The Dimensionality of Language Ability in Young Children

Language and Reading Research Consortium

Accepted for publication in Child Development

Date of acceptance: 04/05/2015

Author Note

This paper was prepared by a Task Force of the Language and Reading Research Consortium (LARRC) consisting of Jill Pentimonti (Convener), Kate Cain, Laura Justice, and Ann O'Connell. LARRC project sites and investigators are as follows:

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

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

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

University of Kansas (Lawrence, KS): Hugh Catts[1] (Site PI), Mindy Bridges, Diane Nielsen.

University of Nebraska-Lincoln (Lincoln, NE): Tiffany Hogan[2] (Site PI), Jim Bovaird, J. Ron Nelson.[3]

1. Hugh Catts is now at Florida State University.

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

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

This work was supported by grant # R305F100002 of the Institute of Education Sciences' Reading for Understanding Initiative. We are deeply grateful to the numerous staff, research associates, school administrators, teachers, children, and families who participated. Key personnel at study sites include: Garey Berry, Beau Bevens, Jennifer Bostic, Shara Brinkley,

Lori Chleborad, Dawn Davis, Michel Eltschinger, Tamarine Foreman, Rashaun Geter, Sara Gilliam, Miki Herman, Trudy Kuo, Gustavo Lujan, Carol Mesa, Denise Meyer, Marcie Mutters, Maria Moratto, Trevor Rey, and Stephanie Williams.

 Correspondence concerning this work should be sent to Jill Pentimonti, email: pentimonti.1@osu.edu.

Abstract

The purpose of this study was to empirically examine the dimensionality of language ability for young children (4 to 8 years) from pre-kindergarten to third grade (n = 915), theorizing that measures of vocabulary and grammar ability will represent a unitary trait across these ages, and to determine whether discourse skills represents an additional source of variance in language ability. Results demonstrated emergent dimensionality of language across development with distinct factors of vocabulary, grammar, and discourse skills by third grade, confirming that discourse skills are an important source of variance in children's language ability and represent an important additional dimension to be accounted for in studying growth in language skills over the course of childhood.

The Dimensionality of Language Ability in Young Children

Scholars within a number of disciplines (e.g., speech-hearing sciences, developmental psychology, linguistics) tend to view language ability as a complex system consisting of multiple dimensions. As a result, textbooks and discussion on language development and disorders are often organized by different dimensions of language (e.g., Bishop, 1997; Schwartz, 2009). The presumed multi-dimensionality of language ability has practical import as well. For instance, the diagnosis of disordered language in children is based on evidence of symptoms (delays or deficits) across one or more of the dimensions of vocabulary, grammar, narrative, discourse, or pragmatics (American Psychiatric Association, 2012). Such practices further reinforce the perspective that language ability comprises multiple dimensions and that these dimensions represent independent abilities. Further, it has been shown that early weaknesses in vocabulary, grammar, and discourse comprehension and production are related to later reading comprehension difficulties (Catts, Fey, Zhang, & Tomblin, 1999; Nation, Cocksey, Taylor, & Bishop, 2010). Thus, our understanding of the structure of language extends beyond oral language to literacy and educational attainment.

A premise relevant to the present study is that such conceptualizations of language ability are largely theoretical in nature. To date, there has been limited empirical assessment of the extent to which the various theorized dimensions of language do indeed represent latent abilities at a given point in a child's development. Importantly, there is limited empirical justification for measuring language ability (and diagnosing language disorders) as a set of distinct psychological dimensions (using subtests for interpretation) or as a unitary psychological trait (using the composite for interpretation; see Tomblin & Zhang, 1996). The present study empirically examined the dimensionality of language ability for consecutive age groups from four to eight years. Using a cross-sectional sample and measures of the language dimensions of vocabulary,

grammar and discourse, study outcomes help to advance our fundamental understanding of the nature of language ability among children and inform practices that depend upon this understanding.

Dimensionality of Language Ability

Language is a complex system; thus, it comes as no surprise that there are different ways to conceptualize its dimensions. One approach is to distinguish between language performance in different *modalities*, that is comprehension (what is understood) versus production (how language is used). In support of this approach, there are clear dissociations between comprehension and production during early language development: for example, children between 8 to 16 months understand many more spoken words than they produce (Bates, Dale, & Thal, 1995), and comprehension and production result in specific and different patterns of brain activity in one-year-olds (Mills, Coffey, & Neville, 1993). However, research empirically examining the distinctiveness of the comprehension versus production modality, at least for children five years of age and older, fails to support these modalities as true dimensions of language ability (see Tomblin & Zhang, 2006). Thus, the research reported here cuts across modality to examine the distinctiveness of theorized dimensions of language, and considers whether these should be conceptualized as a single system or separate subsystems in the early primary grades.

The majority of research that speaks to this question has examined the relationship between vocabulary and grammar. One school of thought is that vocabulary and grammar represent separate modular systems each governed by different underlying systems or mechanisms (Pinker, 1998). Certainly we find that in languages such as English, one-word production precedes the use of grammar in the form of inflections on single words or

combinations of words. This temporal asynchrony in lexical and grammatical development could be taken as evidence that different modular learning mechanisms for words and grammar exist. However, the computational approach for simulating language development has demonstrated that the same learning mechanism could be used for both vocabulary and grammar acquisition (Plunkett & Marchman, 1993). Further, grammatical development depends upon a critical lexical base and different phases of grammar development are linked to different phases of lexical growth (Marchman & Bates, 1994). These two strands of evidence lead some to argue that the apparent dimensionality of different language systems might be an emergent, rather than innate, property of a developing language system (Elsabbagh & Karmiloff-Smith, 2004).

Empirical Investigations of Language Dimensionality

The period of language development examined in this paper is 4 to 8 years of age, which is typically later than the age range examined in the modularity debates on vocabulary and grammar (e.g., Bates et al., 1995). During this period, extensive language development occurs and many language and language-related disorders are diagnosed, such as specific language impairment (SLI) and dyslexia. For these reasons, we consider it essential to understand how best to conceptualize language within this age range to inform not only theory, but also such practices as assessment, intervention, and instruction.

We know of only one prior empirical investigation of the dimensionality of language ability in typically developing young children, conducted by Tomblin and colleagues (Tomblin & Zhang, 2006; Tomblin, Nippold, Fey, & Zhang, 2014). Using data collected as part of a longitudinal epidemiological study of language disorders among kindergarten children, Tomblin and Zhang (2006) examined whether children's language ability at kindergarten, second, fourth, and eighth grades represented a uni- or multidimensional trait, and whether representation of this

trait (or traits) changed as children grew older. The dimensions of language ability assessed were vocabulary and grammar and were measured for both modalities of comprehension and production.

From an initial sample of 1,929 kindergarteners, about 600 children were followed over time (Tomblin & Zhang, 2006). Modality (comprehension and production) was best considered to be uni-dimensional. Critically, there was evidence of emergent dimensionality: in the linguistic dimension, the correlations between factors representing vocabulary and grammar at each grade (kindergarten, second grade, fourth grade and eighth grade respectively) showed that these were high (e.g., *r*s = .94, .93, .90, and .78 with increasing grade). However, the lower value (.78) seen at eighth grade suggested that language ability might be considered multi-dimensional in the later grades. The possibility of emergent dimensionality was supported by confirmatory factor analysis. At the kindergarten, second grade, and fourth grade time-points, the one-dimensional model fit well and represented the most parsimonious approach to conceptualizing dimensionality of language ability at these grades. In contrast, at eighth grade, the two-dimensional model separating vocabulary and grammar was the better fitting model. This work suggests that a uni-dimensional model may not be the most accurate description of language abilities across development.

Language dimensionality: A broader perspective

Tomblin and Zhang (2006) note that they did not include measures of language that are often included in the dimension of discourse (or text) and, therefore, that their findings cannot be generalized more broadly. An understanding of the structure of language including discourse is critical for both theoretical and practical reasons. Theoretically, vocabulary and grammar have been conceptualized as lower-level skills that are important, but not sufficient, for discourse

comprehension, which has been called a higher-level dimension of language (Kim, in press; Lepola, Lynch, Laakkonen, Silven, & Niemi, 2012). Measures of discourse include the ability to understand narrative, generate inferences, and monitor comprehension language. Notably, these discourse-level language skills develop early: for example, comprehension monitoring during a narrative is evident between 30 to 39 months (Skarakis-Doyle & Dempsey, 2008) and infants' early word learning depends on inferential processing (Tomasello, Carpenter, & Liszkowski, 2007). Furthermore, when discourse level language skills such as inference, narrative skills, and comprehension monitoring are included alongside vocabulary and grammar, each predicts variance in listening comprehension in 4- to 6-year-olds (Kim, in press; Lepola et al., 2012; Silva & Cain, in press). This literature suggests that if the dimension of discourse were included in an examination of language dimensionality, it might form a separable dimension to vocabulary and grammar. On the practical side, there is a strong body of work demonstrating the language bases of literacy outcomes (Berninger & Abbott, 2010; Catts, Hogan, & Fey, 2003; Hogan, Cain, & Bridges, 2012). More accurate models of language structure in pre-readers and beginner readers therefore have broader implications beyond simply language development and difficulties: it might better inform our identification of children at risk of later literacy difficulties.

