
The Dimensionality of Language Ability in School-Age Children

J. Bruce Tomblin
Xuyang Zhang
University of Iowa, Iowa City

Purpose: This study asked if children's performance on language tests reflects different dimensions of language and if this dimensionality changes with development.

Method: Children were given standardized language batteries at kindergarten and at second, fourth, and eighth grades. A revised modified parallel analysis was used to determine the dimensionality of these items at each grade level. A confirmatory factor analysis was also performed on the subtest scores to evaluate alternate models of dimensionality.

Results: The revised modified parallel analysis revealed a single dimension across items with evidence of either test specific or language area specific minor dimensions at different ages. The confirmatory factor analysis tested models involving modality (receptive or expressive) and domain (vocabulary or sentence use) against a single-dimension model. The 2-dimensional model involving domains of vocabulary and sentence use fit the data better than the single-dimensional model; however, the single-dimension model also fit the data well in the lower grades.

Conclusions: Much of the variance in standardized measures of language appears to be attributable to a single common factor or trait. There is a developmental trend during middle childhood for grammatical abilities and vocabulary abilities to become differentiated. These measures do not provide differential information concerning receptive and expressive abilities.

KEY WORDS: assessment, item response theory, language

There is little that those studying language will agree on, including the very notion of what language actually is. Most would agree, however, that it is a complex system of knowledge used, among other things, for conveying ideas to others via conventionalized behaviors. Once we venture past this general statement, we quickly confront numerous controversies concerning the composition of this complex system of knowledge. Traditionally, this system has been viewed as having several different components often arranged in a hierarchical relationship. Drawing on linguistic notions, one set of components has been identified as phonology, syntax, morphology, semantics, and pragmatics. Within the language sciences, there have been long-standing debates as to the extent to which these distinctions actually represent different types of language representations involving different types of learning or acquisition systems or whether most of these systems are built from the same mechanisms and stored in the brain in much the same way. The extent to which these language domains reflect different underlying systems has been the center of theoretical and empirical debates for some time. Pinker (1997, 1998) has proposed that the distinction between vocabulary and grammar is based on different systems. Specifically, language is composed of a mental dictionary of memorized words and a mental grammar of rules used to create novel forms. However, Bates and Goodman (2001)

asserted that the case for a modular distinction between grammar and the lexicon in language development has been overstated. The evidence they reviewed supports a unified lexicalist view.

Extending from this theoretical uncertainty concerning the dimensionality of language are the practical issues of language measurement and the characterization of individual differences in language, especially those that make up clinically significant deficits in language development—those that would count as language impairment. The content of most language assessment batteries for children with language impairment, as well as the manner in which language impairment is characterized, can be found in many current textbooks on language disorders of children. In nearly all cases, language is characterized as a multidimensional system consisting of the intersection of different linguistic domains and different modalities. For example, Paul (2001) described an intralinguistic framework for language assessment based on Miller's (1981) previous work in which children were profiled according to the domains of receptive syntax and semantics and expressive phonology, syntax, semantics, and pragmatics. A very similar scheme was proposed by Tomblin, Records, and Zhang (1996) for the purposes of diagnosing specific language impairment. In this case, the diagnosis was determined by measures in each of the three domains (e.g., vocabulary, grammar, and narration) and in each of the two modalities of listening and speaking as measured across the three language domains. This prevailing approach to language assessment assumes that these domains represent independent or partially independent aspects of language development such that children can have deficits in one domain and remain relatively strong in another domain. Likewise, most commercial test batteries provide subtests reflecting, presumably, different language ability domains. Subtest scores are provided, and users are often encouraged to use discrepancies among these subtest scores for diagnosis.

Within modern psychological measurement theory, these subtest scores may be viewed as potentially distinctive psychological traits, thus reflecting a view of language that is multidimensional and componential. It is also common, however, for these subtests to be combined into composite scores, suggesting that a unitary language ability contributes to performance on all tasks involving listening and speaking. Recently, de Villers (2003) questioned the validity of this practice of aggregating test items without sufficient justification from linguistic theory. A similar viewpoint can be found in the psychometric literature as well. Crocker and Algina (1986) noted that "measurement, even though it is based on observable responses, would have little meaning or usefulness unless it could be interpreted in light of the underlying theoretical construct" (p. 7). Thus, the

manner in which measurement items are aggregated requires both theoretical and empirical groundwork. In this way, appropriate measurement of language requires that the instruments that are used be empirically tested against theoretical models to determine what it is that they are measuring.

