, such as listening comprehension, expressive and receptive vocabulary, syntax, grammar, morphology, and discourse skill, among a wide age of populations from preschool through high school (Foorman, Herrera, Petscher, Mitchell, & Truckenmiller, 2015; LARRC, 2015; Lonigan & Millburn, 2017; Tomblin & Zhang, 2006). The consistent finding from these studies is that the dimensionality of oral language changes with development. Specifically, oral language is a unidimensional construct for young children, but differentiates into multiple dimensions for older children, even though the shifting period is not same across different studies. For example, Tomblin & Zhang (2006) found that the best-fitting model was a one-factor model in kindergarten, second, and fourth grades, but a two-factor model (r = 0.78 between two latent factors) in eighth grade by testing expressive/receptive vocabulary and expressive/receptive grammar. The Language and Reading Research Consortium (LARRC, 2015) examined the structure of language tasks by testing vocabulary, grammar, and discourse-level language skills. Their results showed that vocabulary and grammar were best represented

by a single factor from prekindergarten to second grade, but separate factors in third grade (r = 0.90 between the latent vocabulary factor and the latent grammar factor), while discourse skills loaded on a distinct factor from first grade, despite the high correlations (r = 0.70 to 0.95) with the latent factors represented by vocabulary and grammar from prekindergarten through third grade. Foorman, Herrera, Petscher, Mitchell, and Truckenmiller (2015) also found that the best-fitting model varied depending on grade level, with one factor accounting for all language measures (vocabulary, syntax, and listening comprehension) in kindergarten, but a second-order model accounting for three lower-order factors (vocabulary, syntax, and listening comprehension) in first grade and a second-order model accounting for two lower-order factors (vocabulary, syntax/listening comprehension) in second grade. Lonigan & Milburn (2017) tested 19 to 20 measures of oral language on a sample of 1,895 children from preschool to fifth grade, finding a two-factor model with highly correlated (r = 0.90 to 0.94) vocabulary (expressive/receptive vocabulary, vocabulary depth) and syntax (expressive/receptive syntax/grammar, listening comprehension) factors across all grades.

Even though the multi-factor model was chosen as the best-fitting model for older children in the above studies, the correlations between the separate factors were very high (r = 0.70 to 0.94), suggesting large common variation among the separate factors. In other words, there might be a strong general oral language factor among various oral language measures. Thus, the multi-factor model found in the previous studies might be better represented as a bi-factor model with a general oral language factor and several specific factors (such as vocabulary, syntax). This hypothesis was supported by two recent studies (Foorman, Koon, Petcher, Mitchell, & Truckenmiller, 2015; Kieffer, Petscher, Proctor, & Silverman, 2016), which found that the bi-factor model was the best-fitting model for all children from third to tenth grade. The

general oral language factor was strong and reliable. It explained vast amounts of variance in reading comprehension across third to tenth grade. The specific factors of the language components were less strong and reliable, and they had smaller effects on reading comprehension beyond the general oral language factor. These results of the dimensionality of oral language are consistent with Carroll's (1993) finding that there is a strong general factor among all language measures with several specific factors for each language skill. Longitudinally, the language skills gradually become differentiated, but the results demonstrate the existence of the general oral language factor, at least, from prekindergarten through tenth grade—also supporting Carroll's hypothesis of divergence of abilities over time.

While single observed measures of language and reading comprehension measures have moderate to high correlations, the latent oral language factor is strongly correlated with reading comprehension. For example, Kershaw and Schatschneider (2012) found that the correlation between the oral language factor (listening comprehension and vocabulary) and the reading comprehension factor was 0.83 for third grade, 0.87 for seventh grade and 0.92 for tenth grade. Braze et al (2016) showed that the correlation was even higher (r = 0.96) among older participants aged from 16 to 25 years old. Foorman, Petscher, and Herrera (2018) showed that the oral language factor represented by listening comprehension, vocabulary, and syntax shared 92% to > 99% of variation with reading comprehension from fourth to tenth grades, while the shared variance between the oral language factor and reading comprehension was only moderate to high (42% - 69%) for first to third grades. Using multilevel confirmatory factor analysis, Mehta, Foorman, Branum-Martin, and Taylor (2005) found that the measures of phonological awareness, word recognition, spelling, reading comprehension, and writing skills represented a unitary literacy factor, which had a perfect correlation with a language factor represented by

vocabulary and general language skills at the classroom level, and a correlation of 0.7 at the student level. These studies provided valuable information to evaluate the relation between oral language and reading comprehension, but none of them directly tested whether a general language factor could encompass both oral language and reading comprehension. Moreover, contrary to Carroll's (1993) developmental hypothesis of cognitive abilities diverging at later ages, previous studies have shown that the relationship between oral language and reading comprehension increases as children get older.

1.3 Reading Comprehension Assessment

Reading comprehension is the ability to process text, understand its meaning, and to integrate with what the reader already knows. It is usually tested using standardized or unstandardized reading comprehension measures that require children to read a passage and answer several related questions. Most researchers claim that reading comprehension is a complex, multidimensional construct, but there is no direct evidence for such a claim from a measurement model (e.g., confirmatory factor analysis yielding multiple factors for reading). Studies using regression analyses revealed that different reading comprehension measures might tap a different array of cognitive processes (Betjemann, Keenan, Olson, & Defries, 2011; Cutting & Scarborough, 2006; Keenan, Betjemann, & Olso, 2008; Keenan & Meenan, 2014; Kendeou, Papadopoulos, & Spaneoudis, 2012). For example, Cutting and Scarborough (2006) compared three commonly used tests of reading comprehension, Wechsler Individual Achievement Test (WIAT), Gates-MacGinitie (G-M) and Gray Oral Reading Test (GORT), and found that the unique contribution of word decoding skill varied across all three tests, with nearly twice as much variance accounted for in WIAT scores than in G-M and GORT scores. The percentage of variance uniquely explained by oral language proficiency was similar for the WIAT and GORT

but substantially higher for the G-M. Keenan, Betjemann, and Olson (2008) also reported dramatic differences among four reading comprehension measures (Woodcock-Johnson Passage Comprehension (WJPC), Peabody Individual Achievement Test (PIAT), Gray Oral Reading Test (GORT), and Qualitative Reading Inventory (QRI)) in the degree to which performance is explained by word decoding versus listening comprehension and showed that the differences are not just a function of using a cloze-test format. The average correlation among these four reading comprehension measures was only 0.54. Word decoding accounted for far more variance than listening comprehension when the measure was either the WJPC or PIAT, but the reverse pattern was found for the other two measures. Although WJPC is a cloze test, the PIAT is not, suggesting that the extent to which individual differences in reading comprehension tests are largely accounted for by word reading is not so much a function of test formats as of passage length. Both WJPC and PIAT use sentence-length passages, while QRI and GORT use longer passages that might increase dependence on higher-level language skills. Kendeou, Papadopoulos, and Spaneoudis (2012) found that three reading comprehension tests (WJPC, Curriculum-based measure test, and a recall test) posed different processing demands on young readers (second graders). Specifically, the WJPC test exerts processing demands predominantly on orthographic processing and working memory. The CBM test exerts processing demands on fluency and vocabulary, whereas the recall test exerts processing demands on phonological processing, orthographic processing and working memory.

Using a twin design, Betjemann, Keenan, Olson, and Defries (2011) found that different reading comprehension tests used to measure the same construct may manifest very different patterns of genetic covariation. Specifically, WJPC and PIAT shared most genetic variance with decoding, and QRI, GORT, passage retelling and open-end comprehension questions shared

most with listening comprehension. Thus, the observed score based on a single reading comprehension measure cannot fully represent an individual's reading comprehension ability, and different reading comprehension measures may tap different facets of an individual's reading comprehension ability. However, using confirmatory factor analysis (CFA), some studies have demonstrated that different reading comprehension measures can be accounted for with a unitary construct (Foorman, Petscher, & Herrera, 2018; Francis, Fletcher, Catts, & Tomblin, 2005; Kershaw, & Schatschneider, 2012). Therefore, the result based on one single reading comprehension measure cannot necessarily generalize to other measures. A latent variable defined by multiple measures is required to better represent a reading comprehension construct which can generalize across tests (Francis, Kulesz, & Benoit, 2018).

The content of what is assessed by reading comprehension measures also changes with age and reading ability. Keenan, Betjemann, and Olson (2008) found that, for the same test, word decoding accounted for more variance in reading comprehension among younger children (or children with low-level reading skill) than older children (or children with high-level reading skill), while listening comprehension explained more variance in reading comprehension for older children (or children with high-level reading skill). This disparity was quite large for tests assessing comprehension with short texts (PIAT and WJPC), but less dramatic for tests with longer passages (GORT and QRI). Thus, detection of developmental differences is also influenced by test differences. Using quantile regression, Hua and Keenan (2017) showed that the contribution of word recognition and listening comprehension vary as a function of reading comprehension skill. For example, for the GORT, although there is considerable variance in word recognition skills among those who score at 10th quantile, it accounts for very little of their performance. Word recognition is more important in explaining poor comprehenders' scores than

typical or high performers' scores for one QRI test. Thus, the content of reading comprehension measures also influences the relation between oral language and reading comprehension. As children get older or achieve higher reading levels, the relation between oral language and reading comprehension would be stronger.

However, typically developing adults are understudied in previous studies. The adult samples in most studies have been struggling readers or nonnative English speakers. Mellard et al. (2010) found that among low-literacy adults, reading comprehension strongly relied on decoding skill with vocabulary and language comprehension skills contributing less to reading comprehension using a path analysis, which demonstrated that these low-literacy adults had not made the shift from reliance on decoding to oral language. Using confirmatory factor analysis, Fritz (2015) explored the reading construct among struggling readers by focusing on oral language, reading comprehension, decoding, and fluency. In the resulting models, the latent correlations between oral language and reading comprehension were lower among struggling readers (r = 0.67 for elementary students, r = 0.51 for middle school students, and r = 0.72 for adults), compared with the latent correlations among typically developing children (Foorman et al., 2018; Kershaw & Schatschneider, 2012). Nanda et al. (2010) also explored the reading construct among struggling adult readers by testing reading comprehension, vocabulary, decoding, reading fluency, and phonological awareness using confirmatory factor analysis, but found that the child-based theoretical models failed to fit to both native English-speaking adults struggling with reading and nonnative English-speaking adults struggling with reading. They speculated that the poor model fit was due to low correlations among the measures. These results suggest that the relation between oral language and reading comprehension is low for struggling

adult readers regardless of whether they are native English speakers or English learners from other languages.

In addition, two meta-analysis studies examined the correlations between reading comprehension and its components in two populations (struggling adult readers and second language learners). Tighe and Schatschneider (2016) found moderate correlation between oral language and reading comprehension for struggling adult readers. Specifically, the average correlation with reading comprehension was 0.55 for language comprehension, 0.52 for vocabulary, and 0.59 for morphological awareness. Jeon and Yamashita (2014) found that the average correlation between oral language and reading comprehension was higher for second language learners with most studies having English as the second language. The average correlation with reading comprehension was 0.77 for listening comprehension, 0.79 for vocabulary, 0.85 for morphosyntactic knowledge, and 0.61 for morphological awareness. Age and second language proficiency were not significant moderators, except on vocabulary with older participants (13 years old or older) having higher correlation between reading comprehension and vocabulary. Therefore, for younger or typically developing second language learners, the relation between oral language and reading comprehension might be higher, compared with struggling adult readers.

1.4 Purpose of the Current Study

To summarize, the three-stratum factor structure suggests that oral language and reading comprehension are both specific facets of a general language factor, especially for developing readers. As children get older, they might be more differentiated. The SVR assumes that the relation between listening comprehension and reading comprehension is lower among younger children or children with lower decoding skill, but higher among older children or children with

developed decoding skill. Nevertheless, the SVR does not have assumptions regarding how other oral language skills, such as vocabulary, morphological awareness and syntactic skill, relate to reading and listening comprehension. Previous studies have shown that measures of oral language skills and listening comprehension are all good indicators of a strong general oral language factor. Using confirmatory factor analysis, some studies have also found a high correlation between the oral language factor and the reading comprehension factor. However, the measurement question of whether oral language and reading comprehension are two distinct constructs has not received much attention. Therefore, the current study aims to examine whether multiple oral language and reading comprehension measures can represent two distinct and measurable constructs, and whether their relation differs across samples with different ages, different reading abilities and whether or not native English speakers, by reanalyzing the summary statistics reported in published or unpublished studies.

2 METHOD

2.1 Inclusion Criteria for Studies

In order to examine the structural relation between oral language and reading comprehension in a secondary data analysis, we needed studies that could provide sufficient summary statistics to test our hypothesized models. That is, the studies should include multiple cognitive measures for testing both oral language and reading comprehension skills and report at least the correlations among these measures (means and SD are not required for testing structural hypotheses). In the current study, we also limited the languages and samples in the studies, by focusing on English only and typically developing populations in Grade 3 or above. Specifically, each study was required to meet the following criteria:

(a) English with reported measures. The study was written in English, had full text available, and had measures tested in English, because the main purpose of the current study was to examine the relation between oral language and reading comprehension in English.

(b) Cognitive test data. The study was a primary study that included cognitive test data on reading and language performance. Meta-analyses, corrigenda, and studies of non-test data (e.g., eye movements or brain functions) were excluded.

(c) Typically developing. At least 90% of participants were not from special populations, such as individuals with intellectual disability or Down's syndrome, individuals with vision or hearing impairments, individuals with brain injuries, aphasia or other abnormalities, individuals with autism spectrum disorder, attention deficit/hyperactivity disorder, or other behavior problems.

(d) Reading age. The study contained a sample in grade 3 or above (or 9 years old and older). Grade 3 was selected because children move from learning to read to reading to learn beginning in Grade 3 (Chall, 1967). If the study had wide grade range, it would be included if the median grade was Grade 3 or above (e.g., Grade 2-4).

(e) Multiple measures for each construct. The study included at least two oral language measures and two reading comprehension measures. Multiple measures were required to represent each latent construct in order to control for the measure error. For testing oral language ability, we included the expressive and receptive vocabulary, morphology, syntax, and listening comprehension measures, including those that were in written format. For testing reading comprehension ability, we included sentence-level and passage-level reading comprehension measures. In addition, we excluded timed oral language and reading comprehension measures from our analysis, because the score of the timed measure was confounded with additional complicated ability (e.g., fluency, speed).

(f) Total number of measures. The study had at least five measures in total, because a bi-factor model required at least five measures.

(g) Summary statistics. The study reported Pearson correlations among all oral language and reading comprehension measures. Other non-parametric rank correlations (e.g., Tau correlations) were not included, because in reading research the rank correlations were usually reported when the variables were skewed.

2.2 Literature Search

In the current study, we searched the studies in two phases. The first phase was a standard search and the second phase was implemented with additional criteria to gain better

representation, especially among adult participants. In the first phase, we searched PsycINFO, ERIC, and ProQuest (Dissertations and Theses, PQTD) databases in July 2019, using the key words "reading" in Title and "reading comprehension" OR "text comprehension" OR "passage comprehension" OR "sentence comprehension" in Abstract and "language" OR "listening comprehension" OR "linguistic" OR "vocabulary" OR "synta*" OR "morpholog*" OR "discourse" in Abstract. The key words "fMRI" OR "ERP" OR "brain" OR "genetic" OR "eye-movement" OR "eye-tracking" OR "case study" were not allowed to show in the title, in order to exclude studies unlikely to include cognitive measures of language and reading. We limited the studies to only those that had been published between 2009 and 2019. ProQuest also contains unpublished reports, thesis and dissertations. The initial search yielded 3,434 studies with 1,344 studies in PsycINFO, 1,160 studies in ERIC, and 930 studies in ProQuest. After removing duplicates (1,373 studies), a total of 2,061 studies were identified.

Table 1 shows the results of the study selection process. According to inclusion criterion (a), we excluded 349 studies without full text available online, 289 studies testing in languages other than English, and 26 studies written in languages other than English. According to inclusion criterion (b), we excluded 223 studies that were meta-analyses, systematic reviews, case studies, or corrigenda, and 40 studies that collected data using functional magnetic resonance imaging (fMRI), electroencephalogram (EEG) or eye movements and did not have cognitive test data. According to inclusion criterion (c), 85 studies were excluded because their samples were restricted to special populations only, or more than 10% of the participants were special populations. According to inclusion criterion (d), 122 studies were excluded due to their younger sample of pre-kindergarten to Grade 2 or median grade below Grade 3. According to inclusion criterion (e), we excluded 205 studies that did not test oral language and reading

comprehension, 305 studies that only tested one ability (either oral language or reading comprehension), and 372 studies where only a single measure was used to test oral language or reading comprehension. According to inclusion criterion (f), we excluded eight studies that included less than five oral language and reading comprehension measures in total. According to inclusion criterion (g), we excluded 11 studies that did not report Pearson correlations among oral language and reading comprehension measures.