Aims of the Present Study

The overarching aim of the present study was to further our understanding of the dimensionality of language ability among children in the 4- to 8-year-old age range. Two interests were addressed, the first of which was to examine dimensionality of language ability with respect to vocabulary and grammar. Prior work has indicated that these two dimensions of language represent a unitary trait across these ages (see Tomblin & Zhang, 2006); however, a distinguishing feature of the present work is that we extend the study of language dimensionality

to children in pre-kindergarten, and include a wider collection of multiple indicators of language ability for each proposed dimension to derive latent variables. The second interest was to determine if discourse skills represent an additional source of variance in language ability that is unique from vocabulary and grammar at these ages. Despite theoretical representation in the literature of the distinctiveness of vocabulary and grammar vs. discourse language skills in the literature on reading development (Cain, Oakhill, & Bryant, 2004; Dickinson, Golinkoff, & Hirsh-Pasek, 2010; Kim, in press; Lepola et al., 2012) the present work is the first effort of which we are aware to empirically determine whether these theorized dimensions in fact represent valid dimensions of language ability. In addition, we assessed whether these dimensions become more salient across development, as found by Tomblin and Zhang (2006) theorizing that the emergence of discourse language skills as a dimension of language ability might only be observed for the older children in our sample.

Method

Data for the present study were collected as part of a multi-site, five-year longitudinal research project conducted by the Language and Reading Research Consortium (LARRC). The purpose of the LARRC longitudinal study was to identify and model language processes important for reading comprehension in children enrolled in pre-kindergarten (PK) to grade three (G3). The longitudinal study featured a cohort design, such that in Year 1 five cohorts of children in each of grades PK and G3 were ascertained and then followed until G3; children thus graduated from the study annually. The study design thus yields both cross-sectional data and longitudinal data. In the present study, we utilized the cross-sectional data collected for five waves of children (PK-G3) during the 2011-2012 academic year.

Participants

Participants were 915 children in PK through G3 recruited from 69 schools and/or preschool centers at four data collection sites (Arizona, Kansas, Nebraska, Ohio). To recruit participants into the study, teachers provided children's caregivers with recruitment packets that contained literature on LARRC, a caregiver consent form, a family screener questionnaire, and a return envelope. Caregivers could return the forms directly back to the teacher or mail them to the research team. This process was used until each site recruited up to 110 PK children and at least 30 (but no more than 50) children from each of K-G3; the number of participants per site was designed to be approximately equal. The children in our sample were similar to the relative population of children at recruitment sites across key demographic characteristics (e.g., eligibility to receive free/reduced price lunch and membership in racial/ethnic categories). Upon receipt of consent for a given child, research staff reviewed the family screener to ensure that consented children could appropriately participate in assessments selected for the general sample. This included being rated as understanding and speaking English fluently and not having severe or profound disabilities. In addition, for a PK child to be enrolled in the sample, it was expected that they would matriculate to K the following year. In instances when consent was received for the sibling of a selected child, they were excluded from the sample.

Of the 915 participating children, 420 were in PK, 124 were in K, 125 were in G1, 123 were in G2, and 123 were in G3. The PK cohort was purposefully designed to be the largest, as it would be followed for five years and attrition was expected. Children's ages ranged from 4 years, 6 months to 8 years 6 months. The sample included slightly more boys than girls (53% versus 47%) and the majority of children were White/Caucasian. Specifically, 85% of children were White/Caucasian, 8% were Black, 5% were Asian, and 2% were other. About 10% of children were Hispanic. Seventy-eight percent of families reported speaking primarily English at

home; other languages spoken at home included Spanish, Chinese, Amharic, and Vietnamese. Seventy percent of children resided in two-parent households. Nearly 10% of children had Individual Education Plans (IEPs), and about 16% qualified for free/reduced lunch.

Procedure

The primary procedure for this study, beyond recruiting the longitudinal cohorts of children, was that of assessment. Children completed a comprehensive assessment battery in the latter half of the academic year (January through May). The battery required an average of 5 to 6 hours to complete, with measures administered in 15- to 40-minute blocks. For the measures utilized in the present study, all children were assessed individually by trained assessors in quiet locations at their schools, a university laboratory, or other public facility (e.g., library). Assessors were certified for a given measure following completion of an extensive standardized training program that included, for instance, completion of a written quiz concerning administration and scoring procedures (required 100% correct), and completion of two live administrations that were observed by an experienced assessor (90% accuracy to administration and scoring procedures based on rubrics developed for this purpose). Measures that were deemed to be too complex to be scored reliably in the field were audio-recorded using digital recorders and scored by trained research staff subsequently in a lab setting.

Measures

The present study utilized a subset of measures from the larger assessment battery that represent children's language ability with regard to vocabulary, grammar, and discourse skills. Table 1 provides a list of measures administered by grade, as well as information on which measures were audio recorded and post-scored for scoring accuracy. Inter-rater reliability was acceptable for all post-scored measures, with ICC's ranging from .86 to .99 across measures and

grades. Missing data percentages ranges from 0 – 30%. The average missing data percentage across all grades and all measures was 7.96%.

Vocabulary Three standardized measures of vocabulary were administered. Among the vocabulary measures utilized across all grades, all sample-specific reliabilities exceeded 0.70 with the exception of the CELF Word Classes subtest for the G3 sample (with $\alpha = 0.69$). . Descriptive statistics are provided in Table 1.

Vocabulary: Peabody Picture Vocabulary Test (PPVT). The Peabody Picture Vocabulary Test – IV (PPVT; Dunn & Dunn, 2007), specifically Form A, assessed children's comprehension of single words. The assessor asked the child to point to the picture, out of four, that matched a verbally-presented stimulus word. The total raw score was utilized in analyses.

Vocabulary: Expressive Vocabulary Test (EVT). The Expressive Vocabulary Test: 2nd edition (EVT-2; Williams, 1997-2007), Form A, assessed children's productive vocabularies. The assessor asked the child to provide a single word label for a picture or to provide a single word synonym for a target word. The total raw score was utilized in analyses.

Vocabulary: Word Classes Subtest (CFWCr and CFWCe). The Word Classes subtests of the Clinical Evaluation of Language Fundamentals, 4th Edition (CELF; Semel, Wiig, & Secord, 2003) assessed children's abilities to understand relationships among words related by semantic class features and produce these similarities and differences verbally. Word Classes 1 was administered to children in PK through G2, and Word Classes 2 was administered to children in G3. To administer this task, the assessor read a set of words, which were also represented pictorially for the younger children assessed with Word Classes 1. To assess comprehension of words, the child was asked to select the two words that went together. For the productive portion, the child was asked to verbalize how the two words were related. The total

number of correct responses on both portions was tallied to compute the raw score utilized in analyses.

Grammar. Six standardized measures of grammar were administered. Different measures were appropriate for different age ranges; thus, not all grade cohorts received the same measures, as shown in Table 1. Among the grammar measures utilized across all grades, all sample-specific reliabilities exceeded 0.70 with the exception of the CELF Word Structure subtest for the G3 sample (with $\alpha = 0.63$).

Grammar: Word Structure Subtest (CFWS). The Word Structure subtest of the CELF (Semel et al., 2003) assessed children's abilities to apply word structure rules to indicate inflections, derivations, and comparison and to select appropriate pronouns to refer to people, objects, and possessive relationships. The child listened to a model sentence and then supplied a missing word, using appropriate inflectional morphology, in a second sentence that mirrored the first. All children completed this subtest, although a stop rule of eight incorrect responses was utilized for PK children. The total number of correct responses was tallied as the raw score.

Grammar: Recalling Sentences Subtest (CFRS). The Recalling Sentences subtest of the CELF (Semel et al., 2003) assessed children's abilities to listen to spoken sentences of increasing length and complexity and repeat these sentences without changing word meanings, inflections, derivations or comparisons, or sentence structure. All children completed this subtest, although two items were added to better accommodate PK children. Specifically, the first two items from the Recalling Sentences subtest of the CELF: Preschool, 2nd edition (Wiig, Secord, & Semel, 2004) were administered as the first two test items, followed by the test items on the Recalling Sentences subtest of the CELF in the designated order. For each item, the child was asked to

listen to a sentence read by the assessor and to repeat it verbatim. Total points were summed to create the raw scores utilized in analyses.

Grammar: Past Tense Probe (TEGT). The Past Tense probe of the Rice/Wexler Test of Early Grammatical Impairment (TEGI; Rice & Wexler, 2001) assessed children's production of regular and irregular past tense verbs, and was administered to PK and K children only. The child was presented with a picture and verbal description of a boy or girl completing an action using the present participle verb form. The child was then asked to respond as to what the boy or girl did, requiring use of the past tense form of the verb. The total raw score was utilized in analyses.

Grammar: Third Person Singular Probe (TEGS). The Third Person Singular probe of the TEGI (Rice & Wexler, 2001) assessed children's abilities to produce /-s/ or /-z/ in present tense verb forms with singular subjects. This probe was administered to PK and K children only. The child was presented with a picture and name of a person with a specific job (e.g., teacher) and asked what that person does (e.g., teaches). The total raw score was utilized in analyses.

Grammar: Test for Reception of Grammar (TRG). The Test for Reception of Grammar – Version 2 (TROG-2; D. Bishop, 2003) assessed children's comprehension of English grammatical contrasts marked by inflections, function words, and word order. Test items are arranged in 20 blocks of four, each block assessing knowledge of the same grammatical contrast. For each item, the child was shown four pictures and the assessor read the accompanying sentence and asked the child to point to the picture for that sentence. For each sentence, one picture represented the target meaning and the other three pictures represented grammatical or lexical foils. The child was awarded one point if all item responses within a block were correct; the total number of correct blocks was used for analyses.