Despite the widespread practice of assuming multidimensionality of language as measured by traditional clinical measures of language, there is little empirical evidence in support of this. Newcomer and Hammill (1988) acknowledged that the linguistic components on the Test of Oral Language Development—2: Primary (TOLD-2:P) were interrelated, although they were described and evaluated as independent aspects of language. The purpose of treating them as independent domains was to promote understanding and to facilitate the diagnosis of deficits. However, it must be asked whether there is utility in providing distinctions that are not reliable. In this respect, a theoretical distinction used in the test construction does not necessarily result in an empirically justifiable distinction. Thus, a factor analysis of the Clinical Evaluation of Language Fundamentals—3 (CELF-3; Semel, Wiig, & Secord, 1995) showed that the first factor derived from 10 CELF-3 subtests accounted for 53% of the total variance for the age group of 6–8 years old and 49% for the age group of 9 years or older (Semel et al., 1995). The second factor accounted for only 8% of the variance for the younger age group and 10% for the older age group. Technical manuals of language tests such as CELF-3 usually report the results of exploratory factor analyses on subtest scores; however, there have not been studies evaluating the validity of the hypothetical dimensions that these tests are intended to measure. Such testing requires that the hypothesized dimensionality be tested using confirmatory factor analysis, wherein a theoretical model is tested against the latent structure of the manifest measures represented by the test scores. Additionally, current psychometric theory stresses the importance of examining the measures at the item level using item response theory (IRT). Recently, Schatschneider, Francis, Foorman, Fletcher, and Mehta (1999) conducted a modified parallel analysis on 105 phonological awareness item scores (0 or 1) and a confirmatory factor analysis on 6 subtest scores, reporting that phonological awareness appeared to be a unidimensional construct. The results of this study, therefore, supported the validity of a single composite score derived from the subtest scores. A similar type of analysis, however, has not been performed in other domains of language.

The purpose of this study was as follows:

1. To examine the dimensionality of language ability as measured by a set of standardized language measures using modified parallel analysis and confirmatory factor analysis in order to

determine whether language ability as measured on common language tests is unidimensional or multidimensional.

2. To examine whether there are any changes in language dimensionality from kindergarten and second, fourth, and eighth grades.
3. If there is more than one dimension in language ability, how the test items group themselves into these dimensions.

These objectives were specifically motivated by research we are conducting concerning the long-term development of language in typical and atypical language learners. If we were to describe the language growth of these children, it was necessary to determine the extent to which the measures used reflected different dimensions that each needed to be described separately, or one dimension. Thus, as described next, in order to address these questions, we examined existing item level data from several standardized language measures that had been administered to a large group of children spanning the age ranges of 5 years to 14 years of age who were participants in this longitudinal study.

Method

Participants

The participants in this study came from two samples of children. The first and largest sample of children had initially been part of a population sample of 7,218 children who participated in a cross-sectional epidemiologic study (see Tomblin et al., 1997). These children were screened for language impairment using a 40-item test constructed from items from the TOLD-2:P. All children who failed the screen and a similar number of those who passed were then seen for a more comprehensive language evaluation that resulted in 420 children with language disorder and 1,509 children with normal language status. All of these children were screened for normal hearing and had no neurodevelopmental

disorders according to parental report. Following participation in this cross-sectional study, all children who were diagnosed as having language impairment in this study and an equal number of those diagnosed as having normal language status were invited to participate in a longitudinal study that began 2 years after the cross-sectional study. As a result, a longitudinal cohort comprising 604 children who had been members of the cross-sectional study was formed 2 years after these children had been in kindergarten. In addition to the 2-year follow-up postkindergarten, these children were seen again 4 ($n = 570$) and 8 years ($n = 527$) after their initial participation as kindergarteners.

Table 1 provides a description of the participants who contributed data for each of the four test batteries. Contained within this table are mean z scores for the receptive vocabulary and the Block Design subtests from the Wechsler Preschool and Primary Scale of Intelligence—Revised or the Wechsler Intelligence Scale for Children—Third Edition (Wechsler, 1989a, 1989b). These scores show that the samples used for each battery were largely a group of typical children. The kindergarten sample was the largest sample and was very comparable in general ability to age mates. The samples at second, fourth, and eighth grades averaged about one third of a standard deviation below the mean for their age group. In all cases, the samples were representative of the variation in ability levels expected in a general population sample as shown by the standard deviations of these z scores.