Table 1 Paper Selection Process and Criteria in First Phase of Search

Criteria	Paper Selection
(a)	349 studies were excluded because full texts were not available online
	289 studies with tests only in languages other than English were excluded
	26 studies written in languages other than English were excluded
(b)	223 studies were excluded including meta-analyses, systematic reviews, case studies, and corrigenda
	40 studies with only non-behavior data (e.g., fMRI, EEG, eye-movement) were excluded
(c)	85 studies including more than 10% of the participants as special populations were excluded
(d)	122 studies were excluded due to their younger sample of pre-kindergarten to Grade 2 or median grade below Grade 3
(e)	205 studies were excluded because they did not test oral language and reading comprehension
	305 studies were excluded because they only tested one ability (oral language or reading comprehension)
	372 studies were excluded because a single measure was used to test oral language or reading comprehension
(f)	Eight studies were excluded due to less than five measures in total
(g)	11 studies were excluded which did not report Pearson correlations
Resulting Sample	24 studies met the inclusion criteria (2 duplicates removed)

Finally, we found 26 studies meeting the inclusion criteria in the first phase of search. Among the 26 studies, we further removed two duplicate studies. Guo and Roehrig (2011) used the same data set with Guo (2009), which was a dissertation. Foorman et al. (2015) also analyzed the same data set with Foorman et al. (2018) with fourth to tenth graders. Foorman et al. (2015) reported the correlation table for the whole sample, while Foorman et al. (2018) reported the correlation table for each grade separately. Thus, we chose to analyze the summary data sets in

Foorman et al. (2018), in order to better examine the effect of age on the relation between oral language and reading comprehension. After these two duplicate studies were removed, there were 24 studies included in our analyses.

In the second phase of the search, we reviewed the references of a meta-analysis study (Quinn, 2016), which examined the relations between reading comprehension and other constructs in the years of 1990 to 2016. We found four studies that met our criteria from these additional references. Of these four studies, Tighe, Wagner, and Schatschneider (2015) used the same data set as Kershaw and Schatschneider (2012) that had been identified through the first phase of search. Tighe, Wagner, and Schatschneider (2015) reported the correlations between the total score of listening comprehension and other reading and language measures, while Kershaw and Schatschneider (2012) reported the correlations between each listening comprehension subtest and other measures. Thus, the summary data set in Kershaw and Schatschneider (2012) was reanalyzed in the current analyses. We also added one dissertation (Fritz, 2015) that explored reading constructs among struggling readers in elementary school, middle school, and adulthood by focusing on oral language, reading comprehension, decoding, and fluency. The data sets for middle school students and adults were included in the current analyses from this dissertation, but the data set for elementary students was not included, because these students were from first to third grade which failed to meet inclusion criterion (c).

Through the two phrases of search, 28 studies comprised of 25 journal articles and three dissertations were included in our analyses. These 28 studies represented a total of 12,367 participants and reported 44 correlation tables in total. The characteristics of these data sets are reported in Appendix A. Appendix B shows the measures for testing reading comprehension,

listening comprehension, vocabulary, morphological awareness and syntactic skill separately in each study. These 28 studies are also denoted with asterisks within the References.

Among the 44 summary data sets, 20 data sets were reported for elementary students, nine data sets for middle school students, three data sets for high school students, three data sets for children with wide age range (8-18 years old), and nine data sets for adults with wide age range (16 - 68 years old). Among 20 data sets for elementary students, only two samples were reported having low reading ability (Cho, Capin, Roberts, Roberts, & Vaughn, 2019). In addition, 15 samples were native English speakers. Only one sample was English learners, one sample was Chinese (Cantonese) and English bilinguals, one sample was Spanish and English bilinguals, one sample was a composite sample of 58.5% English Learners and 45.5% Spanish and English bilinguals, and one sample was a composite sample of 56% native English speakers and 44% Spanish and English bilinguals.

Among nine data sets for middle school students, only one sample was reported as struggling readers (Fritz, 2015), and one sample was native Chinese speakers learning English as a second language (Li & Kirby, 2015). Among three data sets for high school students, no sample was reported as having low reading ability or English learners. Among three data sets including children with a wide age range, one sample was native English speakers aged from 9 to 15 (Spencer, Richmond, & Cutting, 2019), one sample was native English speakers with a history of reading difficulty aged from 8 to 18 (Betjemann et al., 2011), and the third sample was a composite sample of typically developing children and children with low reading abilities aged from 9 to 15 (Cutting, Materek, Cole, Levine, & Mahone, 2009).

Among nine data sets for the adults, four samples were native English speakers with two samples having low reading abilities, four samples were native Chinese speakers learning

English as a second language, and one sample was native Spanish speakers learning English as a second language and had low reading ability.

2.3 Models

The current study used a secondary data analysis approach to examine the structural relation between oral language and reading comprehension. Confirmatory factor analysis (CFA) was used to fit the a priori factor models to the summary statistics from reported studies, using the sample size reported in each study (Bollen, 1989; Kline, 2015; MacCallum, Wegener, Uchino, & Fabrigar, 1993). The summary statistics included the correlations with means and standard deviations (if means and standard deviations were available). First, each model was evaluated in model fit indices. Second, a chi-square difference test was used to compare the one-factor model and the two-factor model for each data set. If the one-factor model fit better, multiple oral language and reading comprehension measures represented a unidimensional ability. If the two-factor model fit better, we further examined the discriminant validity between two latent factors. Third, a bi-factor model derived from Carroll (1993) represented a general language proficiency factor among all oral language and reading comprehension measures, along with a specific factor of oral language or reading comprehension. The strength (convergent validity) of the general factor and the specific factors were evaluated against each other. All models were fit using Mplus 7 (Muthén, & Muthén, 1998-2017).

In this study, we conducted two sets of analyses. In the first set of analyses, we fit and compared the one-factor model and the two-factor model. The one-factor model (Figure 3a) implies that all oral language and reading comprehension measures represent a unidimensional ability (i.e., general language). The two-factor model (Figure 3b) indicates that oral language and reading comprehension might be two distinct and measurable constructs. In addition, two

previous studies found that oral language was better represented by a bi-factor model with a general oral language factor and several specific factors (e.g., vocabulary, syntax; Foorman et al., 2015; Kieffer et al., 2016). Thus, we also fit another two-factor model with oral language represented as a bi-factor model (Figure 3c) for 13 data sets that tested multiple oral language skills (e.g., vocabulary and syntax) with at least five measures. The latent factor *Sf* in the two-factor model in Figure 3c could indicate a specific vocabulary factor, or a specific listening comprehension factor, or a specific syntax factor, or a specific morphology factor, depending on the measures used in each study. These specific factors may represent sub-traits or method effects (Maul, 2013), based on the type of tests used in the study.

According to Carroll (1993), the language structure is better represented as a bi-factor model with a strong general language factor and several specific factors. Thus, in the second set of analyses, we fit a bi-factor model to each data set (Figure 3d) where all oral language and reading comprehension measures represented a general language factor, and beyond the general factor, the specific factor of oral language or reading comprehension may also exist among their respective measures. The existence of the specific factor of oral language or of reading comprehension depended on the measures and samples in each data set. Thus, we used the dashed lines for the specific factors in the diagram of the bi-factor model in Figure 3d.

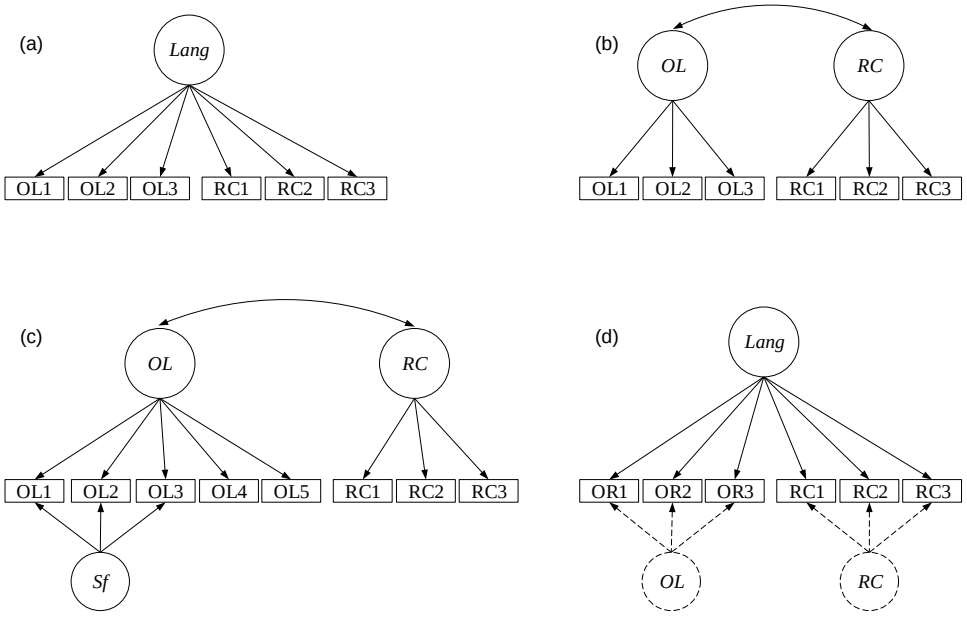

Figure 3 Diagrams of One-Factor Model, Two-Factor Model, Two-Factor Model with Oral Language Represented as a Bi-factor Model and Bi-Factor Model
Note. Lang=Language; OL=Oral Language; RC=Reading Comprehension; *Sf*=Specific factor. (a) one-factor model; (b) two-factor model; (c) two-factor model with oral language represented as a bi-factor model; (d) bi-factor model.

In addition, some measures were taken from a single test battery but used in different modalities (e.g., a reading test given orally). Four studies were found in which modality was adapted. To account for a possible method effect for these four studies, we added a residual correlation between the parallel listening comprehension and reading comprehension measures. Specifically, in Braze et al. (2016) and Van Dyke et al. (2014), half of the items of the Peabody Individual Achievement Test were presented orally to test listening comprehension and half of the items were in written format to test reading comprehension. In Cho et al. (2019), two scores (one for odd items, one for even items) were calculated from the Woodcock Johnson oral comprehension subtest and two scores (one for word knowledge, one for world knowledge) were calculated from the verbal knowledge subtest of the Kaufman Brief Intelligence Test. In Betjemann et al. (2011), the Qualitative Reading Inventory test was used to measure both

listening comprehension and reading comprehension. In each of these four studies, residual covariances were added to the respective models to account for the additional method-based effects of using the same measure in two modalities.

The chi-square (χ^2), comparative fit index (*CFI*), Tucker-Lewis Index (*TLI*), and root-mean-square error of approximation (*RMSEA*) were reported to evaluate model fit. The model was considered as not fitting if the *CFI* and *TLI* values were smaller than 0.90, and the *RMSEA* value was larger than 0.10 (Browne & Cudeck, 1993; Marsh, Hau, & Grayson, 2005). The chi-square difference test was used to compare the nested models. If the chi-square test was not statistically significant, the more parsimonious model fit equivalently with the more complex model, and the more parsimonious model was considered to be the best-fitting model. If the chi-square test was statically significant, the more complex model fit better and was then considered to be the best-fitting model.

In evaluating model implications, besides the model fit indices, reliability (convergent validity) and discriminant validity were also important. Convergent validity is the degree to which the multiple measures designed to measure the same construct are related. Discriminant validity is the degree to which two measures designed to measure conceptually different constructs are unrelated (Campbell & Fiske, 1959). Fornell and Larcker (1981) proposed a more stringent way to evaluate convergent and discriminant validity using the average variance extracted (AVE) and the squared factor correlation (SFC), respectively. AVE is the average of the amount variance across all measures that is explained by a factor by squaring the standardized factor loadings and then averaging them. The higher the AVE value, the stronger the latent factor, indicating good convergent validity of the measures. SFC is the shared variance between latent factors. The higher the SFC value, the closer relation between the latent factors. If

the AVE is less than the SFC, there is little support for discrimination between the latent factors (Fornell & Larcker, 1981; LARRC, 2015; Netemeyer, Bearden, & Sharma, 2003). In the one-factor model, all the indicators were loaded on a single latent factor. Thus, the SFC was fixed at 1, and the AVE was calculated for the oral language measures and the reading comprehension measures separately. For the bi-factor model, we calculated the AVE for the general language factor and the AVE for the specific factor (reading comprehension or oral language). If the AVE for the general factor was much higher than the AVE for the specific factor, it supported the three-stratum factor structure (Carroll, 1993). Otherwise, if the general language factor was weaker, it could be concluded that the two-dimensional model might better represent the language structure. These measures will be reported and evaluated.

3 RESULTS

3.1 One-factor and Two-factor Model Results for Elementary Students

Table 2 shows the model fit indices of the one-factor models for 20 data sets in the elementary students. Of the 20 one-factor models, nine fit well, nine had a lack of fit in *TLI* or *RMSEA* or both, and two had no indication of reasonable fit with *CFI* < 0.90, *TLI* < 0.90 and *RMSEA* > 0.1. Table 3 shows the model fit indices of the two-factor models for the 20 data sets. 12 two-factor models fit well with latent correlations ranging from 0.75 to 0.96 (see Table 3). Two models had a lack of fit in all three indices (*CFI*, *TLI*, and *RMSEA*). Three models only had a lack of fit in *RMSEA*. Among the remaining three models, two models had latent correlations greater than one, as highlighted using the superscript (^{e}) in Table 3, indicating that the unidimensional model is sufficient to represent the language structure in these two samples (Leider et al., 2013; Proctor et al., 2012). One failed to converge in the non-English learners in Cho et al. (2019), which included children who performed below a standard score of 85 on the Gates-MacGinitie reading comprehension test. The correlation results in Cho et al. (2019) showed very low correlations between this test and the oral language measures (r = -0.14 - 0.02) and between this test and the passage comprehension subtest of the Woodcock Johnson Tests of Achievement test (r = 0.18). Even though the two-factor model fit well in the English learners in Cho et al. (2019), a similar correlation pattern was also shown in the English Learners.

Comparing the one-factor model to the two-factor model for each data set, the chi-square test (see Table 2) suggested that the one-factor fit was equivalent with the two-factor model in two samples of native English Speakers (Foorman et al., 2017; Grade 3 and Grade 5), and two composite samples with one including Spanish and English bilinguals and English Learners (Leider et al., 2013), and the other including Spanish and English bilinguals and native English

speakers (Proctor et al., 2012). The two-factor model fit better in 15 data sets with latent correlations ranging from 0.75 to 0.94. In addition, four data sets tested multiple oral language skills using at least five measures. We also fit a two-factor model with oral language represented as a bi-factor model for these four data sets. The model fit indices and the chi-square tests (see Table 3) show that the two-factor model with oral language represented as a bi-factor model fit better than the two-factor model without oral language represented as a bi-factor model for each data set, except that one two-factor model with oral language represented as a bi-factor model had negative residual problem in Foorman et al. (2018; Grade 4). Thus, the two-factor model with oral language represented as a bi-factor model was considered the best-fitting model in Kershaw and Schatschneider (2012; Grade 3), Foorman et al. (2018; Grade 5), and Kieffer et al. (2016; Grade 3-5). All the best-fitting models are highlighted with the superscript ([a]) in Table 2 and 3.

Based on the above best-fitting models, we calculated the SFC, the AVE for the oral language factor and the AVE for the reading comprehension factor for each data set. Figure 4 shows the scatter plot of the SFCs and the AVEs for the oral language factor, and Figure 5 shows the scatter plot of the SFCs and the AVEs for the reading comprehension factor. The filled circle indicates that the sample was limited to one grade, and the empty circle indicates that the sample included participants across several grades. As shown in Figure 4, the AVEs for the oral language factor were lower than the SFCs in all the data sets, except for Kieffer et al. (2016; Grade 3-5) where AVEs for the oral language factor were equivalent with the SFCs. Figure 5 shows that the AVEs for the reading comprehension factor were lower than the SFCs in 13 out of 20 data sets, but higher than the SFCs for the remaining seven data sets. There was an outlying study in Figure 5 where the SFC was one and the AVE for the reading comprehension factor was

0.06 (Cho et al., 2019). This discrepant estimate indicates the correlations were low among oral language and reading comprehension measures for the non-English learners in Cho et al. (2019).

The results shown in Figures 4 and 5 suggest that the discrimination between the two latent factors was low in 13 data sets when the AVEs for two latent factors were both lower than the SFCs. In six data sets, the AVEs for the oral language factor were lower than the SFCs, but the AVEs for the reading comprehension factor were not lower than the SFCs. It was interpretable that the AVEs for the reading comprehension factor were higher, because most reading comprehension measures used the same test paradigm asking participants to read a passage and then answer multiple-choice questions, which lead to higher correlations within these measures. Only in Kieffer et al. (2016; Grade 3-5) where 74% of the sample was native English speakers and 26% was English learners, the AVEs for two latent factors were not lower than the SFCs, implying that the two-dimensional model might be optimal to represent the language structure for this composite sample.