Grammar: Morphological Derivation Task (MDR). A morphological derivation task described by Wagner and colleagues (Wagner, n.d.) assessed children's knowledge of derivational morphology. Only children in G1- 3 completed this measure. The assessor presented the child with a base word (e.g., farm) and an incomplete sentence for which he or she provided a derived form of the base word (e.g., My uncle is a _____). Responses were tallied to provide the raw score utilized in analyses.

Discourse. Five measures of discourse skills were administered to examine comprehension monitoring, text structure knowledge (specific to narratives), and inferencing. Measures given differed slightly by grade, as shown in Table 1. We provide reliability data for these researcher-designed measures.

Discourse: Comprehension Monitoring - Knowledge Violations Task (KVT). A researcher-developed measure based on work of Cain and Oakhill (Cain & Oakhill, 2006; Oakhill & Cain, 2012) was used to assess comprehension monitoring in PK and K children. The child listened to short stories that were either entirely consistent or included inconsistent information, as in a story about a rabbit in a vegetable patch who "especially liked the ice cream that grew there." After each story, the child was asked whether the story made sense and, if not, what was wrong with the story. The task included five practice stories and seven test stories, five of which included inconsistent information. Children received one point for each inconsistent story for which they correctly answered the sense question and correctly identified the incorrect information in the story. These points were tallied to create the raw score utilized in analyses (possible range: 0 to 10). Internal consistency (Cronbach's alpha) in the present sample ranged from 0.80 to 0.81 across grades.

Discourse: Comprehension Monitoring - Detecting Inconsistencies Task (DI). A researcher-developed measure based on the work of Cain and Oakhill (Cain & Oakhill, 2006; Oakhill & Cain, 2011) was used to assess comprehension monitoring for children in grades G1 to G3. The tasks included five practice stories and twelve test stories that were either entirely consistent or included inconsistent information; eight test stories were inconsistent. Similar to the Knowledge Violations Task, the child listened to short stories and was asked whether the story makes sense and, if not, what was wrong with the story. The task was more difficult than the Knowledge Violations Task, because identification of inconsistent information required integration of two elements across two separate sentences (e.g., On her way she lost her purse. When she got to the store, she took out her purse and bought her favorite candy). Children received a point for each inconsistent story for which they correctly identified the incorrect information within the story. These points were tallied to create the raw score utilized in analyses (possible range: 0 to 8). Internal consistency (Cronbach's alpha) in the present sample was 0.70 for G1, 0.73 for G2 and 0.54 for G3.

Discourse: Narrative Structure - Picture Arrangement Task (PAT). The first 13 items of the Picture Arrangement Test from the Wechsler Intelligence Scale for Children (WISC-III; Wechsler, 1992) were adapted to create a measure that assessed children's text structure knowledge, specific to ordering narrative events into a causally- and temporally-coherent sequence. This task was administered to children in grades PK through G2. The child was told that she would view some pictures that tell a story, but the story is out of order. The assessor then showed the child a set of three to five picture cards in a fixed order and read a sentence that described each. The child was asked to rearrange the pictures to put them in the correct sequence. One practice item and 12 test items were administered; a ceiling rule of five incorrect stories was

applied. Internal consistency (Cronbach's alpha) in the present sample ranged from 0.76 to 0.88 across grades.

Discourse: Narrative Structure - Sentence Arrangement Task (SAT). A researcher-developed measure based on work by Oakhill and Cain (2011) and Stein and Glenn (1982) assessed older children's text structure knowledge, specific to ordering narrative events into a causally- and temporally-coherent sequence. This measure was administered to children in G2 and G3. The child was told that she would read some sentences that tell a story, but the story is out of order. The assessor then showed the child a set of 6 to 12 cards, with one sentence typed on each card, in a fixed order and read each sentence aloud to the child. The child was asked to rearrange the sentences to put them in the correct sequence. This measure consists of 1 practice story and 4 test stories. Internal consistency (Cronbach's alpha) in the present sample was 0.67 to 0.69.

Discourse: Inferencing (InfBK and InfInt). A researcher-developed measure based on work by Cain and Oakhill (1999) and Oakhill and Cain (2011) was used to assess children's ability to generate two types of inferences from short narrative texts: inferences that require *integration* of two premises, and inferences that require integration of information in the text with background knowledge to fill in missing details. Following administration of a practice story, children listened to two stories, after which the assessor asked eight questions, reflecting four questions per inference type. The total number of correct responses for the two types of inference were summed and averaged to produce two scores, one for background knowledge and one for integration (possible range: 0 to 2 for each). Internal consistency (Cronbach's alpha) for the averaged score across all items in the present sample ranged from 0.64 to 0.78 across grades.

Results

Means and standard deviations for the primary study measures are shown in Table 1. As expected, mean scores tended to increase across grades. Through inspection of skewness and kurtosis criteria, histograms, and boxplots of the data, the majority of variables showed adequate distribution with no severe departures from normality. Further, no extreme outliers were identified within the data across any of the grades.

Two interests were addressed in the primary analyses. The first interest was to examine dimensionality of language ability with respect to vocabulary, grammar and discourse skills as a unitary construct or as two distinct dimensions consisting of lower-level language (vocabulary and grammar) skills versus discourse skills. The second interest was to examine if discourse, vocabulary, and grammar skills represent three distinct dimensions of language ability. Measures indicating each of the three language dimensions assessed are identified in Table 1. Both interests were addressed simultaneously by examining three models to determine the best conceptualization of language dimensionality across the grades PK to G3. Specifically, we fitted a taxonomy of latent-variable models allowing for competing configurations of dimensionality. These confirmatory factor analyses (CFA) were conducted in MPlus v7.11 (Muthén & Muthén, 1998-2012) using the clustering option and maximum likelihood estimation with robust standard errors (MLR) to adjust for non-independence within classrooms and slight non-normality of the data. Three models were compared at each grade for quality of fit: a uni-dimensional model of language; a two-dimensional model differentiating a lower-level language factor (vocabulary and grammar) from a factor representing discourse; and a three-dimensional model separating vocabulary, grammar, and discourse ability. Each lower-dimension model can be obtained from the higher-dimension models by fixing the corresponding factor correlations to 1.0; as such,

these three models form a set of nested models and relative model fit can be assessed through the Chi-square difference test. In each confirmatory model we allowed for methods effects for measures from the same instrument (e.g. both the comprehension and production parts of the Word Classes, and also for the One-Word Picture Vocabulary Tests), but no other modifications were made. A graphical depiction of these confirmatory models for the PK data can be found in Figure 1; similar models were constructed for corresponding measures in K and each of G1 through G3.

Recent recommendations on model fit and assessment of model quality have called for a consideration of several fit criteria rather than a strict reliance on any single criteria, and this is the strategy through which we evaluated our models (e.g., Hancock & Mueller, 2010, 2013; Hu & Bentler, 1995, 1999; Kline, 2013; Lomax, 2013; McCoach, Black & O'Connell, 2007; Moran, Marsh & Nagengast, 2013; Tomarken & Waller, 2003, 2005). Correspondingly, we used several approaches in our assessment of language dimensionality for each grade. First, we considered a combination of absolute, parsimony correction, and comparative indices of model fit (Brown, 2006; Byrne, 2012). These included the root mean square error of approximation (RMSEA), which ranges from 0 to 1; values less than .08 or, more conservatively, .05, suggest acceptable model fit (Browne & Cudeck, 1993; MacCallum & Austin, 2000). We report 90% confidence intervals for the RMSEA and results of the closeness of fit test (Browne & Cudek, 1993) which tests the null hypothesis that RMSEA is less than or equal to .05; this test should be n.s.). We also used the comparative fit index (CFI), an incremental measure that compares the fit of the theoretical model to a null model which assumes the indicator variables in the model are uncorrelated (i.e., an independence model). CFI values greater than .90 indicate a good fitting model (Hu & Bentler, 1999; Lomax, 2013; Schumacker & Lomax, 2004). Comparison of nested

models based on the CFI can provide reasonable support for the more parsimonious model if the incremental change in CFI is less than .01 (Moran, Marsh & Nagengast, 2013). The standardized root mean square residual (SRMR) is an absolute measure of model fit and is the standardized difference between the observed and predicted correlations; a model that fits well to the data should have an SRMR less than .05 (Byrne, 2012). Model fit was also evaluated using Akaike's Information Criterion (AIC), a log-likelihood measure of fit useful for comparing competing (typically non-nested) models and for which smaller AICs indicate better fit (Kline, 2005).

In addition to examination of the model fit indices, we used the Satorra-Bentler rescaled Chi-square difference test for statistically comparing nested models (S-B χ^2, Satorra & Bentler, 1994). If the scaled Chi-square difference is not statistically significant, the more parsimonious (i.e., the nested or more restrictive) model is retained as having model fit no worse than the more complex model. If the difference is statistically significant, we are not able to retain the more parsimonious model.

In evaluating model quality, we note that descriptive fit statistics as well as results of statistical comparisons of nested models are susceptible to sample size and other design features (McCoach, Black, & O'Connell, 2007; Tomarken & Waller, 2005). Thus, as advocated in the literature (e.g., Lomax, 2013; Mueller & Hancock, 2010), we viewed the collection of model fit indices as a guide rather than rely on a single indicator or test as indicative of the "best" representation of language. We also considered substantive interpretation of the CFAs, particularly regarding discrimination between latent factors for models with increasing dimensionality.