Language Measures

The measures of language in this study were all derived from commonly used standardized tests. The measures to be used consisted of either the total raw score for each test within each battery or selected items within each of these tests. The tests and thus the items for each battery were assigned to be either receptive or expressive measures and likewise vocabulary or sentence measures according to the test manual and content of the

Table 1. Characteristics of research participants who were administered each test battery.

| Test battery | n | Age | Boys (%) | Receptive Vocabulary | Block Design |
|--------------|-------|--------------|----------|----------------------|--------------|
| Kindergarten | 1,929 | 6.04 (0.38) | 55 | 0.07 (0.89) | -0.20 (1.02) |
| Second grade | 604 | 7.47 (0.50) | 56 | -0.34 (1.14) | -0.23 (1.17) |
| Fourth grade | 570 | 9.44 (0.50) | 56 | -0.30 (1.06) | |
| Eighth grade | 527 | 13.43 (0.49) | 56 | -0.39 (1.04) | -0.28 (1.20) |

Note. Mean receptive vocabulary is represented in z -score units ($M = 0$, $SD = 1$; standard deviations are presented in parentheses in table) and was measured by the Picture Vocabulary Test of the Test of Oral Language Development—2: Primary in kindergarten and by the Peabody Picture Vocabulary Test—Revised in the remaining batteries. Block design is also represented by z -score values and was measured by the Wechsler Preschool and Primary Scale of Intelligence—Revised in kindergarten and the Wechsler Intelligence Scale for Children—Third Edition for the remaining grades. Block design scores were not obtained in fourth grade.

items. These represented the potential two dimensions to be evaluated in this study. We have chosen to use the term *sentence* rather than *grammar*, although several of these subtests were described as measures of grammar because these sentence level tests may also be influenced by the lexical items used to form the sentences.

One way in which these measures were evaluated required that individual items be examined. In order that each potential dimension had the same weight, it was necessary to have the same number of items in each language area and because the number of items in these tests differed, it was necessary to select items. Across the batteries and language areas, there were usually at least 14 items available. Therefore, 14 items were selected for each language area across the batteries except for eighth grade battery in which there were only 9 expressive vocabulary test items. The selection of the items was based on their item difficulty provided by an item analysis. Items with medium difficulty for their grade level are ideal for modified parallel analysis as these items are less likely to result in extremely high correlations with other items.

At kindergarten, the four language areas were all measured by specific subtests of the TOLD-2:P. Specifically, receptive vocabulary was measured by Picture Identification of TOLD-2:P, expressive vocabulary by Oral Vocabulary, receptive sentence use by Grammatical Understanding, and expressive use by Grammatical Completion and Sentence Imitation. Items selected for use in this study and areas measured by the items were listed in Table 1 for all four grade levels.

At second grade, receptive vocabulary was measured by the Peabody Picture Vocabulary Test—Revised (PPVT-R; Dunn & Dunn, 1981). Expressive vocabulary was measured by the Expressive Vocabulary subtest of the Comprehensive Receptive and Expressive Vocabulary Test (CREVT; Wallace & Hammill, 1994). These two tests were also used to measure receptive and expressive vocabulary, respectively, at fourth and eighth grades. Second grade receptive sentence use was measured by the Concepts and Directions subtest and the Sentence Structure subtest of Clinical Evaluation of Language Fundamentals—III (CELF—III; Semel et al., 1995). Expressive sentence use was measured by the Word Structure subtest and the Recalling Sentences subtest of the CELF—III.

Fourth grade receptive sentence use was measured by the Concepts and Directions subtest of the CELF—III and expressive grammar was measured by the Recalling Sentences subtest and Formulating Sentences subtest of the CELF—III. The Concepts and Directions subtest and Recalling Sentences subtest of the CELF—III were also used at eighth grade to measure receptive grammar and expressive grammar, respectively.

Procedures

All tests were administered individually by four trained examiners who administered and scored all tests according to the test manuals. Prior to administering these tests, the examiners were trained as a group and were then observed by the trainer who was a certified speech-language pathologist. Throughout the study, this trainer also observed and co-scored 5% of the sessions of each of the examiners ensure consistency across examiners and across time. Correlation coefficients computed between composite scores of the trainer and the examiners were above .95 for all tests within all four batteries.

Statistical Analysis

The dimensionality of language test composite scores or items contained in these tests was examined in two ways: revised modified parallel analysis and confirmatory factor analysis. The four language areas were receptive vocabulary, expressive vocabulary, receptive sentence use, and expressive sentence use. The revised modified parallel analysis is an analysis approach based on exploratory factor analysis and was performed on item data to determine the extent to which there was evidence of one versus multiple latent dimensions underlying the individual items of these tests. The confirmatory factor analysis tested whether there were particular latent variables (receptive or expressive, vocabulary or sentence use) underlying the test scores in these language batteries. The composite scores for the confirmatory factor analysis were determined using the standard procedures described in the test manual. For parallel analysis, only selected items were included. This rule applied to all four grade levels.