Table 2 Fit Statistics of One-factor Models for 20 Data Sets in Elementary Students

Sample	Grade	χ^2	*df*	*CFI*	*TLI*	*RMSEA*	*p*(diff)
Chiu (2018)	3	78.49	9	0.91	0.86	0.16	< 0.01
Foorman et al. (2017; Grade 3) [a]	3	14.26	5	0.99	0.99	0.06	0.15
Foorman et al. (2018; Grade 3)	3	153.27	9	0.91	0.85	0.18	< 0.01
Kershaw & Schatschneider (2012; Grade 3)	3	98.93	14	0.90	0.85	0.17	< 0.01
Kim & Wagner (2015; Grade 3)	3	43.95	5	0.94	0.88	0.17	< 0.01
Siu & Ho (2015)	3	23.08	5	0.94	0.88	0.13	< 0.01
Tannenbaum et al. (2006)	3	67.19	9	0.91	0.85	0.18	< 0.01
Tannenbaum (2009; Grade 3)	3	61.85	9	0.95	0.92	0.15	< 0.01
Cho et al. (2019; English Learners)	4	14.21	7	0.99	0.97	0.07	0.03
Cho et al. (2019; Non-English Learners) [a]	4	15.68	7	0.96	0.92	0.08	—
Foorman et al. (2017; Grade 4)	4	7.79	5	1.00	1.00	0.03	0.03
Foorman et al. (2018; Grade 4)	4	198.67	14	0.83	0.75	0.22	< 0.01
Kim & Wagner (2015; Grade 4)	4	10.34	5	0.99	0.98	0.07	0.01
Harlaar et al. (2010)	4 [b]	25.86	9	0.99	0.98	0.07	< 0.01
Foorman et al. (2017; Grade 5) [a]	5	10.91	5	1.00	0.99	0.04	0.14
Foorman et al. (2018; Grade 5)	5	196.40	14	0.88	0.82	0.20	< 0.01
Proctor et al. (2012) [a]	2-4	9.45	9	1.00	1.00	0.01	0.15
Kieffer et al. (2016; Grade 3-5)	3-5	97.36	14	0.93	0.90	0.14	< 0.01
Leider et al. (2013) [a]	3-5	14.16	9	0.99	0.98	0.07	0.49
Lesaux et al. (2010)	4-5	10.94	5	0.97	0.93	0.12	< 0.01

Note. *p*(diff) is the *p* value for the chi-square test of model fit for each data set, compared to the respective two-factor model in Table 3; [a] indicates the best-fitting model for that data set; [b] indicates the sample's mean age was reported in the study, and converted to grade by subtracting the mean age from six; Dash indicates that the chi-square test was not available.

Table 3 Fit Statistics of Two-factor Model and Two-Factor Model with Oral Language Represented as a Bi-factor Model for 20 Data Sets in Elementary Students

Model, Sample	Grade	χ^2	*df*	*CFI*	*TLI*	*RMSEA*	*r(RC,OL)*	*p*(diff)
Two-factor Model								
Chiu (2018) [a]	3	27.30	8	0.98	0.96	0.09	0.80	
Foorman et al. (2017; Grade 3)	3	12.22	4	1.00	0.99	0.06	0.93	
Foorman et al. (2018; Grade 3) [a]	3	70.97	8	0.96	0.93	0.13	0.86	
Kershaw & Schatschneider (2012; Grade 3)	3	34.96	13	0.97	0.96	0.09	0.82	
Kim & Wagner (2015; Grade 3) [a]	3	22.93	4	0.97	0.93	0.14	0.82	
Siu & Ho (2015) [a]	3	11.57	4	0.98	0.94	0.10	0.81	
Tannenbaum et al. (2006) [a]	3	13.88	8	0.99	0.98	0.06	0.81	
Tannenbaum (2009; Grade 3) [a]	3	17.47	8	0.99	0.98	0.07	0.88	
Cho et al. (2019; English Learners) [a]	4	9.69	6	0.99	0.98	0.05	0.75	
Cho et al. (2019; Non-English Learners)	4	—	—	—	—	—	—	
Foorman et al. (2017; Grade 4) [a]	4	3.11	4	1.00	1.00	< 0.01	0.91	
Foorman et al. (2018; Grade 4) [a]	4	173.18	13	0.85	0.77	0.21	0.89	
Kim & Wagner (2015; Grade 4) [a]	4	4.10	4	1.00	1.00	0.01	0.90	
Harlaar et al. (2010) [a]	10 [b]	17.51	8	0.99	0.99	0.05	0.94	
Foorman et al. (2017; Grade 5)	5	8.78	4	1.00	0.99	0.04	0.96	
Foorman et al. (2018; Grade 5)	5	178.00	13	0.89	0.83	0.20	0.94	
Proctor et al. (2012) [e]	2-4	7.38	8	1.00	1.00	< 0.01	> 1	
Kieffer et al. (2016; Grade 3-5)	3-5	61.90	13	0.96	0.94	0.11	0.83	
Leider et al. (2013) [e]	3-5	13.69	8	0.99	0.98	0.08	> 1	
Lesaux et al. (2010) [a]	4-5	2.14	4	1.00	1.03	< 0.01	0.83	
Two-factor Model with Oral Language Represented as a Bi-factor Model								
Kershaw & Schatschneider (2012; Grade 3) [a]	3	9.87	10	1.00	1.00	< 0.01	0.85	< 0.01
Foorman et al. (2018; Grade 4) [d]	4	60.82	9	0.95	0.89	0.15	0.83	< 0.01
Foorman et al. (2018; Grade 5) [a]	5	23.08	9	0.99	0.98	0.07	0.81	< 0.01
Kieffer et al. (2016; Grade 3-5) [a]	3-5	21.75	10	0.99	0.98	0.06	0.77	< 0.01

Note. *r(RC,OL)* indicates the latent correlation between the oral language factor and the reading comprehension factor; *p*(diff) is the *p* value for the chi-square test of model fit for each data set, compared to the respective two-factor model in Table 3. [a] indicates the best-fitting model for that data set; [b] indicates the sample's mean age was reported in the study, and converted to grade by subtracting the mean age from six; [d] indicates negative residual variance; [e] indicates that the correlation between two latent factors is larger than 1; Dashes indicate that the model did not converge for that data set.

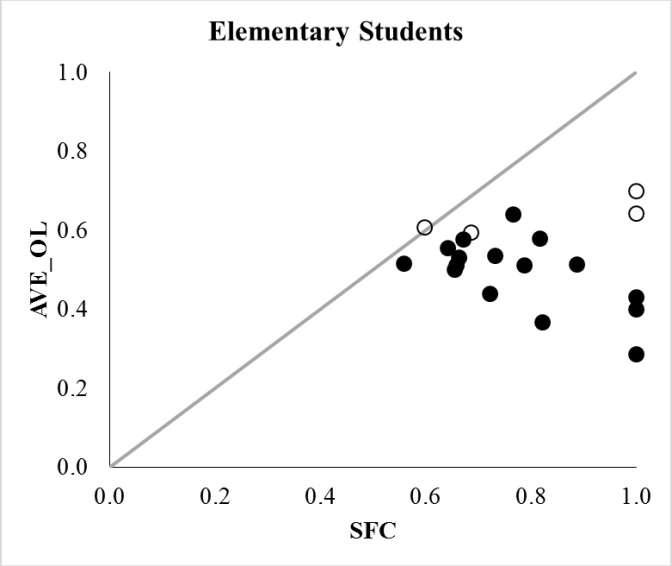

Figure 4 Scatter Plot of the SFC and the AVE for the Oral Language Factor in the Elementary Students

Note. Filled circles indicate that the sample was in one grade, while empty circles indicate that the sample was from several grades. The diagonal line is shown for reference.

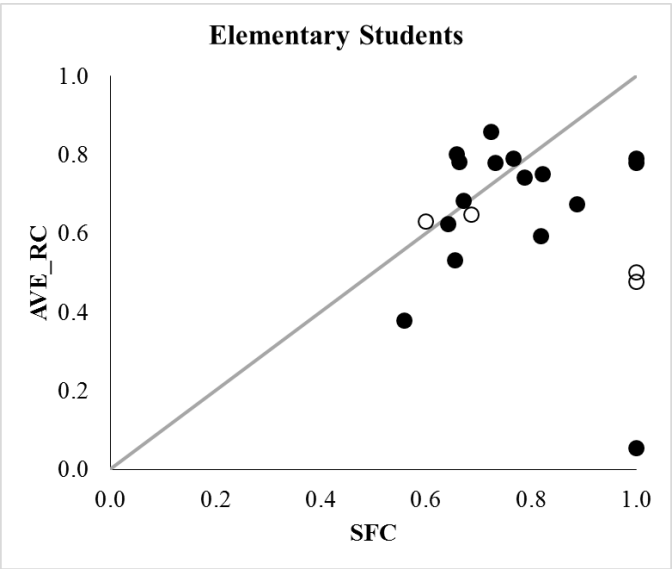

Figure 5 Scatter Plot of the SFC and the AVE for the Reading Comprehension Factor in the Elementary Students

Note. Filled circles indicate that the sample was in one grade, while empty circles indicate that the sample was from several grades. The diagonal line is shown for reference.

3.2 One-factor and Two-factor Model Results for Middle School and High School Students

In this section, we present the model results for middle school and high school students, as well as three samples with wider age ranges including Betjemann et al. (2011; 8-18 years old), Cutting et al. (2009; 9-15 years old) and Spencer et al. (2019; 9-15 years old). Table 4 shows the model fit indices of the one-factor models for 15 data sets in middle school and high school students. According to the model fit criteria, the one-factor model fit well in four datasets but worse in four data sets with *CFI* < 0.90, *TLI* < 0.90 and *RMSEA* > 0.10. In the remaining seven data sets, the model had a lack of fit in *TLI* or *RMSEA* or both. Table 5 shows the model fit indices of the two-factor models for the 15 data sets. The two-factor model fit well in four data sets with latent correlations ranging from 0.80 to 0.93 (see Table 5), but had a lack of fit in all three indices (*CFI*, *TLI*, and *RMSEA*) with latent correlations of 0.97 and 0.93 in Foorman et al. (2018; Grade 8 and 10). Eight models had a lack of fit in *CFI* or *RMSEA* or both with latent correlations ranging from 0.68 to 1.00. One model had a latent correlation greater than one in Foorman et al. (2017; Grade 8), indicating that the unidimensional model was sufficient to represent the language structure in this sample.

Comparing the one-factor model to the two-factor model for each data set, the chi-square test (see Table 4) suggested that the one-factor fit equivalent with the two-factor model in three samples of native English Speakers (Cutting et al., 2009; Foorman et al., 2017; Grade 8; Foorman et al., 2018; Grade 8), and one sample of Chinese students learning English as a second language (Li & Kirby, 2015). The two-factor model fit better in 11 data sets with latent correlations ranging from 0.68 to 0.95. In addition, nine data sets tested multiple oral language skills using at least five measures. We also fit the two-factor model with oral language

represented as a bi-factor model for these data sets. The model fit indices and the chi-square tests (see Table 5) show that the two-factor model with oral language represented as a bi-factor model fit better than the two-factor model without oral language represented as a bi-factor model for each data set. However, one two-factor model with oral language represented as a bi-factor model had negative residual in Kershaw and Schatschneider (2012; Grade 7), one model had a non-positive definite latent covariance matrix in Foorman et al, (2019; Grade 9) and one model failed to converge in Cutting et al. (2009). Thus, the two-factor model with oral language represented as a bi-factor model was considered the best-fitting model in the remaining six data sets. All the best-fitting models are highlighted with the superscript ([a]) in Tables 4 and 5.

Based on the above best-fitting models in middle school and high school students, we calculated the SFC, the AVE for the oral language factor and the AVE for the reading comprehension factor for each data set. Figure 6 shows the scatter plot of the SFCs and the AVEs for the oral language factor, and Figure 7 shows the scatter plot of the SFCs and the AVEs for the reading comprehension factor. The filled circle indicates that the sample was limited to one grade, and the empty circle indicates that the sample included participants across several grades. Figure 6 shows that the AVEs for the oral language factor were lower than the SFCs in all the data sets, except for Fritz (2015; Grade 6-8), Tannenbaum (2009; Grade 7) and Foorman et al. (2018; Grade 10) where the AVEs for the oral language factor were slightly higher with the SFCs. Figure 7 shows that the AVEs for the reading comprehension factor were lower than the SFCs in nine out of 15 data sets, but higher than the SFCs in the remaining six data sets. Overall, the AVEs for two latent factors were lower than the SFCs in nine data sets, suggesting low discriminant validity. In three data sets, the AVEs for the oral language factors were lower than the SFCs, but the AVEs for the reading comprehension factor were not lower than the SFCs. In

three data sets, the AVEs for two latent factors were not lower than the SFCs, implying that the two-dimensional model might be optimal to represent the language structure in these three data sets. The samples were all native English speakers in Fritz (2015; Grade 6-8), Tannenbaum (2009; Grade 7) and Foorman et al. (2018; Grade 10). Only Fritz (2015; Grade 6-8) reported a sample of struggling readers.

Table 4 Fit Statistics of One-factor Models for 15 Data Sets in Middle School and High School Students

Sample	Grade	χ^2	*df*	*CFI*	*TLI*	*RMSEA*	*p*(diff)
Foorman et al. (2018; Grade 6)	6	174.51	14	0.88	0.83	0.19	< 0.01
Sabatini et al. (2014)	6	27.24	5	0.98	0.96	0.14	< 0.01
Foorman et al. (2018; Grade 7)	7	130.17	14	0.92	0.88	0.17	< 0.01
Kershaw & Schatschneider (2012; Grade 7)	7	72.24	14	0.91	0.87	0.15	< 0.01
Tannenbaum (2009; Grade 7)	7	74.89	9	0.90	0.83	0.21	< 0.01
Li & Kirby (2015) [a]	8	23.91	9	0.93	0.88	0.08	1.00
Foorman et al. (2017; Grade 8) [a]	8	4.27	5	1.00	1.00	< 0.01	0.50
Foorman et al. (2018; Grade 8)	8	180.46	14	0.85	0.77	0.23	0.30
Foorman et al. (2018; Grade 9)	9	130.34	14	0.92	0.87	0.19	< 0.01
Foorman et al. (2018; Grade 10)	10	95.06	14	0.88	0.82	0.22	0.04
Kershaw & Schatschneider (2012; Grade 10)	10	32.04	14	0.96	0.94	0.09	0.02
Fritz (2015; Grade 6-8)	6-8	104.14	5	0.87	0.74	0.18	< 0.01
Betjemann et al. (2011)	2-12 [b]	243.90	25	0.90	0.86	0.12	< 0.01
Cutting et al. (2009) [a]	3-9 [b]	22.41	14	0.93	0.90	0.10	0.96
Spencer et al. (2019)	3-9 [b]	120.05	35	0.96	0.95	0.10	< 0.01

Note. *p*(diff) is the *p* value for the chi-square test of model fit for each data set, compared to the respective two-factor model in Table 5; [a] indicates the best-fitting model for that data set; [b] indicates the sample's mean age was reported in the study, and converted to grade by subtracting the mean age from six.

Table 5 Fit Statistics of Two-factor Model and Two-factor Model with Oral Language Represented as a Bi-factor Model for 15 Data Sets in Middle School and High School Students

Model, Sample	Grade	χ^2	*df*	*CFI*	*TLI*	*RMSEA*	*r(RC,OL)*	*p*(diff)
Two-factor Model								
Foorman et al. (2018; Grade 6)	6	147.28	13	0.90	0.84	0.18	0.90	
Sabatini et al. (2014) [a]	6	7.59	4	1.00	0.99	0.06	0.93	
Foorman et al. (2018; Grade 7)	7	119.08	13	0.93	0.88	0.17	0.95	
Kershaw & Schatschneider (2012; Grade 7) [a]	7	44.94	13	0.95	0.92	0.11	0.86	
Tannenbaum, (2009; Grade 7) [a]	7	14.37	8	0.99	0.98	0.07	0.80	
Li & Kirby (2015)	8	23.91	8	0.92	0.85	0.09	1.00	
Foorman et al. (2017; Grade 8) [e]	8	3.81	4	1.00	1.00	< 0.01	> 1	
Foorman et al. (2018; Grade 8)	8	179.38	13	0.85	0.76	0.24	0.97	
Foorman et al. (2018; Grade 9) [a]	9	100.39	13	0.94	0.90	0.17	0.92	
Foorman et al. (2018; Grade 10)	10	90.89	13	0.88	0.81	0.22	0.93	
Kershaw & Schatschneider (2012; Grade 10)	10	26.59	13	0.97	0.95	0.08	0.92	
Fritz (2015; Grade 6-8) [a]	6-8	33.43	4	0.96	0.90	0.11	0.68	
Betjemann et al. (2011) [a]	2-12 [b]	189.90	24	0.93	0.89	0.10	0.86	
Cutting et al. (2009)	3-9 [b]	22.41	13	0.92	0.88	0.11	1.00	
Spencer et al. (2019)	3-9 [b]	96.29	34	0.97	0.96	0.08	0.93	
Two-factor Model with Oral Language Represented as a Bi-factor Model								
Foorman et al. (2018; Grade 6) [a]	6	77.69	9	0.95	0.88	0.16	0.72	< 0.01
Foorman et al. (2018; Grade 7) [a]	7	29.49	9	0.99	0.97	0.09	0.94	< 0.01
Kershaw & Schatschneider (2012; Grade 7) [d]	7	9.44	10	1.00	1.00	< 0.01	0.88	< 0.01
Foorman et al. (2018; Grade 8) [a]	8	28.05	9	0.98	0.96	0.10	0.87	< 0.01
Foorman et al. (2018; Grade 9) [f]	9	36.68	9	0.98	0.95	0.11	0.83	< 0.01
Foorman et al. (2018; Grade 10) [a]	10	33.64	9	0.96	0.91	0.15	0.75	< 0.01
Kershaw & Schatschneider (2012; Grade 10) [a]	10	18.22	10	0.98	0.96	0.07	0.77	0.04
Cutting et al. (2009)	3-9 [b]	—	—	—	—	—	—	—
Spencer et al. (2019) [a]	3-9 [b]	57.30	27	0.99	0.98	0.06	0.92	< 0.01

Note. *r(RC,OL)* indicates the latent correlation between the oral language factor and the reading comprehension factor; *p*(diff) is the *p* value for the chi-square test of model fit for each data set, compared to the respective two-factor model in Table 5; [a] indicates the best-fitting model for that data set; [b] indicates the sample's mean age was reported in the study, and converted to grade by subtracting the mean age from six; [d] indicates negative residual; [e] indicates that the correlation between two latent factors is larger than 1; [f] indicates that the latent covariance matrix is not positive definite; Dashes indicate that the model did not converge for that data set.