Accordingly, in our final approach to examining support for increasing dimensionality across grades, we considered the extent to which the latent factors in the multidimensional

models represented distinct components of language. For each competing model by grade, we reviewed the factor correlations. We also calculated the amount of variance in a measure that is explained by a factor by squaring the standardized factor loadings and then averaged these to estimate the Average Variance Extracted (AVE) (Hair, Black, Babin & Anderson, 2010; Netemeyer, Bearden, & Sharma, 2003). Comparing the degree to which measure variance is captured by a factor (AVE) relative to the shared variance between constructs (squared factor correlations; SFCs) for the two- and three-dimensional models provides information about how well the constructs are distinct within each model. If the proportion of variance extracted by a factor is less than the proportion of variance shared between those factors, there is little support for discrimination between the factors (Fornell & Larcker, 1981; Hair et al., 2010; Netemeyer, et al., 2003). Consequently, our decisions on the best dimensional representation of language for each grade were based on the overall constellation of model fit criteria, consideration of discriminability between latent factors, and tests of nested model comparisons.

Model fit statistics and results for the adjusted Chi-square comparison tests across models within each grade are provided in Table 2. Table 3 contains information on the AVE for the set of measures theoretically presumed to relate to each factor in the models, as well as the squared factor correlations (SFCs) for the two- and three-dimensional models. In what follows, we present results for each of the five grades examined in this study. We provide the greatest depth of interpretation for the PK results, as details provided therein can then be extrapolated for the other grades. Construct reliability (Coefficient *H*) for the final model we selected within each grade is also reported (Hancock & Mueller, 2001).

Pre-Kindergarten: Dimensionality of Language Ability

Model fit criteria. The constellation of fit statistics (see Table 2) was very similar across the three models of dimensionality being compared. For example, RMSEA was lowest for the two-dimensional model at .062, but the confidence intervals for all three models were similar, and the RMSEA estimates for the uni-dimensional and three-dimensional models were similar at .065 and .063 respectively Corresponding tests for closeness of fit (i.e., RMSEA $\leq$ .05) were not significant for any of the three models, with *p-close* values of .019, .045, and .036, respectively. Non-statistically significant values for the "closeness" test indicate acceptable fit, although Brown (2006) mentions that there is a tendency among methodologists to prefer a *p-close* value greater than .50 in order to claim that the RMSEA is sufficiently small. None of these models achieved that boundary. Among the three models for PK, the CFI and SRMR values were the same. The work of Moran et al., (2013) suggests that the more parsimonious of these models can reasonably be supported given the lack of incremental change in the CFI. Overall, however, this collection of fit indices did not point to unequivocal support for one model of language dimensionality at PK over another.

Nested model comparisons. Next we examined the (adjusted) S-B χ^2 difference tests for each pair of models (see right-hand section of Table 2, column titled Adjusted $\Delta\chi^{2)}$. This test assesses whether the fit of the simpler (lower-dimensioned) model is appreciably worse than that of the more complex (higher-dimensioned) model. Results indicated that the one-dimensional model provided slightly worse fit relative to the two-dimensional and three-dimensional models, ($\chi^2(1) = 9.13$, $p < .01$ and $\chi^2(3) = 13.34$, $p < .01$, respectively), and there was no statistical difference between the two- and three-dimensional models ($\chi^2(2) = 1.88$, $p > .05$). Thus, statistically speaking, the two-dimensional model would be supported. Given the known problems with distribution and sensitivity to sample size of the chi-square difference test

(Tomarken & Waller, 2003), we next examined the factor correlations and discriminability for the multi-dimensional solutions.

Discrimination between factors. For the two-factor solution, the correlation between the constructs of Lower-Level Language (LLL, consisting of grammar and vocabulary measures) and Discourse skill (D) was very high at .91. Brown (2006) recommends factor correlations lower than .80 or .85 for good discrimination of latent constructs. Thus, discrimination among the two constructs can be characterized as poor, and it is not clear that a two-dimensional representation of language ability provided a *substantive* improvement over that of the uni-dimensional model. We found similar results for the factors in the three-dimensional model representing Grammar (G), Vocabulary (V), and Discourse (D). All factor correlations were very high at .90 or larger for each pair of latent variables, suggesting that discrimination among these constructs was weak for the three-dimensional model as well.

Standardized factor loadings in all three models of language dimensionality (not shown) were statistically significant and the majority of loadings in each CFA were greater than .70. All but one measure had standardized loadings greater than .50; the lowest standardized loading in the PK CFAs was .49 in the uni-dimensional model for the past-tense probe of the TEGI. For each factor, the factor loadings can be squared and summed to find the AVE; when compared to the squared factor correlations (SFCs), the AVE provides a more rigorous assessment of discrimination among factors than simple review of the factor correlations (Fornell & Larcker, 1981; Hair et al., 2010; Netemeyer, et al., 2003).

The top row of Table 3 contains the AVE for the PK data across the factors for the one-, two- and three-dimensional models. For the uni-dimensional model, the AVE for the single factor was 45.4%. In the two-dimensional model, the Lower-Level Language construct is able to

explain an average of 48.9% of the variance in the set of its indicator Grammar and Vocabulary measures, with Discourse explaining an average of 43.2% of the variance in its set of discourse skill indicators. These AVEs are substantially lower than the variance shared between the constructs of LLL and D which was 83.4%, and thus discrimination between factors is not supported (e.g., Netemeyer, et al., 2003). This finding is similar to the results for the three-factor model. All three AVEs for the three-factor Grammar, Vocabulary, and Discourse model are substantially smaller than the shared variance between these factors. The latent variables in the two- and three-dimensional representations of language considered here shared more variance between them than they were able to extract from their corresponding set of measures; thus the latent factors in the multidimensional models are not reasonably distinct.

Summary: PK results. For the three specific models we tested, results of the Chi-square comparison tests supports the two-dimensional model for language in PK. However, we note some ambiguity in claiming that the collection of evidence is strong in support of this model. First, fit criteria were acceptable and nearly equal across all of the models, and the CFI in particular was no different for these models, providing some evidence in favor or the more parsimonious solution (i.e., one-dimensional model) (Moran, et al., 2013). Second, factor correlations are very high for the two-dimensional model, and investigation of the average variance extracted by each factor relative to the squared multiple correlations between factors suggests that meaningful differences between the two dimensions are not readily discerned. We estimated construct reliability for the uni- and two-dimensional models using Coefficient *H* (Hancock & Mueller, 2001). For the uni-dimensional model, reliability was high with $H = .93$. For the two-dimensional model, reliability was still high for the Lower Level Language factor (grammar and vocabulary skills), $H = .91$, but was lower for the Discourse skills factor, $H =$

.782. Although an argument could be made in support of the two-dimensional model based strictly on the scaled Chi-square difference tests, it seems unlikely that the two-dimensional model yields a stronger substantive interpretation to the data over the parsimony available through the one-dimensional model, at least for the specific models that we have compared here at PK. Thus, from a substantive perspective, the uni-dimensional model at PK seems preferable to the two separate but highly correlated factors ($r_{LLL\text{-}D}$ = .91, with r^2 = 83.4%) obtained for the two-dimensional model.

Kindergarten: Dimensionality of Language Ability

Model fit criteria. Examination of model fit criteria showed no clear preference for one representation of language over another (with more consistency across models than the fit results for the PK data: see Table 2).

Nested model comparisons. The uni-dimensional model of language in kindergarten was retained, as none of the adjusted χ^2 difference tests were statistically significant (Table 2, Adjusted $\Delta\chi^2$).

Discrimination between factors. Further evidence supporting uni-dimensionality of language for the Kindergarten data comes from examination of the factor correlations in the multidimensional models, which are all very large (last column of Table 2) and are above the recommendations for good discrimination of latent constructs in a multi-factor model. Standardized factor loadings in all three models of language dimensionality (not shown) were statistically significant and the majority of loadings in each CFA were greater than .70. Three measures had loadings below .50 in the uni-dimensional and two-dimensional models, and only two in the three-dimensional model. As shown in Table 3, the AVE's for the latent factors in the two- and three-dimensional models are less than the large proportion of variance each factor

shares with the other factor(s) in that model (SFC's); thus, in Kindergarten there is no evidence or support for distinct constructs of Lower-Level Language versus Discourse skill, or for Grammar versus Vocabulary versus Discourse ability.

Summary: K results. Findings suggest that the best representation of dimensionality of language ability at kindergarten is the uni-dimensional model. Construct reliability was high, with $H = .94$.

Grade One: Dimensionality of Language Ability

Model fit criteria. Examination of model fit criteria suggests that we can discount the three-factor model due to inadmissible values for one of the factor correlations, with a correlation greater than 1.0 between the latent constructs for Grammar and Vocabulary ($r_{G\text{-}V} = 1.06$). This is not an atypical occurrence in CFA and suggests that model modifications are warranted. Given our specific series of theoretical and confirmatory models to be compared, we did not consider adaptations to the three-dimensional model. The two-dimensional model fit criteria were slightly improved over the one-dimensional model (Table 2, third section). For example, the Chi-square value was smaller, the *p-close* was larger, the CFI was larger, and the AIC was smaller for the two-dimensional model. SRMR was similar for the uni- and two-dimensional models. None of the models attain suggested cutoffs for the RMSEA. Despite larger values on the RMSEA for all three values, the remaining criteria suggest acceptable model fit for the two-dimensional model: Lower-Level Language skill versus Discourse ability.