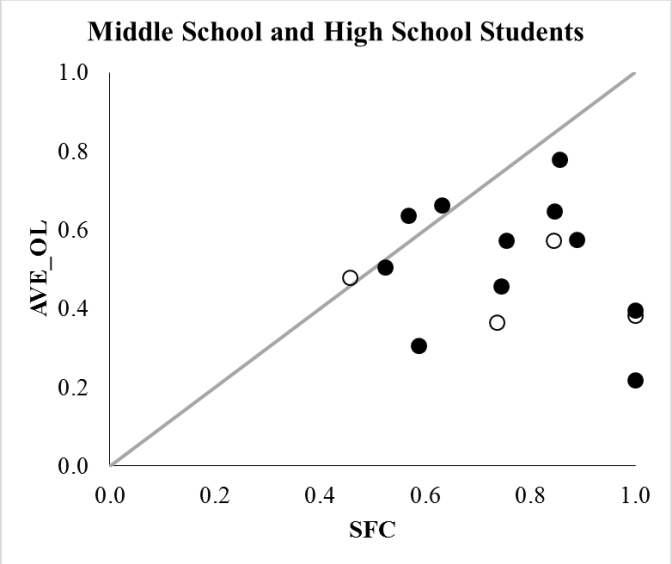

Figure 6 Scatter Plot of the SFC and the AVE for the Oral Language Factor in Middle School and High School Students

Note. Filled circles indicate that the sample was in one grade, while empty circles indicate that the sample was from several grades. The diagonal line is shown for reference.

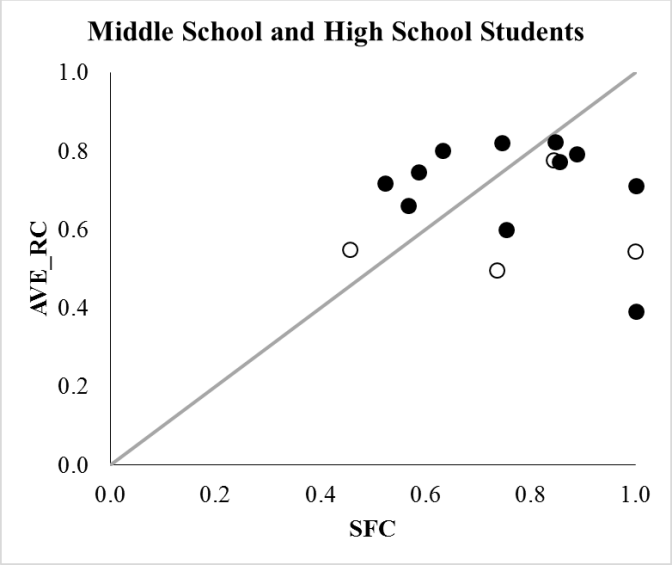

Figure 7 Scatter Plot of the SFC and the AVE for the Reading Comprehension Factor in Middle School and High School Students

Note. Filled circles indicate that the sample was in one grade, while empty circles indicate that the sample was from several grades. The diagonal line is shown for reference.

3.3 One-factor and Two-factor Model Results for Adults

Table 6 shows the model fit indices of one-factor and two-factor models in nine data sets in adults. The one-factor model fit well in four data sets, but worse in one sample of native English speakers (Guo et al., 2011) and one sample of Chinese students learning English as a second language (Zhang & Koda, 2012). In the remaining three data sets, the model had a lack of fit in *TLI* or *RMSEA* or both. The two-factor model fit well in six data sets with latent correlations ranging from 0.65 to 0.97 (see Table 6). One model had a lack of fit in *TLI* and *RMSEA* with latent correlation of 0.76 in Fritz (2015; Age 16-68) and one model had a lack of fit in *TLI* with latent correlation of 0.48 in Zhang and Koda (2012). One model had a latent correlation greater than one in Guo (2018), indicating a unidimensional language structure. The chi-square test (see Table 6) suggested that the one-factor fit equivalently with the two-factor model in one sample of native English Speakers (Van Dyke et al., 2014), and two samples of Chinese students learning English as a second language (Guo & Roehrig, 2011; Guo, 2018). The two-factor model fit better in the remaining six data sets.

Based on the above best-fitting models in adults, we calculated the SFC, the AVE for the oral language factor and the AVE for the reading comprehension factor for each data set. Figure 8 shows the scatter plot of the SFCs and the AVEs for the oral language factors, and Figure 9 shows the scatter plot of the SFCs and the AVEs for the reading comprehension factor. The empty circles indicate that all adult samples had a wider age range. Figure 8 shows that the AVEs for the oral language factor were lower than the SFCs in six data sets, but were equivalent in Fritz (2015; Age 16-68) and slightly higher in Rodriguez (2010) and Zhang and Koda (2012). Figure 9 shows that the AVEs for the reading comprehension factor were lower than the SFCs in seven data sets, but higher in Guo et al. (2011) and Rodriguez (2010). Overall, the SFCs were

higher in four data sets that might not support a two-dimensional language structure. Moreover, the AVEs for two latent factors were much lower in most data sets, indicating low convergent validity. This result suggested that the language structure might diverge into more specialized abilities in adults.

In addition, Figures 8 and 9 also show that the SFC and the AVEs had larger variation across studies, compared with the studies in elementary, middle school and high school students. One reason might be that the adult samples were more heterogeneous due to the wider range of student ages. Another reason for the wide variation in results might be the adult samples differed in reading levels and native languages across studies. Among nine data sets in adults, four samples were native English speakers with two samples having low reading abilities, four samples were Chinese students learning English as a second language, and one sample was Spanish adults learning English as a second language and had low reading ability.

Table 6 Fit Statistics of One-factor and Two-factor Models for Nine Data Sets in Adults

Model, Sample	Age	χ^2	*df*	*CFI*	*TLI*	*RMSEA*	*r(RC,OL)*	*p*(diff)
One-factor Model								
Van Dyke et al. (2014) [a]	16-24	10.52	13	1.00	1.01	< 0.01	—	0.30
Braze et al. (2016)	16-25	20.02	13	1.00	0.99	0.04	—	< 0.01
Guo & Roehrig (2011) [a]	18-23	12.10	9	0.99	0.98	0.04	—	0.15
Guo (2018) [a]	20-25	25.24	9	0.99	0.99	0.08	—	0.70
Rodriguez (2010)	17-40	16.05	5	0.91	0.82	0.17	—	< 0.01
Guo et al. (2011)	18-45	36.54	9	0.86	0.77	0.14	—	< 0.01
Fritz (2015; Age 16-68)	16-68	28.32	5	0.94	0.87	0.14	—	< 0.01
Zhang & Koda (2012)	[c]	61.59	27	0.76	0.68	0.11	—	< 0.01
Zhang (2012)	[c]	18.19	9	0.93	0.88	0.08	—	< 0.01
Two-factor Model								
Van Dyke et al. (2014)	16-24	9.44	12	1.00	1.01	< 0.01	0.97	
Braze et al. (2016) [a]	16-25	9.32	12	1.00	1.00	< 0.01	0.97	
Guo & Roehrig (2011)	18-23	10.05	8	0.99	0.99	0.03	0.90	
Guo (2018) [e]	20-25	25.09	8	0.99	0.98	0.09	> 1	
Rodriguez (2010) [a]	17-40	7.30	4	0.97	0.93	0.10	0.66	
Guo et al. (2011) [a]	18-45	16.00	8	0.96	0.93	0.08	0.65	
Fritz (2015; Age 16-68) [a]	16-68	20.06	4	0.96	0.89	0.13	0.76	
Zhang & Koda (2012) [a]	[c]	39.90	26	0.90	0.87	0.07	0.48	
Zhang (2012) [a]	[c]	8.22	8	1.00	1.00	0.01	0.68	

Note. *r(RC,OL)* indicates the latent correlation between the oral language factor and the reading comprehension factor; *p*(diff) is the *p* value for the chi-square test of model fit for each data set, compared to the respective two-factor model in Table 6; [a] indicates the best-fitting model for that data set; [c] indicates the sample was comprised of graduate students without an age range reported in the study; [e] indicates that the correlation between two latent factors is larger than 1. Dashes indicate that there was no latent correlation in the on-factor model.

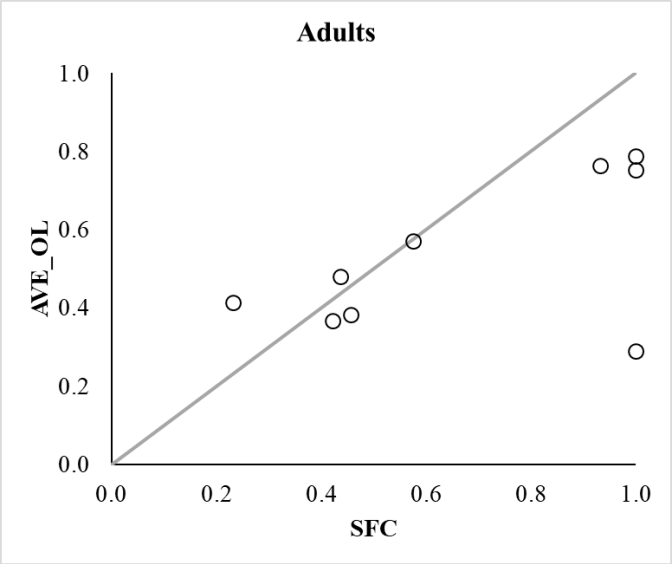

Figure 8 Scatter Plot of the SFC and the AVE for the Oral Language Factor in Adults
Note. Empty circles indicate that the sample included a range of ages.

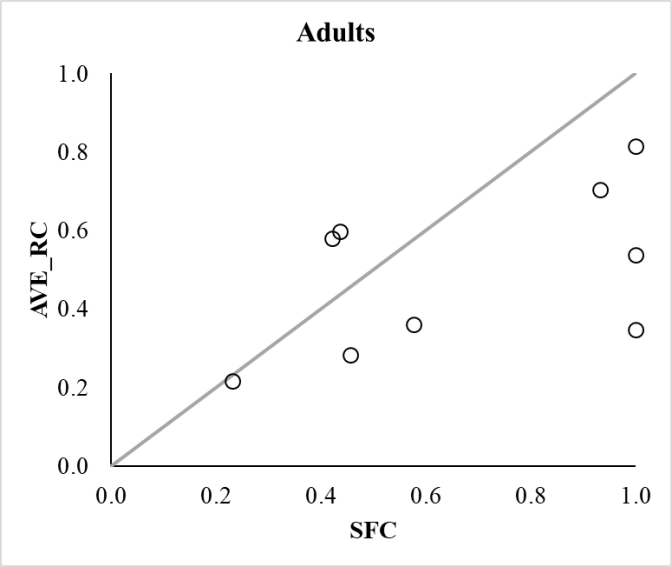

Figure 9 Scatter Plot of the SFC and the AVE for the Reading Comprehension Factor in Adults
Note. Empty circles indicate that the sample included a range of ages.

3.4 Bi-factor Model Results for All Samples

By comparing the unidimensional and two-dimensional models, we found that a unidimensional language structure could explain the data well in 11 out of 44 data sets. The two-dimensional language structure fit better in 33 data sets, whereas the discrimination between the two latent factors was low in most data sets. According to Carroll (1993), the language structure might be better represented as a bi-factor model, especially for developing readers. Thus, we also fit a bi-factor model for each of the 44 data sets. The model fit indices are shown in Table 7 for elementary students and in Table 8 for middle school and high school students and adults. The bi-factor model failed to converge in seven data sets. For these seven data sets, a one-factor model was considered to be the best-fitting model for five data sets and the two-factor model was the best-fitting model for two data sets with latent correlations of 0.75 in Cho et al., (2019; English Learners) and 0.89 in Foorman et al. (2018; Grade 4) according to the above results. The bi-factor model also had negative residual variances in five data sets, which were highlighted with a superscript (^{d}) in Tables 7 and 8. According to the above results, the one-factor model was sufficient to explain the data in Cutting et al. (2009), and the two-factor models (with or without oral language represented as a bi-factor model) fit better in the remaining four data sets with latent correlations ranging from 0.72 to 0.81. In addition, one bi-factor model had a non-positive definite latent correlation matrix in Guo et al. (2011) where the two-factor model fit well with latent correlation of 0.65. Among the remaining 31 data sets, the bi-factor model fit well in 27 data sets but had a lack of fit in at least one index in four data sets.

We calculated the AVE for the general language factor and the AVE for the specific factor of oral language or reading comprehension for the 31 data sets where the bi-factor models had no estimation problems (see Table 7, Table 8 and Figure 10). In Figure 10, the AVE for the

general language factor is represented by a filled square, the AVE for the specific oral language factor is represented by a filled triangle, and the AVE for the specific reading comprehension factor is represented by a filled circle. In elementary students, the AVEs for the general language factor were higher than the AVEs for the specific factor in all 13 data sets. The AVEs for the general language factor were all above 0.4 with an average value of 0.53, while the AVEs for the specific factors were all below 0.3 with an average value of 0.16.

In middle school and high school students, the AVEs for the general language factor were also higher than the AVEs for the specific factor in all 12 data sets. The average of the AVEs for the general language factor was 0.52, and the average of the AVEs for the specific factor was 0.18. However, in four data sets, the AVEs for the general language factor were lower and close with the AVEs for the specific factor (Betjemann et al., 2011; Fritz, 2015 (Grade 6-8); Kershaw & Schatschneider, 2012 (Grade 10); Li & Kirby, 2015), compared with the other eight data sets. The samples were Chinese students learning English as a second language in Li and Kirby (2015) and native English speakers in the other three data sets. The sample was typically developing children in Kershaw and Schatschneider (2012) and struggling readers in Fritz (2015; Grade 6-8). The sample in Betjemann et al. (2011) had wide age range (8-18 years old) and 36% of the sample had a history of reading difficulty.

For adults, only two data sets had high AVEs for the general language factor (Braze et al., 2016; Guo, 2018). The sample in Braze et al. (2016) was native English speakers with most participants having low reading scores. In Guo (2018), the sample was Chinese students learning English as a second language. The AVE for the general language factor may have been high in Guo (2018) because most of oral language and reading comprehension measures were from the TOEFL test and were given in written format, which might result in higher correlations between

these measures. In the remaining four data sets, the AVEs for the general language factor were much lower and close to the AVEs for the specific factor, indicating less shared variance among oral language and reading comprehension measures in these data sets. The samples of these four data sets were English second language learners. One sample was Spanish adults learning English as a second language and had low reading ability (Rodriguez, 2010). The other three samples were all Chinese students learning English as a second language (Guo & Roehrig, 2011; Zhang, 2012; Zhang & Koda, 2012).

Table 7 Fit Statistics of Bi-Factor Model in Elementary Students

Sample	χ^2	*df*	*CFI*	*TLI*	*RMSEA*	*AVE_Lang*	*AVE_OL*	*AVE_RC*
Elementary Students								
Harlaar et al. (2010)	10.17	5	1.00	0.99	0.05	0.53	0.08	—
Chiu (2018)	22.69	6	0.98	0.95	0.10	0.49	0.22	—
Foorman et al. (2017; Grade 3)	0.11	2	1.00	1.01	< 0.01	0.54	—	0.11
Foorman et al. (2018; Grade 3)	18.74	5	0.99	0.98	0.07	0.52	0.18	—
Kershaw & Schatschneider (2012; Grade 3)	16.53	9	0.99	0.98	0.06	0.49	0.15	—
Kim & Wagner (2015; Grade 3)	4.93	2	1.00	0.98	0.08	0.52	—	0.25
Siu & Ho (2015) [d]	3.28	2	1.00	0.98	0.06	—	—	—
Tannenbaum et al. (2006)	11.04	5	0.99	0.97	0.08	0.50	0.18	—
Tannenbaum (2009; Grade 3)	4.04	5	1.00	1.00	< 0.01	0.60	0.15	—
Cho et al. (2019; English Learners)	—	—	—	—	—	—	—	—
Cho et al. (2019; Non-English Learners)	—	—	—	—	—	—	—	—
Foorman et al. (2017; Grade 4)	1.99	2	1.00	1.00	< 0.01	0.52	—	0.13
Foorman et al. (2018; Grade 4)	—	—	—	—	—	—	—	—
Kim & Wagner (2015; Grade 4)	0.41	2	1.00	1.02	< 0.01	0.53	—	0.15
Foorman et al. (2017; Grade 5)	2.96	2	1.00	1.00	0.03	0.63	—	0.06
Foorman et al. (2018; Grade 5) [d]	15.59	10	1.00	0.99	0.04	—	—	—
Proctor et al. (2012)	—	—	—	—	—	—	—	—
Kieffer et al. (2016; Grade 3-5)	35.19	9	0.98	0.95	0.10	0.50	0.19	—
Leider et al. (2013)	—	—	—	—	—	—	—	—
Lesaux et al. (2010)	1.75	2	1.00	1.01	< 0.01	0.50	0.19	—

Note. AVE_Lang indicates the AVE for the general language factor; AVE_OL indicates the AVE for the specific oral language factor; AVE_RC indicates the AVE for the specific reading comprehension factor; [d] indicates negative residual; Dashes indicate that the model did not converge for that data set, or the AVEs were not available.

Table 8 Fit Statistics of Bi-Factor Model in Middle School and High School Students and Adults

Sample	χ^2	*df*	*CFI*	*TLI*	*RMSEA*	*AVE_Lang*	*AVE_OL*	*AVE_RC*
Middle School and High School Students								
Foorman et al. (2018; Grade 6) [d]	109.68	11	0.93	0.86	0.17	—	—	—
Sabatini et al. (2014)	5.45	2	1.00	0.98	0.09	0.71	0.11	—
Foorman et al. (2018; Grade 7)	19.33	10	0.99	0.99	0.06	0.60	0.22	—
Kershaw & Schatschneider (2012; Grade 7)	66.34	11	0.92	0.84	0.16	0.51	0.08	—
Tannenbaum (2009; Grade 7)	1.79	5	1.00	1.02	< 0.01	0.55	0.25	—
Li & Kirby (2015)	17.48	6	0.94	0.86	0.09	0.27	0.11	—
Foorman et al. (2017; Grade 8)	—	—	—	—	—	—	—	—
Foorman et al. (2018; Grade 8)	20.01	10	0.99	0.98	0.07	0.57	0.19	—
Foorman et al. (2018; Grade 9)	32.01	9	0.98	0.96	0.10	0.62	0.15	—
Foorman et al. (2018; Grade 10)	19.86	10	0.99	0.97	0.09	0.62	0.24	—
Kershaw & Schatschneider (2012; Grade 10)	17.96	11	0.99	0.97	0.06	0.43	0.29	—
Fritz (2015; Grade 6-8)	25.28	2	0.97	0.85	0.14	0.35	0.25	—
Betjemann et al. (2011)	88.73	21	0.97	0.95	0.07	0.39	0.18	—
Cutting et al. (2009) [d]	10.03	9	0.99	0.98	0.05	—	—	—
Spencer et al. (2019)	86.04	31	0.97	0.96	0.08	0.60	0.08	—
Adults								
Van Dyke et al. (2014)	—	—	—	—	—	—	—	—
Braze et al. (2016)	8.63	9	1.00	1.00	< 0.01	0.71	0.05	—
Guo & Roehrig (2011)	5.64	5	1.00	0.99	0.02	0.29	0.09	—
Guo (2018)	13.37	6	1.00	0.99	0.07	0.80	—	0.03
Rodriguez (2010)	1.15	2	1.00	1.04	< 0.01	0.35	—	0.37
Guo et al. (2011) [f]	5.59	5	1.00	0.99	0.03	—	—	—
Fritz (2015; Age 16-68) [d]	4.29	2	0.99	0.97	0.07	—	—	—
Zhang & Koda (2012)	38.65	23	0.89	0.83	0.08	0.17	0.31	—
Zhang (2012)	4.26	6	1.00	1.03	< 0.01	0.26	—	0.23

Note. AVE_Lang indicates the AVE for the general language factor; AVE_OL indicates the AVE for the specific oral language factor; AVE_RC indicates the AVE for the specific reading comprehension factor; [d] indicates negative residual; [f] indicates that the latent covariance matrix is not positive definite; Dashes indicate that the model did not converge for that data set, or the AVEs were not available.

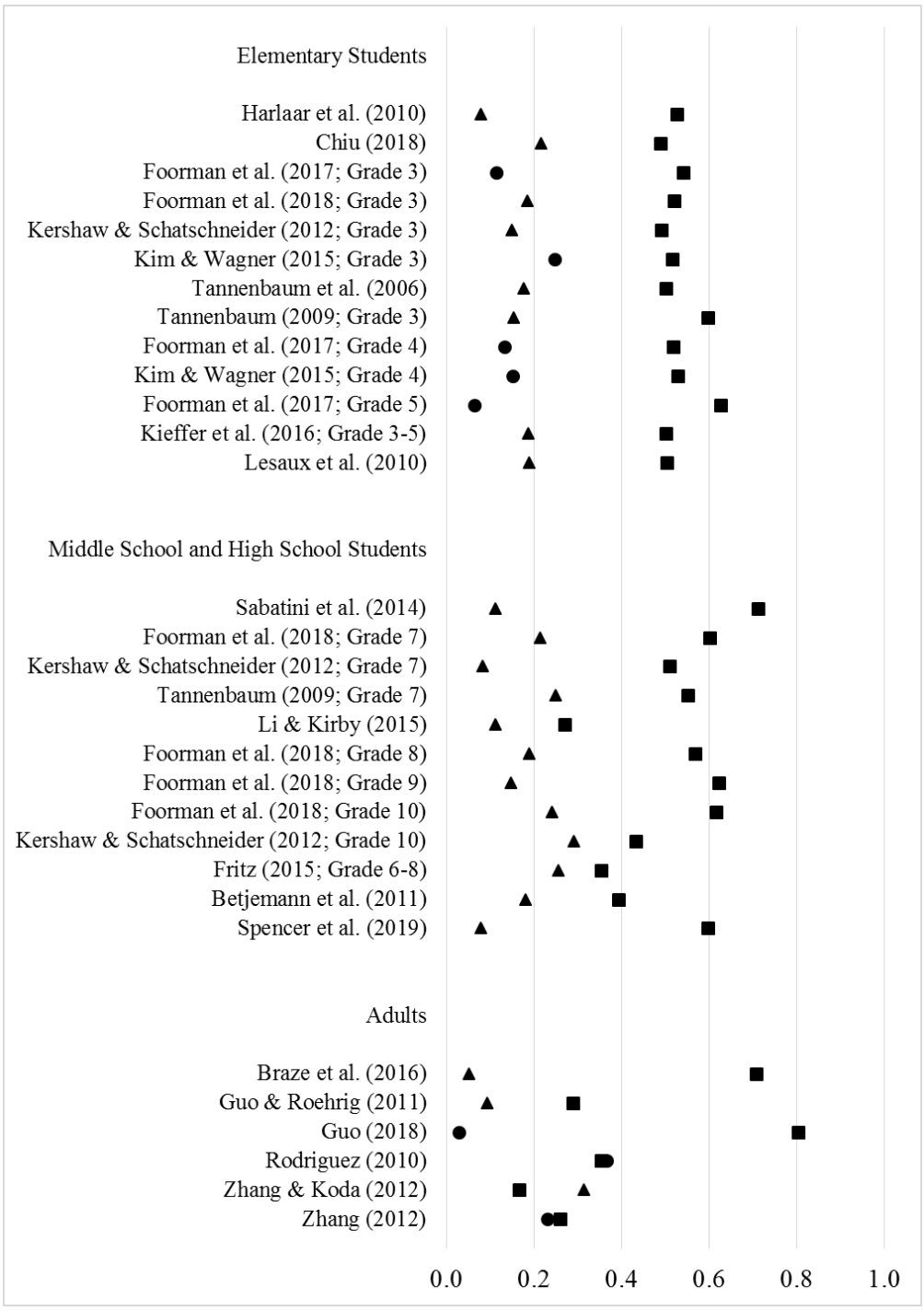

Figure 10 The AVEs of the General Language Factor and the Specific Factor of Oral Language or Reading Comprehension in 31 data sets

Note. Squares indicate the AVE for the general language factor, triangles indicate the AVE for the specific oral language factor, and circles indicate the AVE for the specific reading comprehension factor.

3.5 Effects of the Sample Characteristics on Model Results

Previous studies suggested that the relation between oral language and reading comprehension was different across different sample characteristics, such as age and reading ability (Hua & Keenan, 2017; Keenan, Betjemann, & Olson, 2008). Therefore, we conducted three types of additional comparisons to examine whether the SFCs differed across samples due to their age, reading ability, and whether or not they were native English speakers.

First, with respect to age, the results show that the average SFC was 0.78 with a range of 0.56 to 1.00 in elementary students, 0.76 with a range of 0.46 to 1.00 in middle school and high school students, and 0.67 with a range of 0.23 to 1.00 in adults. Thus, on average, the SFC was similar between the elementary, middle school, and high school students, but lower in adults. The range of the SFC was wider as participants were older.

Second, we examined the model results in the samples which included students with low reading ability. As shown in Table 9, two samples had low reading ability among the elementary students. Both samples included students who performed below a standard score of 85 on the Gates-MacGinitie reading compression test in Cho et al. (2019). The SFC was low for English learners (0.56), compared to the average SFC (0.78) in elementary students. The SFC was high for non-English learners, but the AVE for the reading comprehension factor was very low (0.06), indicating very low convergent validity of the reading comprehension factor. Therefore, the reading comprehension factor was unreliable for this sample.

Among middle school and high school students, three samples included students with low reading ability. In Fritz (2015; Grade 6-8), all students were struggling readers. The SFC was 0.46 for this sample, which was lower than the average SFC (0.76) in middle school and high school students. The other two samples were mixed. Specifically, in Betjemann et al. (2011),

36% of the children had a history of reading difficulty. In Cutting et al. (2009), the sample was mixed with the typically developing children and the children with reading disability. In these two samples, the SFCs were not lower than the average SFC in middle school and high school students. However, the convergent validity for the oral language factor was relatively low in these two mixed samples. The average AVE for the oral language factor was 0.51 in middle school and high school students, while the AVE for the oral language factor was 0.36 in Betjemann et al. (2011) and 0.38 in Cutting et al. (2009).

Among adults, three of the nine samples included participants with low reading ability. In Braze et al. (2016), the sample was recruited from the adult schools, community college campuses, and community gathering places where many people tended to have low reading abilities. However, Braze et al. (2016) did not report whether the final sample had low reading ability. Our results show the SFC and the AVEs for two latent factors were all high for this sample, indicating good convergent validity but low discriminant validity. In Rodriguez (2010), the sample had low reading ability and also English second language learners. The SFC was 0.44, which was lower than the average SFC (0.67) in adults. In Fritz (2015; Age 16-68), all participants were struggling readers. The SFC (0.58) and the AVE (0.36) for the reading comprehension factor were both lower than the average SFC (0.67) and AVE (0.49) for the reading comprehension factor in adults. Therefore, in the samples where all participants had low reading ability, the relation between oral language and reading comprehension appeared lower.

Third, we examined the relation between oral language and reading comprehension in non-native English speakers. As shown in Table 10, among elementary students, six samples included the non-native English speakers. However, the results in Cho et al. (2019; non-English Learners) was not discussed here due to the unreliable reading comprehension factor. In two

bilingual samples, the SFCs were 0.69 in Lesaux et al. (2010; Spanish and English) and 0.65 in Siu and Ho (2015; Chinese (Cantonese) and English). In two composite samples, the SFCs were both 1.00 in Leider et al. (2013; Spanish and English Bilinguals and English Learners) and Proctor et al. (2012; Spanish and English Bilinguals and native English Speakers). In Kieffer et al. (2016, Grade 3-5), the sample included native English speakers with 26% English learners. The SFC for this sample was 0.60. Compared with the average SFC (0.77) in the elementary students, three samples including non-native English speakers show slightly lower SFCs.

Among middle school and high school students, only Li and Kirby (2015) included English second language learners. The SFC for this sample was 1.00, since the one-factor model fit was the best-fitting model. However, the AVEs (0.39, 0.22) were much lower, indicating that the latent factor was much weaker.

Among the adults, four of the nine samples were comprised of native Chinese speakers learning English as a second language. The SFCs were high in two of the Chinese samples (Guo & Roehrig, 2011; Guo, 2018), but low in the two other Chinese samples (Zhang, 2012; Zhang & Koda, 2012). Except for Guo (2018), the AVEs were very low ranging from 0.22 to 0.41 in the Chinese samples. Therefore, in three of the four Chinese samples, there was much less common variance among oral language and reading comprehension measures. The high SFC and AVEs in Guo (2018) might be due to the same test format (written format) and the same test battery (TOEFL test) being used for most measures. In addition, in Rodriguez (2010), the sample was Spanish adults learning English as a second language who also had low reading ability. The SFC was low (0.44) for this sample. Overall, the relation between oral language and reading comprehension was lower in the samples of English second language learners, especially for adults.

Table 9 The SFC and the AVEs for Two Latent Factors in the Samples with Low Reading Ability

Sample	Grade	Reading ability	SFC	AVE_RC	AVE_OL
Elementary students					
Cho et al. (2019; English Learners)	4	Below a standard score of 85 in GMRT Reading Comprehension	0.56	0.38	0.52
Cho et al. (2019; Non-English Learners)	4	Below a standard score of 85 in GMRT Reading Comprehension	1.00	0.06	0.40
Middle School and High School Students					
Fritz (2015; Grade 6-8)	6-8	Struggling Readers	0.46	0.55	0.48
Betjemann et al. (2011)	8-18[b]	36% with history of difficulty	0.74	0.50	0.36
Cutting et al. (2009)	9-15[b]	Typically developing and reading disabilities	1.00	0.54	0.38
Adults					
Braze et al. (2016)	16-25[b]	Most Low	0.93	0.70	0.76
Rodriguez (2010)	17-40[b]	Low level	0.44	0.60	0.48
Fritz (2015; Age 16-68)	16-68[b]	Struggling Readers	0.58	0.36	0.57

Note. GMRT = Gates-MacGinitie Reading Test; [b] indicates the sample's age; AVE_RC indicates the AVE for the reading comprehension factor; AVE_OL indicates the AVE for the oral language factor.

Table 10 The SFC and the AVEs for Two Latent Factors in the Samples of Nan-native English Speakers

Sample	Grade	Native/Non-native	SFC	AVE_RC	AVE_OL
Elementary students					
Siu & Ho (2015)	3	Chinese(Cantonese)–English Bilinguals	0.65	0.53	0.50
Cho et al. (2019; English Learners)	4	English Learners with LEP	0.56	0.38	0.52
Proctor et al. (2012)	2-4	56% native speakers and 44% Spanish–English Bilinguals, 50% LEP	1.00	0.50	0.64
Kieffer et al. (2016; Grade 3-5)	3-5	Native speakers with 26% English Learners	0.60	0.63	0.61
Leider et al. (2013)	3-5	58.5% English Learners, 45.5% Spanish–English Bilinguals	1.00	0.48	0.70
Lesaux et al. (2010)	4-5	Spanish–English Bilinguals	0.69	0.65	0.59
Middle School and High School Students					
Li & Kirby, (2015)	8	English Learners of native Chinese speakers	1.00	0.39	0.22
Adults					
Guo & Roehrig (2011)	18-23[b]	English Learners of native Chinese speakers	1.00	0.35	0.29
Guo (2018)	20-25[b]	English Learners of native Chinese speakers	1.00	0.82	0.79
Rodriguez (2010)	17-40[b]	English learners of native Spanish speakers	0.44	0.60	0.48
Zhang & Koda (2012)	[c]	English Learners of native Chinese speakers	0.23	0.22	0.41
Zhang (2012)	[c]	English Learners of native Chinese speakers	0.46	0.28	0.38

Note. LEP = limited English proficiency; [b] indicates the sample's age; [c] indicates graduate students; AVE_RC indicates the AVE for the reading comprehension factor; AVE_OL indicates the AVE for the oral language factor.

3.6 Sensitivity analyses

Except for sample characteristics, the SFC might also be influenced by other factors, such as sample size, whether or not the study had been published, study quality, and test characteristics. Thus, we further examined whether the SFCs differed across studies with different sample sizes, publication types, reliability for the latent factors, and test format of the oral language measures based on the best-fitting models for each data set.

Sample size and publication type. Figure 11 shows the scatter plot of the sample sizes and the SFCs. Most samples had a sample size between 100 and 300. Only eight samples included more than 400 participants. Overall, the SFCs had larger variation in studies with sample sizes smaller than 300. In the studies with larger sample sizes (> 400), the SFCs were mostly higher except in one sample with low reading ability (Fritz, 2015; Grade 6-8). In addition, the five empty circles in Figure 11 indicate samples that were from unpublished dissertations. The SFCs were relatively low in Fritz (2015) where two samples were comprised of struggling readers, and in Rodriguez (2010) where the sample was English second language learners with low reading ability. Therefore, the effect of reading ability was confounded by a potential publication bias.

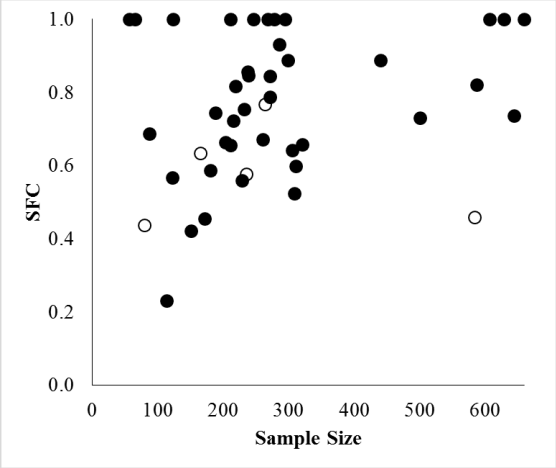

Figure 11 Scatter Plot of Sample Size and the SFC based on the Best-Fitting Model for Each Data Set
Note. Filled circles indicate the journal articles, while empty circles indicate dissertations.

Reliability. Then, we examined the relation between the SFC and study quality indexed by the reliability (AVEs) calculated based on our models. Figure 12 and Figure 13 shows that as the reliability for two latent factors got higher, the SFCs were higher and more homogeneous. In the studies with low reliability, the SFC had large variation. Thus, it should be cautious to interpret the results in these studies. Specifically, the reliability was low for both the oral language factor and the reading comprehension factor in studies with the samples of Chinese students learning English as a second language. In Cho et al. (2019) with the samples having low reading comprehension scores, the reliability was also low, especially for the reading comprehension factor. The reliability was low for the oral language factor in Foorman et al. (2017) where all oral language measures were administrated in written format.