Nested model comparisons. A statistical comparison of the uni- versus two-dimensional models of language (Table 2) found that the uni-dimensional model fit significantly worse that the two-dimensional model, $\chi^2(1) = 17.82$, $p < .01$. Thus, there is support for the two-factor solution.

Discrimination between factors. The correlation between the Lower-Level Language and Discourse ability factors in the two-factor solution is quite high, $r_{LLL\text{-}D} = .85$, calling into question the distinctiveness of these two factors. However, Brown (2006) notes that .85 (or more conservatively, .80) serves as an appropriate upper bound for discriminability among factors; thus, although high, separate contributions to language by these two factors are likely to be substantively meaningful. To further understand this overlap, we reviewed the discriminant validity properties for the two-dimensional model. Standardized factor loadings in all three models of language dimensionality (not shown) were statistically significant and the majority of loadings in each CFA were greater than .70. Two measures had loadings below .50 in the uni-dimensional model, and only one was below this criteria in the two-dimensional and the three-dimensional models. The AVE results (Table 3) indicate that the variance extracted for each of the two factors in the two-dimensional model are still lower than the variance shared between the two factors. Although we would like to see the AVEs be larger than the SFCs for distinct constructs of Lower-Level Language versus Discourse skills, in G1 relative to K and PK, the amount of shared variance (SFCs) in the two-dimensional models is decreasing, and the amount of variance extracted by each factor (AVEs) in these models is somewhat increasing.

Summary: G1 results. Considering the model fit criteria and the chi-square difference test, the two-dimensional model of language ability at G1 is preferred. Construct reliabilities are moderate to high; for LLL, $H = .92$ and for D, $H = .76$. There is considerable overlap between the constructs of Lower-Level Language and Discourse skill ($r^2 = 72.3\%$) but despite this overlap, we begin to see an emergence of language dimensionality specific to LLL and D skills at first grade.

Grade Two: Dimensionality of Language Ability

Model fit criteria. Similar to the G1 results, the three-factor model can be discounted due to inadmissible values for one of the factor correlations ($r_{G\text{-}V}$ = 1.04 between the latent factors for Grammar and Vocabulary in the three-dimensional solution). Comparing across the uni-dimensional and two-dimensional results (Table 2), the model fit criteria were better for the two-dimensional model. The upper CI boundary on the RMSEA was slightly beyond .08, but all other criteria were within acceptable limits. Thus, based strictly on these criteria, there is support for the two-factor solution.

Nested model comparisons. A statistical comparison of the one- versus two-dimensional models of language (Table 2, far-right column) yielded preference for the two-factor solution, $\chi^2(1) = 17.97$, $p < .01$.

Discrimination between factors. As we saw with the G1 sample, the correlation between the two factors of Lower-Level Language and Discourse skill was high, $r_{LLL\text{-}D}$ = .80, but within acceptable range for discriminability between the factors. Standardized factor loadings in all three models of language dimensionality (not shown) were statistically significant and the majority of loadings in each CFA were greater than .70. Two measures had loadings below .50 in the uni-dimensional and two-dimensional model, and only one was below this criteria in the three-dimensional model. The AVE results (Table 3) indicate that the AVEs for each of the two factors in the two-dimensional model (51.6% and 43.7%, respectively) were lower than the variance shared between the two factors (64.5%). This discrepancy was not what we would desire for distinct constructs of Lower-Level versus Discourse language skills, but there was continuing evidence across grades that the amount of shared variance (SFCs) in the two-dimensional models was gradually decreasing, and the strength of these language skills as independent dimensions of language (AVE) was improving.

Summary: G2 results. Based on the model fit criteria and the chi-square difference test, the two-dimensional model best represents language ability at G2. Construct reliabilities were moderate to high; for LLL, $H = .92$ and for D, $H = .77$. We note again the considerable overlap between the constructs of LLL and D skills ($r^2 = 64.5\%$) at second grade, but continue to detect evidence for the emergence of differentiation between these two constructs.

Grade Three: Dimensionality of Language Ability

Model fit criteria. The final section of Table 2 contains model fit statistics for the three confirmatory factor models for the G3 sample. The three-dimensional solution is interpretable here, with admissible parameter estimates for the factor correlations. All of the model fit criteria were best for the three-dimensional solution; we note that the upper bound on the RMSEA confidence interval was, however, outside the recommended .08 cutoff.

Nested model comparisons. The two-dimensional and three-dimensional models yielded better fit compared to the one-dimensional model ($\chi^2(1) = 7.51$, $p < .01$ and $\chi^2(3) = 12.87$, $p < .01$, respectively: see Table 2), but there was also a statistical difference between the two- and three-dimensional models ($\chi^2(2) = 6.19$, $p < .05$). Based on these results, the representation of language in the three-dimensional model that separates Grammar, Vocabulary and Discourse language skills fit the data significantly better than either lower-order solution.

Discrimination between factors. Factor correlations, however, were high in both the two- and three-dimensional solutions, suggesting that while there may be a statistical preference for a three-dimensional model that isolates the language skills of grammar, vocabulary and discourse skills, there were also relatively strong correlations between these constructs. In particular, the strongest correlation in this set was between the constructs of Grammar and Vocabulary in the three-dimensional model ($r_{G\text{-}V} = .90$), formerly part of the same construct in

the two-dimensional models supported in G1 and G2. Standardized factor loadings in all three models of language dimensionality (not shown) were statistically significant and the majority of loadings in each CFA were greater than .70. Two measures had loadings below .50 in the uni-dimensional model, and only one was below this criteria in the two-dimensional model and the three-dimensional models. The AVE and SFC results provided in Table 3 indicate that the AVE for the first two of the three factors in the three-dimensional model (56.4%, 62.6%, and 29.8%, respectively) were somewhat close to but not exceeding the variance shared between the three factors; there was also a very large 80.6% estimate of variance shared between the latent variables of Grammar and Vocabulary in the three-dimensional model. Thus, despite statistical preference for the three-dimensional solution, overlap among at least two factors is large. However, in the three-factor composition these three separate components of language skill – namely, Grammar, Vocabulary, and Discourse skills – seem to be emerging as independent dimensions of language in older children. We note, however, that the average variance explained by Discourse in the three-dimensional model is lower here than was obtained in the previous grades, suggesting that one or more of the indicators of Discourse may not adequately represent this construct well in G3.

Summary: G3 results. Based on the model fit criteria and the chi-square difference test, the three-dimensional model was preferred for third grade, with a cautionary note on the overlap particularly between the Grammar and Vocabulary factors in the three-dimensional model. Construct reliabilities were moderate; for Grammar, $H = .85$, for Vocabulary, $H = .87$, and for Discourse, $H = .65$. In third grade, differentiation of Grammar and Vocabulary as separate skills from Discourse ability appeared to occur, although these skills were correlated.

Discussion

This examination of the dimensionality of language ability in 4- to 8-year-old children significantly extends our understanding of language development during this critical period of early childhood in several important ways. First, we found that when discourse skills were added to the study of language skills, there was some ambiguity in the identification of dimensionality among the youngest children in our study (children in pre-kindergarten). By Kindergarten, our data showed that the constructs we were investigating, namely grammar, vocabulary, and discourse formed a single construct. Second, we demonstrate a potentially emergent dimensional structure of language, resulting in the separation of vocabulary and grammar, which hang together, and discourse at first and second grade, followed by differentiation of vocabulary, grammar, and discourse by 8 years of age.

Concerning the first major finding, this work built upon an earlier examination of the dimensionality of language in vocabulary and grammar (see Tomblin & Zhang, 2006). That work indicated that language ability is initially uni-dimensional, but becomes increasingly multi-dimensional as children move across the primary grades and into middle school. Complementing those initial findings, the present work looked intensively at the dimensionality of language ability for children who were 4– to 8-years of age, corresponding to the time period in which children typically enter formal schooling (at pre-kindergarten) and transition through periods of learning to read to reading to learn (third grade). Our results help us to understand more explicitly the nature of children's language ability during these formative years of schooling and reading development. For the youngest children in our sample, language dimensionality was found to be somewhat unclear. Specifically, for children in pre-kindergarten, a two dimensional representation of language emerged, whereas for kindergarten children language was observed as

a single dimension of development. Although our results can be considered more robust than previous work in this area due to the latent approach utilized in analyses, we acknowledge these findings are tentative given the small size of our kindergarten sample. Considering the instability of language observed for the youngest children in our study, future research, particularly work that is longitudinal in nature, is warranted to disentangle our mixed findings regarding the possibility that discourse level skills may in fact be separable from other language constructs at very early ages. For the older children, language ability was best represented as two or even three related dimensions. In fact, by third grade, we see that vocabulary, grammar, and discourse appear to be emerging as distinct skills, and that discourse language skills, such as inference and comprehension monitoring, are a distinguishable dimension of language.

These findings of a unitary dimension for vocabulary and grammar until third grade are not discrepant with those of the earlier work. Although Tomblin and Zhang (2006) preferred the two-factor solution only for children in eighth grade, the CFI values for their two-factor model for the earlier grades all indicated a good fit. Together, these two studies with different samples and test instruments similarly show the emergence of separate constructs of vocabulary and grammar. Thus, although these aspects of language are often grouped together as lower-level language skills, the two studies indicate that this level of description may be relevant only for younger children. Further, this finding fits with the theoretical conceptualizations of vocabulary and grammar as lower-level or foundational skills that support later reading comprehension (Lepola et al., 2012).