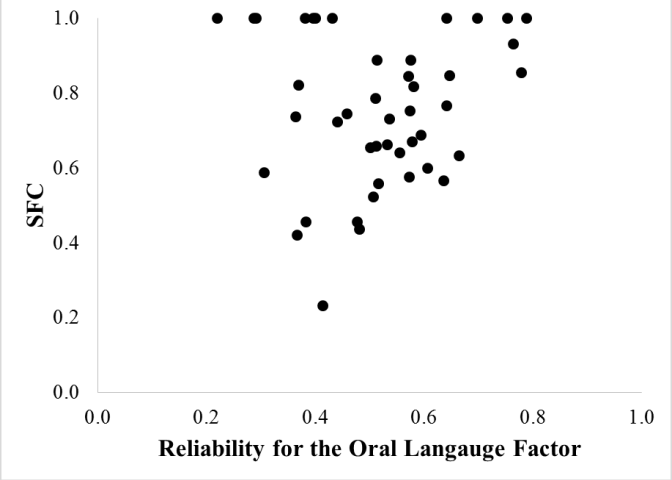

Figure 12 Scatter Plot of the SFC and the Reliability for the Oral Language Factor (AVE) for All 44 Data Sets

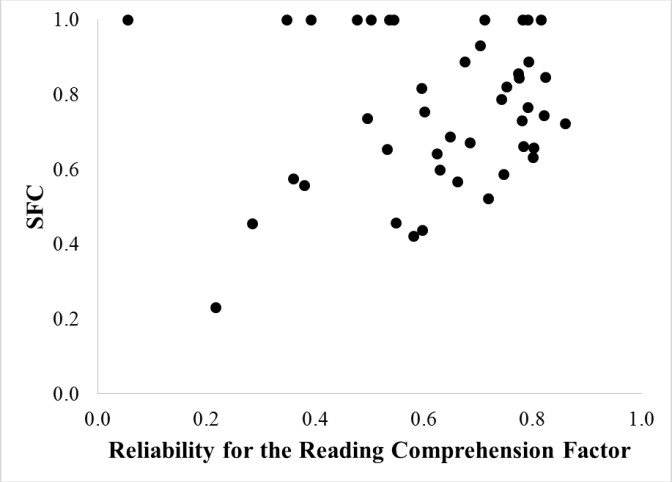

Figure 13 Scatter Plot of the SFC and the Reliability for the Reading Comprehension Factor (AVE) for All 44 Data Sets

Labs or author groups. We also compared the model results across different labs or groups of authors. Since the studies conducted by the same lab or authors were more likely to use the same measures or samples, this might result in groups of more consistent results within labs and different model results across labs or authors. For example, collaborating with other authors, six authors reported two or three data sets involved in our analyses. Cho et al. (2019) also reported two data sets, but they were not examined here due to the unreliable reading comprehension factor for the non-English learners. As shown in Figure 14, the SFCs differed dramatically across six authors. However, the samples were also different within author groups across the data sets reported. Specifically, the samples were Chinese graduate students learning English as a second language in Zhang (2012) and Zhang and Koda (2012). The samples werenative English speakers who were struggling readers in Fritz (2015). The two samples reported by Guo (2018) and Guo and Roehrig (2011) were undergraduate and graduate students with one sample of native speakers, and one sample of English second language learners. The two samples reported in Kim and Wagner (2015) were native English speakers in Grade 3 and 4.

Van Dyke reported one data set in Van Dyke et al. (2014) and another data set in Braze et al. (2016) by collaborating with other authors (both were native English speaking adults). Proctor reported on data set in Proctor et al. (2012), and another data set in Leider et al. (2013), including Spanish–English Bilinguals and native English speakers in Grades 2-4 in Proctor et al. (2012), and mixed with Spanish–English Bilinguals and English Learners in Grades 3-5 in Leider et al. (2013), but the tests used in both studies were the same. Therefore, overlap due to authors was highly confounded with sample characteristics. The differences in SFCs across the authors might be due to the differences in sample characteristics. Moreover, in our analyses, 18 out of 44 data sets were from Florida and used the same reading comprehension tests including the Florida Comprehensive Assessment Test and the Stanford Achievement Test. The average SFC was 0.76 with the range of 0.52 to 1.00 in the data sets from Florida, which was similar with the SFC in the other data sets with the average of 0.78 and range of 0.44 to 1.00.

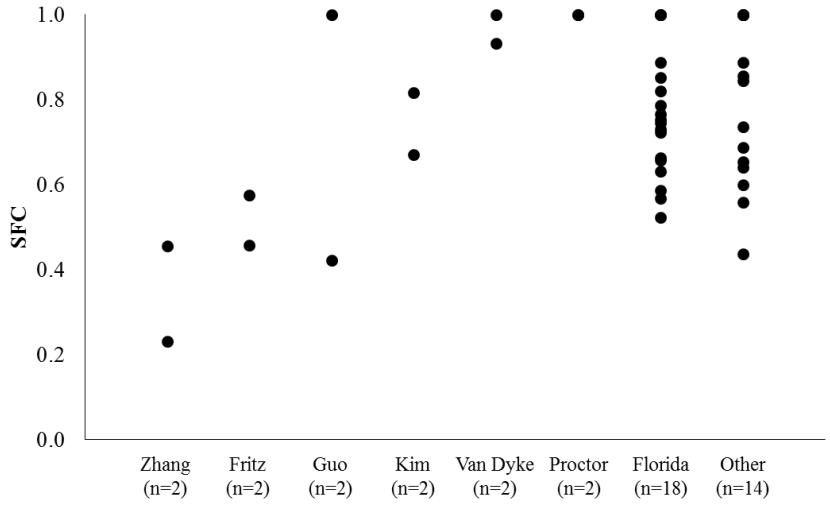

Figure 14 The SFCs by author groups
Note. *n* represents the number of samples (correlation tables) for each author group.

Language tests in written format. In addition, some studies tested oral language ability using measures given in written format, which might inflate the correlations between oral

language and reading comprehension given that reading ability was involved in both oral language and reading comprehension measures. As shown in Table 11, in the first group, nine data sets had all oral language measures administered in written format. One data set had four oral language measures in total with three in written format and one in oral and written format. Among these 10 data sets, the average SFC was 0.84 ranging from 0.23 to 1.00. Specifically, the SFCs were high (above 0.82) in eight data sets, but low (0.23 and 0.46) in two data sets with the samples of Chinese graduate students learning English as a second language (Zhang, 2012; Zhang & Koda, 2012). In the second group, only one oral language measure was administered in oral format in two data sets. The average SFC was 0.54 in these two data sets. In the third group, two out of four oral language measures were administered in oral and written format in two data sets, and two out of five oral language measures were administered in written format in one data set. The average SFC was 0.67 in these three data sets. In the fourth group, only one oral language measure was administered in written format in 10 data sets and only one in oral and written format in one data set, all the other oral language measures were administered in oral format. The average SFC in these 11 data sets was 0.79 ranging from 0.52 to 1.00. In the fifth group, only one out of seven oral language measures was administered in oral format in one data set. It was less likely to influence the SFC, even though the SFC was high (0.84) in this data set. Comparing with the average SFC (0.72) in the other 17 data sets where no oral language measure was administered in written format, the SFCs were mostly high in the data sets in the first group where all oral language measures were administered in written format. Therefore, the relation between oral language and reading comprehension might be inflated if all oral language measures were administered in written format.

Table 11 The SFCs in Samples with the Oral Language Measures in Written Format

Group	Average SFC (range)	Total number of oral language measures	Number of measures in written format	Number of measures in oral and written format	Number of measures in oral format	Number of samples
1	0.84 (0.23 - 1.00)	2	2	0	0	4
		3	3	0	0	3
		4	4	0	0	2
		4	3	1	0	1
2	0.54 (0.42-0.65)	3	2	0	1	1
		3	3	0	1	1
3	0.67 (0.60 – 0.77)	4	0	2	2	2
		5	2	0	3	1
4	0.79 (0.52 – 1.00)	4	1	0	3	3
		5	1	0	4	7
		5	0	1	4	1
5	0.84	7	1	0	6	1

Sensitivity check for written format. Finally, we examined that whether our model results on sample differences were confounded with the test format of the oral language measures, by comparing the model results before and after excluding the ten data sets with all oral language measures in written format. First, there were trivial changes in the average SFC in each age group. Specifically, the average SFC changed from 0.78 to 0.76 among elementary student, 0.76 to 0.71 among middle school and high school students, and did not change among adults. Second, among eight samples with low reading ability, no data set included all oral language measures in written format. Thus, the results on samples with low reading ability were not confounded with the test format of the oral language measures. Third, for nonnative English speakers, five out of 12 samples had all oral language measures administrated in written format. All five samples were Chinese students with one sample in middle school and four samples in college. The SFC largely differed across these five samples with the value of 1 in three samples since the one-factor model was chosen as the besting-fitting model, but low (0.23, 0.46) in the other two samples. Therefore, it was unclear how the SFC in Chinese samples was influenced by

the test format of the oral language measures. However, the AVE values were low in four out of five Chinese samples, indicating low reliability for the latent factors. The reason for the high reliability in Guo et al (2018) might be that most oral language and reading comprehension measures were from the same test battery (i.e., not a straightforward method effect, but a battery effect not modeled here). Fourth, after removing the ten data sets with all oral language measures in written format, the data sets from Florida still had similar average value in SFC with other data sets, but had narrow range because the highest SFC values were in the data sets with all oral language measures in written format.

4 DISCUSSION

4.1 Structural Relation between Oral Language and Reading Comprehension

Using a secondary data analysis approach, the current study examined the structural relation between oral language and reading comprehension by reanalyzing 44 summary data sets reported in 28 studies. Beyond the model fit indices, we also evaluated the discriminant validity of two latent factors by comparing the shared variance between the latent factors (SFC) and the proportion of variance in measure extracted by a factor (AVE). Through reanalysis of 44 summary data sets reported across 25 journal articles and three dissertations, we found that the one-factor model fit well in 11 out of 44 data sets, suggesting that the unidimensional model was adequate to represent the relations among multiple oral language and reading comprehension measures for these 11 data sets, but was not sufficient for the other 33 data sets.

The two-dimensional models (with or without oral language represented as a bi-factor model) fit better in these 33 data sets, but the discriminant validity of two latent factors was frequently low. Specifically, the AVE for the oral language factor was lower than the SFC in 26 out of the 33 data sets, and the AVE for the reading comprehension factor was also lower than the SFC in 18 data sets. Therefore, psychometrically, it was difficult to separate oral language from reading comprehension in these two-dimensional data sets, suggesting that neither the unidimensional model nor the two-dimensional model was the optimal model to represent the language structure tested by oral language and reading comprehension measures. Only in five data sets did the two-dimensional model have acceptable discriminant validity, where the AVEs for the two latent factors were not lower than the SFC.

Recent studies suggest that oral language is better represented as a bi-factor model including a general oral language factor and several specific factors (e.g., vocabulary, syntax)

(Foorman et al., 2015; Kieffer et al., 2016). Our results also support this finding, since the two-factor model with oral language represented as a bi-factor model mostly fit better than the two-factor model without oral language represented as a bi-factor model. Beyond the general oral language factor, the specific factors were weak with the AVEs ranging from 0.10 to 0.31 and less related with reading comprehension with the SFCs ranging from 0.01 to 0.30 (see Appendix C).

However, according to Carroll (1993), not only could oral language be represented as a bi-factor model, all indicators of language and text-based skills might also indicate a general language ability and be better represented as a bi-factor model. We examined this hypothesis by fitting a bi-factor model with a general language factor and one specific factor of oral language or reading comprehension for each data set. The results show that the bi-factor model fit acceptably for 31 out of the 44 data sets with a strong general language factor in elementary, middle, and high school students. Beyond the general factor, the specific factors were much weaker. However, in adults, the general language factor was substantially weaker in four out of the six data sets, indicating that there was much less common variance among oral language and reading comprehension measures. Therefore, the language structure tested by oral language and reading comprehension measures could be represented as a bi-factor model among elementary, middle, and high school students, but not in adults. However, our analyses included few adult samples and most adult samples were English second language learners or had low reading ability. Stronger, more representative evidence is needed from adults who are typically developing and native English speakers.

4.2 Structural Relation between Oral Language and Reading Comprehension across Different Samples

In our analyses, the samples differed in age, reading ability, whether or not they were native English speakers, and tests administered across studies. These factors were mostly confounded with each other, making it difficult to separate their effects on the relation between oral language and reading comprehension. Generally, our results did not suggest a strong age effect in school age children. The relation between oral language and reading comprehension was stronger and more similar in elementary, middle, and high school students. In adults, their relation was mostly lower in the samples of English second language learners. In addition, the general language factor in the bi-factor model did not show a clear, decreasing pattern by age across the elementary, middle, and high school students as suggested by Carroll (1993). However, in the samples with low reading ability or/and non-native English speakers, the relation between oral language and reading comprehension was mostly low, especially in adults. These results were consistent with previous studies (Hoover & Gough, 1990; Lonigan & Burgess, 2017; Keenan, Betjemann, & Olson, 2008; Kershaw & Schatschneider, 2012), which found that the relation between oral language measures and reading comprehension measures was lower for children with low reading ability, but higher for children with high reading ability.

Our results also show that the SFCs were high and homogeneous among studies with high reliability, but had large variation among studies with low reliability. Generally, the reliability was lower in the samples with low reading ability, especially in Cho et al. (2019), and nonnative English speakers. Specifically, the reliability was mostly low in studies with the samples of Chinese students learning English as a second language. The reason might be that the measures used in these studies were unstandardized or standardized based on native English

speakers. Thus, more high quality data is needed to examine the language structure in this population. The reliability was also low on the oral language factor only in Foorman et al. (2017) where all oral language measures were administrated in written format. Theoretically, it was questionable that these measures could represent a latent oral language ability. Statistically, our model results also did not support that these measure represented a reading ability only, because the factor loadings were much lower for these measures but high for the reading comprehension measures, while the one-factor model fit well. The reason might be that a different test paradigm (computer-adaptive tasks) was used to test the language ability in Foorman et al. (2017), which caused low correlations between these tasks and the standardized reading comprehension tests.

Moreover, we examined whether the model results may have been heavily influenced by the labs/authors and the test formats. We found that authors were more likely to conduct studies using similar samples, which made author effects essentially inseparable from sample effects. In addition, the authors in the same lab were likely to use the same measures. In our analyses, 18 out of 44 data sets were from one research group in Florida. The same reading comprehension measures were used in these data sets. By comparing these 18 data sets to other data sets, we did not find compelling visual evidence that the model results differed compared to those outside of this research group.

An additional consideration is that the test format used in oral language measures might bias the model results. Our results show that in studies with all oral language measures administered in written format, the relation between oral language and reading comprehension was considerably higher, since the reading ability was required in both oral language and reading comprehension measures. Therefore, we should be cautious in interpreting the model results for

the 10 data sets where all oral language measures were administered in written format (see Table 11 and Appendix B).

4.3 Limitations and Future Directions

There are several limitations to the current study. First, the paper search was restricted to the last 10 years, 2009 to 2019. Many other studies that were not published during this time period might have provided valuable information for answering the research questions but were excluded from the current analyses.

Second, the current study is only focused on untimed oral language and reading comprehension measures. It is possible that ignoring timed measures could be thought to yield more homogeneous measures, increasing the likelihood of higher factor correlations and unidimensional models. Timed measures raise the possibility of additional cognitive abilities involving speed and fluency, which would have substantially complicated the current investigation. Therefore, it will be necessary to extend or critically evaluate our model results by including timed measures in future studies. To better explore the range of language and literacy abilities, other skills, such as decoding and fluency may be informative.

Third, among the 44 data sets in the current analyses, only nine samples were among adults. Moreover, five of these adult samples were comprised of second language learners and two samples were native English speakers with low reading ability. Thus, more studies testing typically developing adults are needed to make stronger conclusions about the structure of language abilities in adults.

Fourth, the sample characteristics and the test characteristics appeared to be important features influencing the language structure. However, it was difficult to examine each feature's effect on the language structure in our analyses, since these features were confounded with each

other. In particular, the tests were very different across studies. For example, some studies tested reading comprehension with the Woodcock-Johnson passage comprehension subtest and Gates-MacGinitie reading comprehension test, while other studies included the Qualitative Reading Inventory. The current analysis did not have enough distinctive pairings to be able to dependably model effects for test batteries. To answer this question, a meta-SEM (structural equation model; Cheung & Chan, 2005; Cheung & Chan, 2009) would be needed, which would require more studies with sufficient overlap across all possible measures.

Fifth, study quality is always an important issue in meta-analysis and secondary data analysis based on the summary statistics reported in different studies. The data from low quality studies might cause the artificial results, or even bias the findings. Therefore, a systematic review of the study quality and an examination of its influence on the findings were necessary. The current study only addressed the study quality issues on sample size, whether or not the study had been published, and our model-estimated reliability for the latent factors. Other issues of study quality which may limit our conclusions are needed to be addressed in the future studies.

A sixth, related limitation is that while the current models are theoretically based in cognitive and structural expectations, the resulting parameters and cross-study effects are essentially descriptive and analyzed visually—sample sizes and other meta-analytic moderating effects were not modeled across studies. One future extension could be to incorporate standard errors into the estimates and use them in the resulting plots.