The emergent dimensionality seen among young children in their development of language makes it clear that language is a complex construct. Our data support the viewpoint that there are dimensions within this construct that are, to some extent, separable. However, why are

these dimensions not wholly apparent until later in children's development? One explanation for our finding of emergent dimensionality is that a hierarchy of language exists, from words to sentences to discourse (Tomblin & Zhang, 2006). As the dimensions lower down in the hierarchy become consolidated and automatized in use, as occurs early in development, cognitive resources are freed up to support more complex processing, as would be required for comprehension monitoring and inferencing. This may arise because a critical base of vocabulary is needed for grammar and, similarly, a critical base of grammar is needed to understand discourse. As noted in our introduction, different phases of grammar development have been linked to lexical knowledge during the first two years of life (Marchman & Bates, 1994).

We do not believe that this is an adequate explanation because there is evidence for the discourse language skills that we discuss (comprehension monitoring, knowledge of narrative, and inference) in the first 3 to 4 years of life (Skarakis-Doyle & Dempsey, 2008; Tomasello et al., 2007). An alternative explanation as to why this emergent dimensionality may occur at a particular time point, is that language abilities are likely affected by children's literacy experiences within schooling and the increasing demand to apply linguistic resources to higher-level interactions with texts (orally or read) as a major part of reading instruction. To this point, consider that during the course of early language development, the written texts that children experience show gradual increases in the complexity of the morphological structure of words, the syntactic sophistication of sentences, and the number of propositions that have to be integrated across the text as a whole. Children's interactions with these increasingly complex written texts during the first few years of literacy instruction may enhance the structure of language dimensions that we find. Consistent with this argument is that the increasing challenge of texts in the early school years provides one explanation for the identification of a group of poor readers

whose problems only become apparent in fourth grade as these children struggle to comprehend what they read (Catts, Compton, Tomblin, & Bridges, 2012). Similarly, it has been shown that interactions with literacy drive phonemic awareness (Castles & Coltheart, 2004): thus, schooling impacts on the structure and refinement of oral, as well as written language. This noted, the results here do not imply that discourse language skills, such as comprehension monitoring and inferencing, are not observed among young children as we have already discussed; rather, it appears to be that case that around first grade these begin to clearly represent a separable dimension of language ability from lower-level skills.

The second major finding of this work concerns the clear emergence of discourse-level language ability as a specific dimension of language ability among the older children in this sample. At first grade, we found two distinct dimensions of language ability- one representing the lower-level skills of vocabulary and grammar, the other representing discourse-level skills, and this distinctiveness held through third grade, albeit with separation of lower-level skills (vocabulary and grammar) observed for the third graders. This is a unique finding because no previous studies of language dimensionality have included assessment of skills essential to discourse.

Limitations of this work warrant note, and we highlight three that are most salient. The first concerns the measures used to represent discourse skill. Several of these measures have never before been used with children as young as those in this study, and thus further examination of their validity with the younger groups is needed. In addition, reliability measures for three of the 20 measures of discourse used across all grades were found to be below commonly accepted cutoffs. Specifically, the alpha for the experimental discourse measure, Detecting Inconsistencies, was found to be particularly low (Cronbach's alpha = 0.54 in grade 3).

However, in each of the models where discourse skill formed its own dimension, the overall construct reliabilities for the discourse factor as measured through Coefficient *H* (Hancock & Mueller, 2001) were acceptable, at .76 and .77 for grades 1 and 2, respectively, although somewhat lower for grade 3 at .65. While additional measurement work should be undertaken to improve the DI measure, construct reliability is not diminished by addition of a weak indicator to a measurement model that includes a collection of strong indicators (Hancock & Mueller, 2001). Nonetheless, the measures reflecting discourse skill overall could be improved. The second limitation concerns the cross-sectional nature of the data. Prior work examining the dimensionality of language ability has used longitudinal methods rather than the cross-sectional approach used here (Tomblin & Zhang, 2006). A limitation of the cross-sectional design is that it essentially provides only a 'snapshot' of phenomenon at a given point in time. However, given that some of the measures across grades varied, examining the models cross-sectionally is a reasonable approach to understanding dimensionality at each grade. While we may infer that differences seen in the dimensionality of language ability would be borne out with longitudinal assessment, this requires additional empirical investigation.

A third limitation is that different theoretical explanations could be considered. It could be argued that the measures of discourse are simply 'more difficult', perhaps because they place a greater demand on cognitive resources, such as working memory, and that if more challenging vocabulary or grammar assessments were included, the three dimensional split would not arise. There is some evidence for this viewpoint. A recent study of children with SLI found that performance on subtests of standardized oral language assessments did not neatly load onto separate factors of semantics and syntax or comprehension vs production language; instead a

factor structure that reflected the different language processing requirements of the tasks was apparent (Hoffmann, Loeb, Brandel, & Gillam, 2011).

While this work is largely meant to be theoretical and descriptive in nature, designed to advance understanding of the nature of language ability among young children, the results may have implications to practice. First, from an assessment perspective, this study does not support the reliance on subtests designed to assess specific dimension of lower-level language ability (i.e., vocabulary and grammar) for pre-kindergarteners through second graders, as these do not appear to reflect specific, distinguishable abilities. Here, we are echoing a point made by Tomblin and Zhang (2006) for older children. Rather, for children at these ages, their vocabulary and grammar reflect a single underlying trait, and it seems best to rely on omnibus assessments of language skill to explore individual differences among children and, potentially, to identify children who have under-developed language skills warranting attention. Second, with respect to instruction, the study supports the notion that for first- through third-grade children, lower- and discourse-level language skills reflect distinguishable underlying traits. While there is increasing awareness of the importance of discourse-level language skills to reading development, it is not clear that educators often seek to identify children who may have limitations in this dimension of language ability. Research showing that discourse skills appear under-developed in children who have comprehension-specific reading deficits in the later primary grades (e.g., Adlof et al, 2006) argues the importance for monitoring these skills early in development.

Table 1

Descriptive Statistics (M, SD) by Measure and Grade

	PK	K	G1	G2	G3
Lower Level Language (LLL) /Grammar (G)					
CELF-4 Word Structure (CFWS)*	15.59 (5.56)	20.49 (5.22)	24.39 (3.96)	26.23 (3.58)	27.46 (2.90)
CELF-4 Recalling Sentences (CFRS)	32.11 (13.9)	42.63 (13.94)	55.67(13.62)	60.47 (13.72)	65.96 (13.86)
TEGI – Past Tense Probe (TEGT)	8.52 (4.31)	9.31 (4.84)	--	--	--
TEGI – Third P. Sing Probe (TEGS)	6.99 (2.93)	7.56 (2.75)	--	--	--
TROG (TRG)	6.24 (3.76)	9.76 (4.27)	13.86 (3.34)	14.41 (3.58)	15.54 (2.90)
Morphological Derivation (MDR)	--	--	7.89 (4.21)	10.57 (4.56)	13.72 (4.04)
Lower Level Language (LLL)/Vocabulary (V)					
PPVT (PPVT)	93.78 (19.31)	111.46 (18.59)	129.15 (16.95)	137.80 (16.68)	150.86 (16.81)
EVT (EVT)	70.06 (13.79)	82.57 (12.11)	96.94 (14.01)	105.75 (13.77)	113.84 (14.28)
CELF-4 Word Classes Receptive (CFWCr)*	14.29 (4.44)	17.55 (2.91)	18.97 (1.93)	19.77 (1.26)	11.31 (3.21)*
CELF-4 Word Classes Expressive (CFWCe)*	7.82 (4.54)	12.61 (4.16)	14.90 (2.70)	16.38 (2.48)	6.44 (2.72)
Discourse (D)					
Knowledge Violations (KVT)	4.86 (2.88)	5.67 (2.86)	--	--	--
Detecting Inconsistencies (DI)	--	--	4.45 (2.18)	4.75 (2.22)	5.50 (1.74)
Picture Arrangement Task (PAT)	2.61 (2.90)	5.90 (3.75)	8.95 (2.64)	--	--
Sentence Arrangement Task (SAT)	--	--	--	0.87 (1.17)	1.12 (1.26)
Inferencing Background Knowledge (InfBK)*	0.76 (0.40)	1.13 (0.40)	1.16 (0.42)	1.20 (0.44)	1.68 (0.29)
Inferencing Integration (InfInt)*	0.87 (0.49)	0.97 (0.45)	1.07 (0.43)	1.13 (0.48)	1.39 (0.39)

Note. Word Classes 1 was administered to children in PK to G2 (scores ranged from 0 to 21 for the receptive subtests and from 0 to 18 for the expressive subtest) and Word Classes 2 was administered to children in G3 (scores ranged from 0 to 19 for the receptive subtests and from 0 to 13 for the expressive subtest). Standard scores from the PPVT and EVT are available from the first author. * = post scored measure.

Table 2

Results of Confirmatory Models by Grade

Model	χ^2	*df*	*p*	MLR Scaling factor	RMSEA, *p*-close (90% CI)	CFI	SRMR	AIC	Adjusted $\Delta\chi^2$	*r* for latent variables
Pre-Kindergarten (*n* = 420)										
Uni-dimensional	170.84	62	<.001	1.02	.065, *p* = .019 (.053, .076)	.96	.04	26387.03		---
Two-dimensional	159.70	61	<.001	1.02	.062, *p* = .045 (.050, .074)	.96	.04	26376.68	9.13[a] **	$r_{LLL\text{-}D}$ = .91
Three-dimensional	157.40	59	<.001	1.02	.063, *p* = .036 (.051, .075)	.96	.04	26379.11	13.34[b] ** 1.88[c]	$r_{G\text{-}D}$ = .90 $r_{V\text{-}D}$ = .94 $r_{G\text{-}V}$ = .99
Kindergarten (*n* = 124)										
Uni-dimensional	60.53	62	.529	1.01	.000, *p* = .939 (.000, .052)	1.00	.04	7727.97		---
Two-dimensional	59.73	61	.522	1.00	.000, *p* = .935 (.000, .053)	1.00	.04	7728.24	0.89[a]	$r_{LLL\text{-}D}$ = .95*
Three-dimensional	56.58	59	.565	1.00	.000, *p* = .944 (.000, .051)	1.00	.04	7729.30	3.75[b] 3.29[c]	$r_{G\text{-}D}$ =.92* $r_{V\text{-}D}$ = .98* $r_{G\text{-}V}$ = .98*
Grade 1 (*n* = 125)										
Uni-dimensional	85.25	52	.002	0.99	.072, *p* = .101 (.043, .098)	.96	.05	6844.67		---
Two-dimensional	71.86	51	.029	1.01	.057, *p* = .332 (.018, .086)	.98	.05	6834.08	17.82[a] **	$r_{LLL\text{-}D}$ = .85*
Three-dimensional	68.03	49	.037	1.00	.056, *p* = .362 (.014, .086)	.98	.05	6833.88	18.43[b] ** 3.83[c]	$r_{G\text{-}D}$ = .82* $r_{V\text{-}D}$ = .95* $r_{G\text{-}V}$ = 1.06*[d]

Grade 2 (*n* = 123)										
Uni-dimensional	87.66	52	.001	0.90	.080, p = .073 (.046, .101)	.95	.06	6489.01		---
Two-dimensional	68.75	51	.049	0.90	.053, p = .412 (.003, .083)	.98	.05	6473.85	17.97[a] **	r_{LLL-D} = .80*
Three-dimensional	67.21	49	.043	0.90	.055, p = .379 (.010, .085)	.98	.05	6476.52	20.45[b] ** 1.51[c]	r_{G-D} = .79* r_{V-D} = .86* r_{G-V} = 1.03* [d]
Grade 3 (*n* = 122)										
Uni-dimensional	87.74	52	.001	1.02	.075, p = .071 (.047, .102)	.94	.06	6484.43		---
Two-dimensional	79.33	51	.010	1.02	.067, p = .158 (.036, .095)	.96	.06	6477.61	7.51[a] **	r_{LLL-D} = .78*
Three-dimensional	71.33	49	.020	0.99	.061, p = .265 (.025, .090)	.97	.05	6471.82	12.87[b] ** 6.19[c] *	r_{G-D} = .80* r_{V-D} = .70* r_{G-V} = .90*

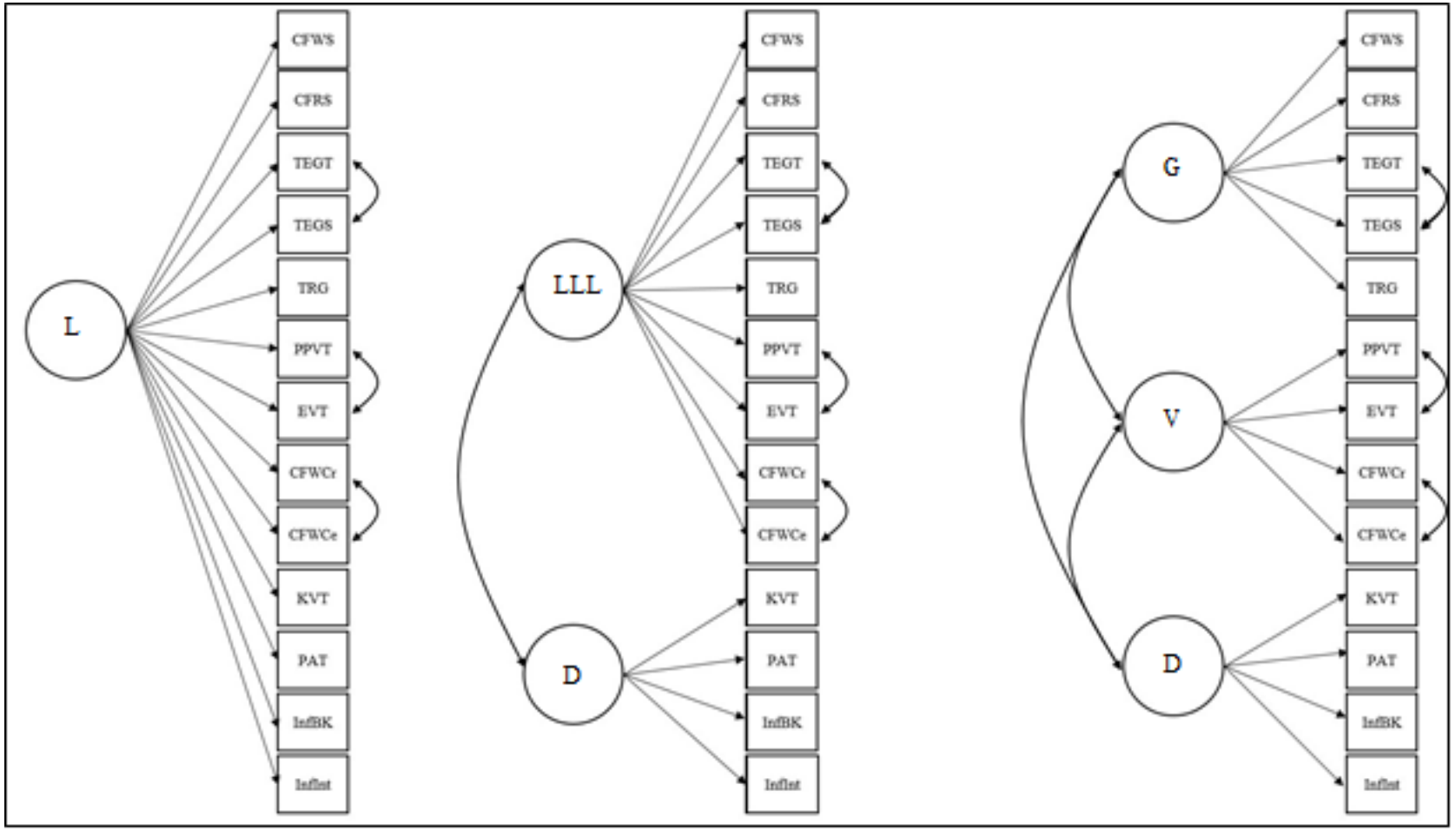
Note. ***p<.001, **p<.01, *p<.05; [a] uni-dimensional vs. two-dimensions (df = 1); [b] uni-dimensional vs. three-dimensions (df = 3); [c]two-dimensions vs. three dimensions (df = 2). [d] inadmissible parameter estimate. MLR = Robust Maximum Likelihood; RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; SRMR = Standardized Root Mean Square Residual; AIC = Akaike Information Criteria

Table 3

Grade-specific Average Variance Extracted (AVE) by Each Factor, and Squared Factor Correlations (SFC's) for the One-, Two-, and Three-Dimensional Models of Language.

Grade	1-Dimensional	2-Dimensional			3-Dimensional					
	AVE	AVE		SFC (%)	AVE			SFC (%)	SFC (%)	SFC (%)
		LLL	D	LLL v D	G	V	D	G v D	V v D	G v V
PK	45.4	48.9	43.2	83.4	47.4	50.7	43.2	80.6	87.6	99.6
K	48.7	52.6	43.6	90.6	52.4	54.6	43.7	83.9	95.5	95.3
G1	46.8	53.5	43.0	72.3	61.6	41.0	43.0	67.7	90.6	> 100.0*
G2	44.6	51.6	43.7	64.5	57.9	43.2	43.8	61.6	73.6	> 100.0*
G3	43.2	55.3	29.8	61.3	56.4	62.6	29.8	65.3	48.9	80.6

Note. * = Inadmissible solution, LLL=Lower Level Language, D = Discourse, G=Grammar, V=Vocabulary.



Note. The figure is an exemplar of confirmatory models, as some measures change across grade (see Table 1 for list of measures).

Figure 1.

Graphical depiction of confirmatory models for the PK data (L=Language, LLL=Lower Level Language, D=Discourse, G=Grammar, V=Vocabulary)

References

Adlof, S., Catts, H., & Little, T. (2006). Should the simple view of reading include a fluency component? *Reading and Writing, 19*(9), 933-958. doi: 10.1007/s11145-006-9024-z

American Psychiatric Association (2012). *Diagnostic and statistical manual of mental disorders.* American Psychiatric Association.

Bates, E., Dale, P. S., & Thal, D. (1995). Individual differences and their implications for theories of language development. *The handbook of child language*, 96-151.