A seventh limitation is that the vast majority of these studies were sampled from schools in which the students shared instruction (Mehta, Foorman, Branum-Martin, & Taylor, 2005) and this clustering in classrooms and schools could not be accounted for here. Ignored classroom clustering could potentially inflate student-level relations, such as the correlation between factors

or the goodness of fit for the bi-factor model.

4.4 Conclusion

To conclude, the current study found a strong general language factor across a wide variety of tests which involved either oral or written language in elementary, middle, and high school students. Beyond the general language factor, the specific factors for oral or written language were relatively weak. However, both general and specific factors were much weaker in adults, especially in the adult samples who were English second language learners or had low reading ability, implying that the language structure might diverge into more specialized abilities in adults.

Implications for research. Overall, our results highlight serious problems for the common practice of treating oral language and reading comprehension as two distinct constructs. The common variance due to a general language proficiency is large. This common variance, if ignored in multiple regression, would result in multicollinearity, leading to artifacts of suppression and conclusions suggesting that measures with otherwise high construct validity would seem distinct, unrelated, or unimportant. Trait versus method or sub-trait/specific factor effects are quite general in research (Maul, 2013). The current study highlights the prevalence and importance of these general ability effects for reading and language research.

Implications for instruction. The results for typically developing schoolchildren suggest that language is a general ability which has widespread effects across diverse measures. The specific aspects of oral versus written abilities suggest that targeting curricula and interventions toward these two specific areas is likely to be effective practice. The results for adults and lower performing students, while based on small, highly selected samples, show lower relations and higher discriminant validity among measures—suggesting that for struggling students,

instruction and intervention might need to be targeted toward individual skills or smaller clusters of specific abilities.

REFERENCES

Adlof, S. M., Catts, H. W., & Little, T. D. (2006). Should the simple view of reading include a fluency component? *Reading and Writing*, *19*(9), 933-958.

*Betjemann, R. S., Keenan, J. M., Olson, R. K., & DeFries, J. C. (2011). Choice of reading comprehension test influences the outcomes of genetic analyses. *Scientific Studies of Reading*, *15*(4), 363-382.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: Wiley.

*Braze, D., Katz, L., Magnuson, J. S., Mencl, W. E., Tabor, W., Van Dyke, J. A., ... & Shankweiler, D. P. (2016). Vocabulary does not complicate the simple view of reading. *Reading and writing*, *29*(3), 435-451.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage focus editions*, *154*, 136-136.

Byle, K. A., & Cucina, J. M. (2014). *Evidence that g isn't a higher-order construct: Bifactor fits better*. Poster presented at the 29th meeting of SIOP (APA Division 14), Honolulu, HI.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of educational psychology*, *96*(1), 31.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.

Catts, H. W., Hogan, T. P., & Adlof, S. M. (2005). Developmental changes in reading and reading disabilities.

Chall, J. (1967). *Learning to read: The great debate*. New York: McGraw-Hill.

Cheung, M. W. L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods, 10*(1), 40-64.

Cheung, M. W. L., & Chan, W. (2009). A two-stage approach to synthesizing covariance matrices in meta-analytic structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(1), 28-53.

*Chiu, Y. D. (2018). The Simple View of Reading across Development: Prediction of Grade 3 Reading Comprehension from Prekindergarten Skills. *Remedial and Special Education*, *39*(5), 289–303.

*Cho, E., Capin, P., Roberts, G., Roberts, G. J., & Vaughn, S. (2019). Examining sources and mechanisms of reading comprehension difficulties: Comparing English learners and non-English learners within the simple view of reading. *Journal of Educational Psychology*, *111*(6), 982–1000.

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.

Cucina, J. M., & Howardson, G. N. (2017). Woodcock-Johnson-III, Kaufman Adolescent and Adult Intelligence Test (KAIT), Kaufman Assessment Battery for Children (KABC), and Differential Ability Scales (DAS) support Carroll but not Cattell-Horn. *Psychological Assessment, 29*(8), 1001-1015.

*Cutting, L. E., Materek, A., Cole, C. A. S., Levine, T. M., & Mahone, E. M. (2009). Effects of fluency, oral language, and executive function on reading comprehension performance. *Annals of Dyslexia, 59*(1), 34–54.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative

contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific studies of reading*, *10*(3), 277-299.

Foorman, B. R., Herrera, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). The structure of oral language and reading and their relation to comprehension in Kindergarten through Grade 2. *Reading and Writing*, *28*(5), 655-681.

Foorman, B. R., Koon, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). Examining general and specific factors in the dimensionality of oral language and reading in 4th–10th grades. *Journal of Educational Psychology*, *107*(3), 884–899.

*Foorman, B. R., Petscher, Y., & Herrera, S. (2018). Unique and common effects of decoding and language factors in predicting reading comprehension in grades 1–10. *Learning and Individual Differences*, *63*, 12–23.

*Foorman, B. R., Petscher, Y., Stanley, C., & Truckenmiller, A. (2017). Latent profiles of reading and language and their association with standardized reading outcomes in kindergarten through tenth grade. *Journal of research on educational effectiveness*, *10*(3), 619-645.

Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of Marketing Research*, 382–388.

Francis, D. J., Fletcher, J. M., Catts, H. W., & Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension. *Children's reading comprehension and assessment*, 369-394.

Francis, D. J., Kulesz, P. A., & Benoit, J. S. (2018). Extending the Simple View of Reading to Account for Variation Within Readers and Across Texts: The Complete View of Reading (CVR i). Remedial and Special Education, 39(5), 274-288.

*Fritz, C. M. (2015). *Modeling Reading Constructs with Struggling Readers at Different Ages*. (Ph.D.). Georgia State University, Atlanta.

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

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

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological review*, *101*(3), 371.

*Guo, L. (2018). Modeling the Relationship of Metacognitive Knowledge, L1 Reading Ability, L2 Language Proficiency and L2 Reading. *Reading in a Foreign Language*, *30*(2), 209–231.

Guo, Y. (2009). *The role of vocabulary knowledge, syntactic awareness and metacognitive awareness in reading comprehension of adult English language learners*. (69). ProQuest Information & Learning.

*Guo, Y., & Roehrig, A. D. (2011). Roles of General versus Second Language (L2) Knowledge in L2 Reading Comprehension. *Reading in a Foreign Language*, *23*(1), 42–64.

*Guo, Y., Roehrig, A. D., & Williams, R. S. (2011). The Relation of Morphological Awareness and Syntactic Awareness to Adults' Reading Comprehension: Is Vocabulary Knowledge a Mediating Variable? *Journal of Literacy Research*, *43*(2), 159–183.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: A global perspective* (7th ed.). Upper Saddle River, NJ: Pearson.

*Harlaar, N., Cutting, L., Deater-Deckard, K., DeThorne, L. S., Justice, L. M., Schatschneider,

C., … Petrill, S. A. (2010). Predicting individual differences in reading comprehension: A twin study. *Annals of Dyslexia*, *60*(2), 265–288.

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

Hua, A. N., & Keenan, J. M. (2017). Interpreting reading comprehension test results: Quantile regression shows that explanatory factors can vary with performance level. *Scientific Studies of Reading*, *21*(3), 225-238.

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160-212.

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.

Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of learning disabilities*, *47*(2), 125-135.

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction*, *22*(5), 354-367.

Kendeou, P., Van den Broek, P., White, M. J., & Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of educational psychology*, *101*(4), 765.'

Kent, W. M. V. (2013). *Intonation and reading skills in fourth-grade students*. (Ph.D.). Wayne State University, Ann Arbor.

*Kershaw, S., & Schatschneider, C. (2012). A Latent Variable Approach to the Simple View of Reading. *Reading and Writing: An Interdisciplinary Journal*, *25*(2), 433–464.

*Kieffer, M. J., Petscher, Y., Proctor, C. P., & Silverman, R. D. (2016). Is the Whole Greater than the Sum of Its Parts? Modeling the Contributions of Language Comprehension Skills to Reading Comprehension in the Upper Elementary Grades. *Scientific Studies of Reading*, *20*(6), 436–454.

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

Kim, Y. S. G. (2017). Why the simple view of reading is not simplistic: Unpacking component skills of reading using a direct and indirect effect model of reading (DIER). *Scientific Studies of Reading, 21*(4), 310-333.

* Kim, Y.-S. G., & Wagner, R. K. (2015). Text (oral) reading fluency as a construct in reading development: An investigation of its mediating role for children from Grades 1 to 4. *Scientific Studies of Reading*, *19*(3), 224–242.

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

Kintsch, W., & Rawson, K. A. (2007). Comprehension. In M. J. Snowling & C. Hulme (Eds.), *The science of reading*: A handbook (pp. 209–226). Oxford, UK: Blackwell.

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (5th ed.). New York, NY: Guilford Press.

Language and Reading Research Consortium. (2015). The dimensionality of language ability in young children. *Child Development*, *86*(6), 1948-1965.

*Leider, C. M., Proctor, C. P., Silverman, R. D., & Harring, J. R. (2013). Examining the Role of Vocabulary Depth, Cross-Linguistic Transfer, and Types of Reading Measures on the

Reading Comprehension of Latino Bilinguals in Elementary School. *Reading and Writing: An Interdisciplinary Journal*, *26*(9), 1459–1485.

*Lesaux, N. K., Crosson, A. C., Kieffer, M. J., & Pierce, M. (2010). Uneven Profiles: Language Minority Learners' Word Reading, Vocabulary, and Reading Comprehension Skills. *Journal of Applied Developmental Psychology*, *31*(6), 475–483.

*Li, M., & Kirby, J. R. (2015). The effects of vocabulary breadth and depth on English reading. *Applied Linguistics*, *36*(5), 611–634.

Lonigan, C. J., & Burgess, S. R. (2017). Dimensionality of reading skills with elementary-school-age children. *Scientific Studies of Reading*, *21*(3), 239-253.

Lonigan, C. J., & Milburn, T. F. (2017). Identifying the dimensionality of oral language skills of children with typical development in preschool through fifth grade. *Journal of Speech, Language, and Hearing Research*, *60*(8), 2185-2198.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, *114*(1), 185–199.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. Mc-Donald* (pp. 275–340). Mahwah, NJ: Erlbaum.

Mehta, P. D., Foorman, B. R., Branum-Martin, L., & Taylor, W. P. (2005). Literacy as a unidimensional multilevel construct: Validation, sources of influence, and implications in a longitudinal study in grades 1 to 4. *Scientific Studies of Reading*, *9*(2), 85-116.

Mellard, D. F., Fall, E., & Woods, K. L. (2010). A path analysis of reading comprehension for adults with low literacy. *Journal of learning disabilities*, *43*(2), 154-165.

Muthén, L.K. and Muthén, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Nanda, A. O., Greenberg, D., & Morris, R. (2010). Modeling child-based theoretical reading constructs with struggling adult readers. *Journal of Learning Disabilities*, *43*(2), 139-153.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures: Issues and applications*. Thousand Oaks, CA: Sage.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. *The science of reading: A handbook*, 227-247.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22-37.

*Proctor, C. P., Silverman, R. D., Harring, J. R., & Montecillo, C. (2012). The Role of Vocabulary Depth in Predicting Reading Comprehension among English Monolingual and Spanish-English Bilingual Children in Elementary School. *Reading and Writing: An Interdisciplinary Journal*, *25*(7), 1635–1664.

Quinn, J. M. (2016). Predictors of Reading Comprehension: A Model-Based Meta-analytic Review. *Unpublished doctoral dissertation). Florida State Univerisity*.

*Rodriguez, A. S. (2010). *The influence of cross-linguistic input and L2 proficiency on L2 reading comprehension among Spanish-speaking adults learning English as a second language*. (Ph.D.). City University of New York, Ann Arbor.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: calculating and interpreting statistical indices. *Psychological methods*, *21*(2), 137.

*Sabatini, J. P., O'Reilly, T., Halderman, L. K., & Bruce, K. (2014). Integrating Scenario-Based and Component Reading Skill Measures to Understand the Reading Behavior of Struggling

Readers. *Learning Disabilities Research & Practice*, *29*(1), 36–43.

*Siu, C. T.-S., & Ho, C. S.-H. (2015). Cross-Language Transfer of Syntactic Skills and Reading Comprehension among Young Cantonese-English Bilingual Students. *Reading Research Quarterly*, *50*(3), 313–336.

*Spencer, M., Richmond, M. C., & Cutting, L. E. (2019). Considering the Role of Executive Function in Reading Comprehension: A Structural Equation Modeling Approach. *Scientific Studies of Reading*, 1-21.

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental psychology*, *38*(6), 934.

*Tannenbaum, K. R. (2009). *Relationships between measures of word knowledge and reading comprehension in third- and seventh-grade children*. (69). ProQuest Information & Learning.

*Tannenbaum, K. R., Torgesen, J. K., & Wagner, R. K. (2006). Relationships between Word Knowledge and Reading Comprehension in Third-Grade Children. *Scientific Studies of Reading*, *10*(4), 381–398.

Tighe, E. L., Wagner, R. K., & Schatschneider, C. (2015). Applying a multiple group causal indicator modeling framework to the reading comprehension skills of third, seventh, and tenth grade students. *Reading and Writing: An Interdisciplinary Journal*, *28*(4), 439–466.

Tighe, E. L., & Schatschneider, C. (2016). Examining the relationships of component reading skills to reading comprehension in struggling adult readers: A meta-analysis. *Journal of learning disabilities*, *49*(4), 395-409.

Tilstra, J., McMaster, K., Van den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: Components of the simple view of reading across grade levels. *Journal of*

Research in Reading, *32*(4), 383–401.

Tomblin, J. B., & Zhang, X. (2006). The dimensionality of language ability in school-age children. *Journal of Speech, Language, and Hearing Research*.

*Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition*, *131*(3), 373–403.

*Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *Modern Language Journal*, *96*(4), 558–575.

*Zhang, D., & Koda, K. (2012). Contribution of morphological awareness and lexical inferencing ability to L2 vocabulary knowledge and reading comprehension among advanced EFL learners: Testing direct and indirect effects. *Reading and Writing: An Interdisciplinary Journal*, *25*(5), 1195–1216.

APPENDICES

Appendix A

Table 12 Table of Sample Characteristics

Sample	N	Age	Grade	Reading ability	Native/Non-native
Betjemann et al. (2011)	644	8-18		36% with history of reading difficulty	Native speakers
Braze et al. (2016)	286	16-25		Most Low	Native speakers
Chiu (2018)	305		3		94.1% Native speakers
Cho et al. (2019; English Learners)	229		4	Below a standard score of 85 in GMRT Reading Comprehension	English Learners with LEP
Cho et al. (2019; Non-English Learns)	211		4	Below a standard score of 85 in GMRT Reading Comprehension	Non English Learners
Cutting et al. (2009)	56	9-15		Typically developing and reading disabilities	Native speakers
Foorman et al. (2017; Grade 3)	607		3	Typically developing	Native speakers with 7.36% English Learner
Foorman et al. (2017; Grade 4)	587		4	Typically developing	Native speakers with 6.91% English Learner
Foorman et al. (2017; Grade 5)	659		5	Typically developing	Native speakers with 8.54% English Learner
Foorman et al. (2017; Grade 8)	629		8	Typically developing	Native speakers with 8.72% English Learner
Foorman et al. (2018; Grade 3)	501		3	Typically developing	Native speakers with 6% LEP
Foorman et al. (2018; Grade 4)	271		4		Native speakers with 8% LEP
Foorman et al. (2018; Grade 5)	321		5		Native speakers with 6% LEP
Foorman et al. (2018; Grade 6)	309		6		Native speakers with 3% LEP
Foorman et al. (2018; Grade 7)	299		7		Native speakers with 3% LEP
Foorman et al. (2018; Grade 8)	232		8		Native speakers with 2% LEP
Foorman et al. (2018; Grade 9)	238		9		Native speakers with 2% LEP
Foorman et al. (2018; Grade 10)	122		10		Native speakers with 3% LEP
Fritz (2015; Age 16-68)	236	16-68		Struggling Readers	Native speakers
Fritz (2015; Grade 6-8)	584		6-8	Struggling Readers	Native speakers
Guo & Roehrig (2011)	278	18-23		Typically developing	English Learners of native Chinese speakers
Guo et al. (2011)	151	18-45		Typically developing	Native speakers
Guo (2018)	268	20-25		Typically developing	English Learners of native Chinese speakers

Harlaar et al. (2010)	440	10			Native speakers
Kershaw & Schatschneider (2012; Grade 10)	180		10	Typically developing	Native speakers
Kershaw & Schatschneider (2012; Grade 3)	215		3	Typically developing	Native speakers
Kershaw & Schatschneider (2012; Grade 7)	188		7	Typically developing	Native speakers
Kieffer et al. (2016; Grade 3-5)	311		3-5		Native speakers with 26% English Learners
Kim & Wagner (2015; Grade 3)	260		3		Native speakers
Kim & Wagner (2015; Grade 4)	219		4		Native speakers
Leider et al. (2013)	123		3-5		58.5% English Learners, 45.5% Spanish–English Bilinguals
Lesaux et al. (2010)	87		4-5		Spanish–English Bilinguals
Li & Kirby, (2015)	246		8	Typically developing	English Learners of native Chinese speakers
Proctor et al. (2012)	294		2-4		56% native speakers and 44% Spanish–English Bilinguals, 50% LEP
Rodriguez (2010)	80	17-40		Low level	English learners of native Spanish speakers
Sabatini et al. (2014)	237		6		
Siu & Ho (2015)	211		3	Typically developing	Chinese(Cantonese)–English Bilinguals
Spencer et al. (2019)	271	9-15			Native speakers
Tannenbaum et al. (2006)	203		3		
Tannenbaum (2009; Grade 3)	264		3	Typically developing	Native speakers
Tannenbaum (2009; Grade 7)	165		7	Typically developing	Native speakers
Van Dyke et al. (2014)	65	16-24		Typically developing	Native speakers
Zhang & Koda (2012)	113			Typically developing	English Learners of native Chinese speakers
Zhang (2012)	172			Typically developing	English Learners of native Chinese speakers

Note. LEP = limited English proficiency; GMRT = Gates-MacGinitie Reading Test; the empty cell indicates that the demographics were not reported in the studies.