Berninger, V. W., & Abbott, R. D. (2010). Listening comprehension, oral expression, reading comprehension, and written expression: Related yet unique language systems in grades 1, 3, 5, and 7. *Journal of educational psychology, 102*, 635-651.

Brown, T.A. & Moore, M.T. (2014). Confirmatory Factor Analysis. In Rick H. Hoyle (Ed.), Handbook of Structural Equation Modeling (pp. 361-379). New York, NY: Guilford Press.

Bishop, D. V. M. (1997). *Uncommon understanding: Development and disorders of language comprehension in children*. Hove, England: Psychology Press.

Bishop, D. V. (2003). *Test for Reception of Grammar* (2nd ed.). Minneapolis, MN: Pearson.

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

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

Byrne, B. M. (2012). *Structural equation modeling with Mplus*. New York, NY: Routledge.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension bility: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96*(1), 31-42. doi: 10.1037/0022-0663.96.1.31

Cain, K., & Oakhill, J. (2006). Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology, 76*, 683-696.

Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the onference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology, 96*(4), 671-681. doi: 10.1037/0022-0663.96.4.671

Castles, A., & Coltheart, M. (2004). Is there a causal link from phonological awareness to success in learning to read? *Cognition, 91*, 77-111.

Catts, H. W., Compton, D., Tomblin, J. B., & Bridges, M. S. (2012). Prevalence and nature of late-emerging poor readers. *Journal of Educational Psychology, 104*(1), 166-181. doi: 10.1037/a0025323

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (1999). Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading, 3*, 331-361.

Catts, H. W., Hogan, T. P., & Fey, M. E. (2003). Subgrouping poor readers on the basis of individual differences in reading-related abilities. *Journal of Learning Disabilities, 36*, 151-164. doi: 10.1177/002221940303600208

Dickinson, D., Golinkoff, R., & Hirsh-Pasek, K. (2010). Speaking out for language: Why language is central to reading development. *Educational Researcher, 39*, 305-310

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test* (4th ed.). Minneapolis, MN: Pearson.

Elsabbagh M., & Karmiloff-Smith, A. (2004). Modularity of mind and language. *The Encyclopaedia of Language and Linguistics*, 218-24.

Finestack, L. H., Sterling, A. M., & Abbeduto, L. (2013). Discriminating Down syndrome and fragile X syndrome based on language ability. *Journal of child language, 40*(1), 244.

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

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education 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), 371-395. doi: 10.1037/0033-295x.101.3.371

Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. E. (20 1 0). Multivariate data analysis: A global perspective. (7th ed.). Upper Saddle River, NJ: Pearson

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. *Structural equation modeling: Present and future*, 195-216.

Hancock, GR & Mueller (2010). Structural equation modeling. In Gregory R. Hancock & Ralph O. Mueller (Eds.), The Reviewer's Guide to Quantitative Methods in the Social Sciences (pp. 371-384). New York, NY: Routledge.

Hancock, GR & Mueller (Eds.) (2013). Structural Equation Modeling: A Second Course (2nd ed.). Charlotte, SC: IAP.

Hoffman, L. M., Loeb, D. F., Brandel, J., & Gillam, R. B. (2011). Concurrent and construct

validity of oral language measures with school-age children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *54*(6), 1597-1608.

Hogan, T. P., Cain, K., & Bridges, M. S. (2012). Young children's oral language abilities and later reading comprehension. In T. Shanahan & C. J. Lonigan (Eds.), *Early Childhood Literacy: The National Early Literacy Panel and Beyond* (pp. 217-232): Brookes Publishing Co

Hu, L.T., & Bentler, P. M. (1995). Evaluating model fit. In Rick H. Hoyle (Ed.), Structural Equation Modeling: Concepts, issues and applications (pp. 76-99). Thousand Oaks, CA: Sage.

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

Kim, Y. S. (in press). Language and cognitive predictors of text comprehension: Evidence from multivariate analysis. *Child Developmen*

Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. Stahl & S. Paris (Eds.), *Children's reading comprehension and assessment* (pp. 71-92). Mahwah, NJ: Lawrence Erlbaum.

Kline, R. (2013). Exploratory and confirmatory factor analysis. In Y. Petscher, C. Schatschneider, & D. Compton (Eds), Applied Quantitative Analysis in Education and the Social Sciences (pp. 171-207). New York, NY: Routledge.

Lepola, J., Lynch, J., Laakkonen, E., Silvén, M., & Niemi, P. (2012). The Role of inference making and other language skills in the development of narrative listening yomprehension in 4–6-year-old children. *Reading Research Quarterly, 47*(3), 259-282. doi: 10.1002/rrq.020

Lomax, R. (2013). Introduction to structural equation modeling. In Y. Petscher, C. Schatschneider & D. Compton (Eds.), *Applied quantitative analysis in education and the social sciences* (pp. 245-264). New York, NY: Routledge.

Lynch, J. S., van den Broek, P., Kremer, K. E., Kendeou, P., White, M. J., & Lorch, E. P. (2008). The development of narrative comprehension and its relation to other early reading skills. *Reading Psychology, 29*(4), 327-365. doi: 10.1080/02702710802165416

MacCallum, R.C., & Austin, J.T. (2000). Applications of structural equation modeling in psychological research. Annual Review of Psychology, 51, 201-226.

Marchman, V. A., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of child language, 21*, 339-339.

Marsh, HW, Hau, K-T, & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and danges in overgeneralizing Hu & Bentler's (1999) findings. Structural Equation Modeling, 11(3), 320-341.

McCoach, D. B., Black, A. C., & O'Connell, A. A. (2007). Errors of inference in structural equation modeling. *Psychology in the Schools, 44*(5), 461-470. doi: 10.1002/pits.20238

Mills, D. L., Coffey-Corina, S., & Neville, H. J. (1997). Language comprehension and cerebral specialization from 13 to 20 months. *Developmental Neuropsychology, 13*(3), 397-445.

Moran, A.J.S., Marsh,H.W., & Nagengast , B. (2013). Exploratory structural equation modeling. In G. R. Hancock & R.O Mueller (Eds.), Structural Equation Modeling: A Second Course (2nd ed.). Charlotte, SC: IAP.

Mueller, R. O. & Hancock, G.R. (2010). Structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 371-383). New York, NY: Routledge.

Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology, 40*(5), 665-681. doi: 10.1037/0012-1649.40.5.665

Muthén, L. K., & Muthén, B. O. (1993-2012). *MPlus User's Guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nation, K., Cocksey, J., Taylor, J. S. H., & Bishop, D. V. M. (2010). A longitudinal investigation of early reading and language skills in children with poor reading comprehension. *Journal of Child Psychology and Psychiatry, 51*, 1031-1039.

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

Oakhill, J. V., & Cain, K. (2012). The precursors of reading ability in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading, 16*(2), 91-121. doi: 10.1080/10888438.2010.529219

Perfetti, C., & Stafura, J. (2013). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18*(1), 22-37. doi: 10.1080/10888438.2013.827687

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

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*(4), 357-383. doi: 10.1080/10888430701530730

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). *The Acquisition of Reading Comprehension Skill* [doi:10.1002/9780470757642.ch13]. Malden: Blackwell Publishing.

Pinker, S. (1998). Words and rules. *Lingua, 106*(1), 219-242.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition, 48*(1), 21-69.

Rice, M. L., & Wexler, K. (2001). *Rice/Wexler Test of Early Grammatical Impairment*. San Antonia, TX: The Psychological Corporation.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. v. E. C. C. Clogg (Ed.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA, US: Sage.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. New York, NY: Psychology Press.

Schwartz, R. G. (2009). *Handbook of Child Language Disorders*. New York: Psychology Press.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals (4th ed.)*. Bloomington, MN: Pearson.

Silva, M. T., & Cain, K. (in press). The relations between lower- and higher-level oral language skills and their role in prediction of early reading comprehension *Journal of Educational Psychology*.

Skarakis-Doyle, E., & Dempsey, L. (2008). The detection and monitoring of comprehension errors by preschool children with and without language impairment. *Journal of Speech, Language, and Hearing Research, 51*, 1227-1243.

Stein, N. L., & Glenn, C. G. (1982). Children's concept of time: The development of a story schema. *The developmental psychology of time*, 255-282.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology, 1*, 31-65.

Tomasello, M., Carpenter, M., & Liszkowski, U. (2007). A new look at infant pointing. *Child development, 78*, 705-722.

Tomblin, J. B., Nippold, M. A., Fey, M. E., & Zhang, X. (2014). The character and course of individual differences in spoken language. *Understanding Individual Differences in Language Development Across the School Years*, 62-93.

Tomblin, J. B., Records, N. L., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 39*(6), 1284-1294. doi: 10.1044/jshr.3906.1284

Tomblin, J. B., & Zhang, X. (2006). The dimensionality of language ability in school-age children. *Journal of Speech, Language, and Hearing Research, 49*(6), 1193-1208.

Wagner, R. K. (n.d.). *Morphological Derivation Task*. Tallahassee, FL: Florida State University.

Wechsler, D., & Kort, W. (2005). *WISC-IIInl: Wechsler intelligence scale for children*. San Diego, CA: Harcourt Assessment.

Wiig, E. H., Secord, W. A., & Semel, E. (2004). *Clinical Evaluation of Language Fundamentals Preschool* (2nd ed.). San Antonio, TX: Harcourt Assessment.

Williams, K. T. (2007). *Expressive vocabulary test* (2nd ed.). Bloomington, MN: Pearson.