Appendix B

Test abbreviations

CASL	Comprehension Assessment of Spoken Language
CELF-4	Clinical Evaluation of Language Fundamentals–Fourth Edition
CREVT	Comprehensive Receptive and Expressive Vocabulary Test
EOWPVT-3	Expressive One-Word Picture Vocabulary Test–Third Edition
FCAT 2.0	Florida Comprehensive Assessment Test 2.0
GISA	Global, Integrated Scenario-based Assessments
GMRT	Gates-MacGinitie Reading Test
GORT-3	Gray Oral Reading Test–Third Edition
GORT-4	Gray Oral Reading Test–Fourth Edition
GSRT-3	Gray Silent Reading Tests–Third Edition
KBIT	Kaufman Brief Intelligence Test
LPT-R	Language Processing Test–Revised
NDRT	Nelson-Denny Reading Test
PIAT-R	Peabody Individual Achievement Test–Revised
PPVT-4	Peabody Picture Vocabulary Test–Fourth Edition
PPVT-III	Peabody Picture Vocabulary Test–Third Edition
QRI-3	Qualitative Reading Inventory–Third Edition
QRI-5	Qualitative Reading Inventory–Fifth Edition
RISE	Reading Inventory and Scholastic Evaluation
ROWPVT-2	Receptive One-Word Picture Vocabulary Test–Second Edition
SARA	Study Aid and Reading Assistant
SAT-10	Stanford Achievement Test–Tenth Edition
SRI-2	Standardized Reading Inventory–Second Edition
TLC-E	Test of Language Competence–Expanded
TMS	Test of Morphological Structure
TNL	Test of Narrative Language
TOAL-4	Test of Adolescent and Adult Language–Fourth Edition
TOLD-I: 3	Test of Oral Language Development, Intermediate–Third Edition
TOWK	Test of Word Knowledge
WASI	Wechsler Abbreviated Scale of Intelligence
WISC-III	Wechsler Intelligence Scale for Children–Third Edition
WJ-III	Woodcock Johnson Tests of Achievement–Third Edition
WLPB-R	Woodcock Language Proficiency Battery-Revised
WMLS-R	Woodcock–Muñoz Language Survey–Revised
WRMT-R/NU	Woodcock Reading Mastery Test–Revised/Normative Update

Table 13 Table of Measures in each Sample

Sample	Reading Comprehension Measure	Listening Comprehension Measures (Test Format)	Vocabulary Measure (Test Format)	Morphology Measures (Test Format)	Syntax Measure (Test Format)
Betjemann et al. (2011)	WJ-III Passage Comprehension PIAT Reading Comprehension GORT-3 Reading Comprehension QRI-3 Reading: Passage Retelling QRI-3 Reading: Open-ended Comprehension Questions	WJ-III Oral Comprehension (Oral) QRI-3 Listening: Passage Retelling (Oral) QRI-3 Listening: Open-ended Comprehension Questions (Oral) KNOW-IT Test (Oral)			
Braze et al. (2016)	WJ-III Passage Comprehension PIAT-R Reading Comprehension (half items) GMRT Level AR Reading Comprehension	WJ-III Oral Comprehension (Oral) PIAT-R Listening Comprehension (half items) (Oral)	WASI Vocabulary (Oral) PPVT-III (Oral)		
Chiu (2018)	GMRT Reading Comprehension WRMT-R/NU Passage Comprehension QRI-5 Reading Comprehension	TNL Narrative Comprehension (Oral) CELF-4 Understanding Spoken Paragraphs (Oral) Adapt from QRI-5 Listening Comprehension (Oral)			
Cho et al. (2019; EL and NonEL)	WJ-III Passage Comprehension GMRT Reading Comprehension	WJ-III Oral Comprehension Odd items (Oral) WJ-III Oral Comprehension Even items (Oral)	KBIT Verbal Knowledge-Word (Oral) KBIT Verbal Knowledge-World (Oral)		
Cutting et al. (2009)	WRMT-R/NU Passage comprehension GORT-4 Reading Comprehension	TLC-E Making Inferences (Oral and Written)	PPVT-III (Oral)		TOLD-I:3 Grammatic Comprehension (Oral) TOLD-I:3 Sentence Combining (Oral)

					TLC-E Ambiguous Sentences (Oral)
Foorman et al. (2017; Grade 3-5, 8)	A computer-adaptive reading comprehension task			A computer-adaptive morphological task (Written)	A computer-adaptive syntactic task (Written)
	SAT-10 Reading Comprehension				
	FCAT 2.0 Reading Comprehension				
Foorman et al. (2018; Grade 3)	GMRT Reading Comprehension	CELF-4 Concepts and Following Directions (Oral)	PPVT-4 (Oral)		CELF-4 Sentence Structure (Oral)
	FCAT 2.0 Reading Comprehension				CELF-4 Recall Sentences (Oral)
					CASL Grammaticality Judgment (Oral)
Foorman et al. (2018; Grade 4-10)	GMRT Reading Comprehension		SARA Vocabulary (Oral)	SARA Morphological Awareness (Written)	CELF-4 Recall Sentences (Oral)
	FCAT 2.0 Reading Comprehension		PPVT-4 (Oral)		CASL Grammaticality Judgment (Oral)
Fritz (2015; Grade 6-8)	WJ-III Passage Comprehension	WJ-III Oral Comprehension (Oral)	Word Test 2-Adolescent: Flexible Word Use (Oral)		
	SRI-2 Passage Comprehension		WASI Vocabulary (Oral)		
Fritz (2015; Age 16-68)	WJ-III Passage Comprehension		PPVT-III (Oral)		TOLD-I: 3 Word Ordering (Oral)
	GORT-4 Reading Comprehension		Boston Naming Test (Oral)		

Guo & Roehrig (2011)	TOEFL Reading Comprehension		Vocabulary Level Test (Written)		TOAL-4 Sentence Combination (Written)
	GSRT-3 Reading Comprehension		Depth of Vocabulary Knowledge (Written)		Syntactic Awareness Questionnaire (Written)
Guo et al. (2011)	NDRT-Form G Reading comprehension		PPVT-III (Written)	Grammatical Application Test–Revised Wug Test (Written)	Syntactic Awareness Questionnaire (Written)
	GMRT Reading Comprehension		CREVT Expressive Vocabulary (Oral)		
Guo (2018)	TOEFL Word Comprehension		Vocabulary Level Test (Written)		TOEFL Structure and Written Expression (Written)
	TOEFL Text Comprehension				Grammaticality Judgment (Written)
	TOEFL Critical Comprehension				
Harlaar et al. (2010)	WRMT-R Passage Comprehension	TNL Narrative Comprehension (Oral)	CELF Word Classes (Oral)		
	PIAT Reading Comprehension	CELF Understanding Spoken Paragraphs (Oral)	Boston Naming Test (Oral)		
Kershaw & Schatschneider (2012; Grade 3, 7, 10)	FCAT Reading Comprehension	Listening comprehension 1 (Oral)	WASI Vocabulary (Oral)		
	SAT-9 Reading Comprehension	Listening comprehension 2 (Oral)	WASI Similarities (Oral or Picture)		
		Listening comprehension 3 (Oral)			
Kieffer et al. (2016; Grade 3-5)	GMRT Reading Comprehension		WMLS Picture Vocabulary (Oral)	Extract the Base test (Written)	CELF Formulated Sentences (Oral)
	WMLS Passage Comprehension		CELF Word Classes 2 (Oral)	Nonword suffix choice (Written)	
Kim & Wagner (2015; Grade 3-4)	WJ-III Passage Comprehension	WJ-III Oral Comprehension (Oral)			

	WRMT-R Passage Comprehension An experimental Reading Comprehension test	An experimental Listening Comprehension test (Oral)			
Leider et al. (2013)	WMLS-R Passage Comprehension GMRT-4 Reading Comprehension		WMLS-R Picture Vocabulary (Oral) CELF Word Class 2 (Oral)	Extract the Base test (Written)	CELF Formulated Sentences (Oral)
Lesaux et al. (2010)	WLPB-R Passage Comprehension GMRT Reading Comprehension	WLPB-R Listening Comprehension (Oral)	PPVT (Oral) WLPB-R Picture Vocabulary (Oral)		
Li & Kirby (2015)	GMRT Reading Comprehension Summary Writing		GMRT Vocabulary test (Written) Adapt from PPVT (Oral and Written) Multiple Meanings Vocabulary test (Written)	Base Identification task (Written)	
Proctor et al. (2012)	WMLS-R Passage Comprehension GMRT Reading Comprehension		WMLS-R Picture Vocabulary (Oral) CELF Word Classes 2 (Oral)	Extract the Base test (Written)	CELF Formulated Sentences (Oral)
Rodriguez (2010)	QRI-5 Reading Comprehension (Open-ended questions) QRI-5 Reading Comprehension (Multiple choice) QRI-5 Reading Comprehension (Retell)		WMLS-R Picture Vocabulary (Oral) WMLS-R Verbal Analogies (Oral)		
Sabatini et al. (2014)	GISA Reading Comprehension		RISE Vocabulary test (Written)	RISE Morphology test (Written)	RISE Sentence Processing test (Written)

	RISE Reading Comprehension				
Siu & Ho (2015)	Sentence comprehension task Passage comprehension task		Short version of PPVT–4 (Oral)	Morphosyntactic correction task (Written)	Word order correction task (Written)
Spencer et al. (2019)	WJ-III Passage Comprehension GMRT Reading Comprehension	KNOW-IT Test (Oral)	TOWK Receptive Vocabulary (Oral) TOWK Expressive Vocabulary (Oral) TOWK Synonyms (Written)	TMS Decomposition (Oral) TMS Derivation (Oral) Test of Morphological Relatedness (Oral)	TLC-E Ambiguous Sentences (Oral)
Tannenbaum et al. (2006)	FCAT Reading Comprehension SAT-9 Reading Comprehension		PPVT–III (Oral) WISC–III Vocabulary (Oral) LPT–R Multiple Meanings (Written) LPT–R Attributes (Oral)		
Tannenbaum (2009; Grade 3, 7)	FCAT Reading Comprehension SAT-10 Reading Comprehension		EOWPVT-3 (Oral) ROWPVT-2 (Oral) TOWK Multiple Contexts (Oral and Written) WORD Test-2 Associations (Oral and Written)		
Van Dyke et al. (2014)	WJ-III Passage Comprehension PIAT-R Reading Comprehension (odd items)	WJ-III Oral Comprehension (Oral) PIAT-R Reading Comprehension (even items) (Oral)	PPVT-R (Oral)		

	GORT-4 Reading Comprehension (passages 5, 7, 9) GMRT-4 Reading Comprehension			
Zhang & Koda (2012)	Reading Comprehension-word supply question Reading Comprehension-conjunction question question Reading Comprehension-co-reference question question Reading Comprehension-textual inference question Reading Comprehension-gist question question	Vocabulary Levels Test (Written) Word Associates Test (Written)	Morphological Awareness task (identify the root) (Written) Lexical inferencing test (Written)	
Zhang (2012)	Reading Comprehension-co-reference question question Reading Comprehension-textual inference question Reading Comprehension-gist question question	Vocabulary Levels Test (Written) Word Associates Test (Written)		Grammatical error correction task (Written)

Appendix C

Among 44 data sets, 13 data sets tested multiple oral language skills using at least five measures. We fit a two-factor model with oral language represented as a bi-factor model for these 13 data sets. The model fit well and was considered to be the besting-fitting model for nine data sets. In seven data sets, oral language was represented as a general oral language factor with a specific vocabulary factor. In one data set, oral language was represented as a general oral language factor with a specific vocabulary factor and a specific listening comprehension factor. In one data set, oral language was represented as a general oral language factor and a specific morphology factor.

Figure 15 displays the SFCs between the reading comprehension factor and the specific factor of vocabulary, listening comprehension, or morphology for each data set. Squares indicate the SFCs between the reading comprehension factor and the specific vocabulary factor. The larger square with error bar indicates the average SFC between the reading comprehension factor and the vocabulary factor was 0.14 with a range of 0.01 to 0.30. Circles indicate the SFCs between the reading comprehension factor and the listening comprehension factor. Triangles indicate the SFCs between the reading comprehension factor and the morphology factor.

Figure 16 displays the AVEs for the specific factors for each data set. Squares indicate the AVEs for the specific vocabulary factor. The larger square with error bar indicates the average AVE for the specific vocabulary factor was 0.24 with a range of 0.10 to 0.31. Circles indicate the AVEs for the specific listening comprehension factor. Triangles indicate the AVE for the specific morphology factor. The results show that both the SFCs and the AVEs were very low for each data set, indicating that beyond the general oral language factor, the specific factor

of vocabulary, listening comprehension or morphology was very weak and had very low relation with the reading comprehension factor.

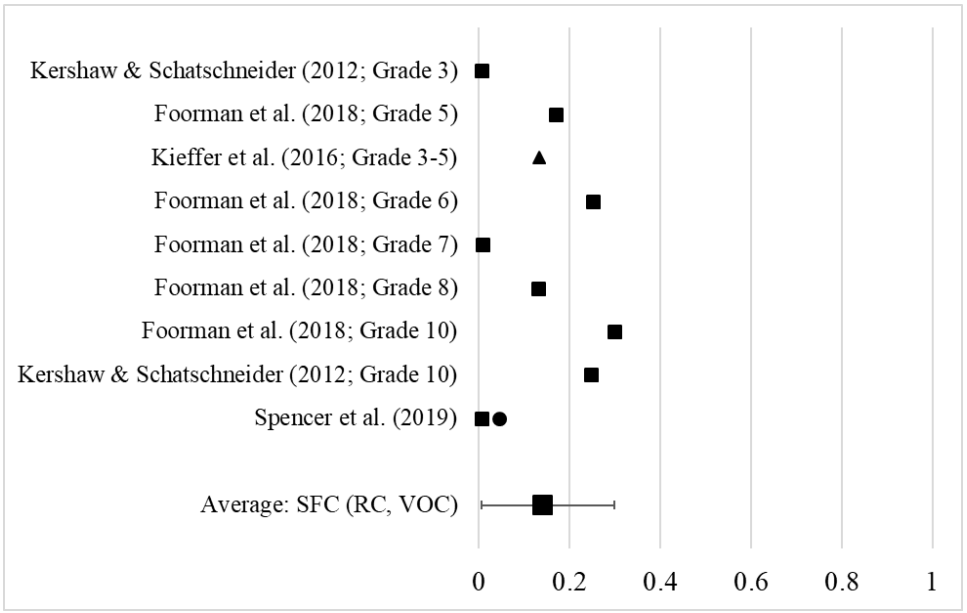

Figure 15 The SFC between Reading Comprehension Factor and the Specific Factor of Oral Language based on Nine Two-Factor Models with Oral Language Represented as a Bi-factor Model

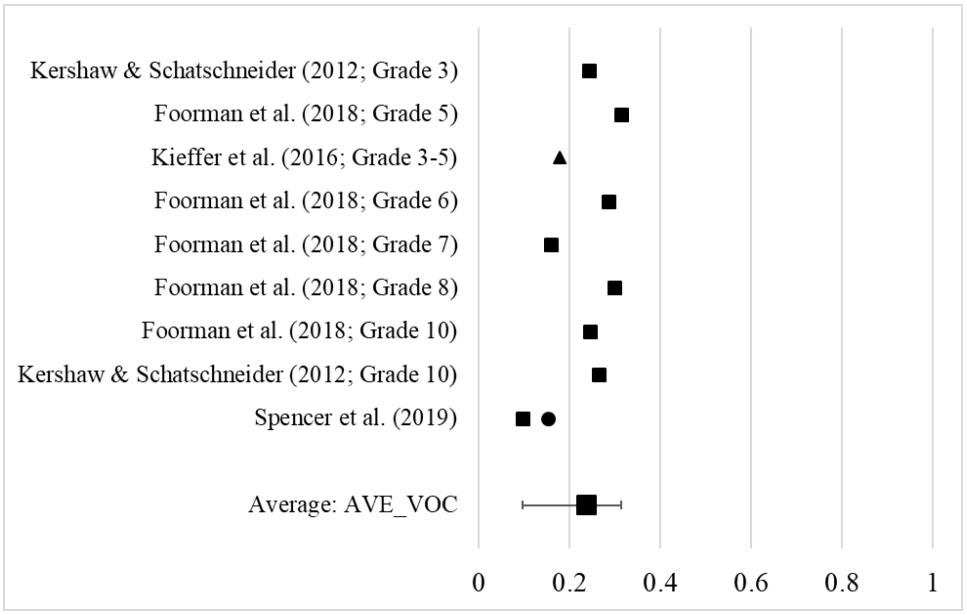

Figure 16 The AVE for the Specific Factor of Oral Language based on Nine Two-Factor Models with Oral Language Represented as a Bi-factor Model