Revised modified parallel analysis (Budesu, Cohen, & BenSimon, 1997). At a conceptual level, it consists of an exploratory factor analysis performed on the individual test items. Exploratory factor analysis provides a test of the assumption that the individual items are not independent of each other, but rather performance on these items is reflective of one or more latent traits or dimensions. The exploratory factor analysis involves first computing the intercorrelation of the obtained test items, and then it analyzes this correlation matrix to produce unrotated eigenvalues. Because all items in this study were scored in a binary fashion, the correlation between items was based on a tetrachoric correlation derived from a 2×2 table as would be used to derive a chi-square with one degree of freedom. This tetrachoric correlation represents whether correct and incorrect responses on one item are associated with correct and incorrect responses on another item. The factor analysis provides a means of reducing the correlation matrix to a smaller set of potential latent variables underlying

these correlations. The eigenvalue for each dimension represents the amount of variance accounted by each dimension. The larger the eigenvalue, the more that particular latent variable accounts for the covariance among items in the matrix.

This type of factor analysis is referred to as *exploratory factor analysis* because no initial model concerning the manner in which items are correlated is being tested. As a result, interpreting the results can be challenging because it is difficult to determine whether one or more than one latent dimension underlies the items being studied. Interpretation can be aided by constructing a parallel factor analysis on data that are generated from data in which the number of latent variables is fixed to one. The obtained analysis can then be compared with the parallel simulated analysis to determine whether a single-dimensional model as shown by the simulation is comparable to the obtained data.

If the comparison is to be valid, it is necessary to generate simulated data item characteristics similar to the actual test given to examinees, except that these items are known to be measuring a single latent trait. These item characteristics can be obtained using IRT that describes items with respect to their difficulty and discrimination. These features can be based on the obtained items and then applied to the simulated items, thus producing a parallel analysis. With these parallel items created, they can then be subjected to the same type of factor analysis as performed on the obtained data. Thus, tetrachoric correlations can be computed for the simulated items and the correlation matrix factor analyzed to produce unrotated eigenvalues. The eigenvalues computed from a tetrachoric correlation matrix of obtained item scores are then compared with the eigenvalues computed from a tetrachoric correlation matrix of item scores from the parallel data set that are known to be unidimensional by visual examination of scree plots. A scree plot for the eigenvalues presents a pattern of declining eigenvalues, and visual inspection can be used to determine how many factors should be retained in the factor analysis. The scree plot from the obtained data in the exploratory analysis can therefore be compared with

the scree plot from the parallel analysis to determine the extent to which the obtained data conform to a single-dimensional model. In a scree plot, the first unrotated factor always accounts for the greatest amount of variance and the last unrotated factor accounts for the least. The decreasing rate from the first eigenvalue to the later ones reflects the dimensionality of the data. If the data are perfectly unidimensional, the first factor would account for all the correlations between all variables. The remaining factors simply reflect random clustering of correlations that represent error. Thus, the scree plot of the simulated unidimensional data provides a model scree plot against which the obtained scree plot can be compared.

The particular method of parallel analysis used in this study was that proposed by Budescu et al. (1997; also see the Appendix for more information). This method skips the step of generating individuals' scores (1 or 0) on each simulated item. Instead, this method generates only an expected 2×2 contingency table for each pair of items (based on item parameters and a unidimensional assumption), with the four cells of the table corresponding to the number of individuals passing both items, passing the first item but failing the second item, failing the first but passing the second, and failing both items (see Table 2). From this contingency table, a tetrachoric correlation can be estimated. After a correlation between all possible item pairs has been estimated, a tetrachoric correlation matrix is formed and a factor analysis can be performed on this correlation matrix.

Specifically, the eigenvalues for the revised modified parallel analysis analysis in this study were derived from tetrachoric correlation matrices formed from the item level data obtained from the participants' performance on the items listed in Table 3 or from the expected contingency tables. These correlations were computed using the SAS macro POLYCHOR (SAS Institute, 2000). To generate the expected contingency table, the difficulty and discrimination power of each item needed to be estimated from the obtained data using IRT. In this case, a two-parameter IRT model consisting of item parameters representing item difficulty (δ) and discrimination power

Table 2. Expected contingency table for numbers of individuals who pass (or fail) item i , item j , and both items i and j in a sample of K individuals.

| | | No. out of K individuals who pass or fail item i | | |
|--|------|--|---------------------|-----------|
| | | Pass | Fail | Total |
| No. out of K individuals who pass or fail item j | pass | A | $M_j - A$ | M_j |
| | fail | $M_i - A$ | $K - M_i - M_j + A$ | $K - M_j$ |
| Total | | M_i | $K - M_i$ | K |

Note. A , M_i , and M_j were computed using the formulas in the Appendix.

Table 3. Items selected from standardized language tests used in kindergarten, second, fourth, and eighth grades for inclusion in modified parallel analysis.

| Test items | Kindergarten | Second grade | Third grade | Fourth grade | No. of items | Modality | Language domain |
|---|--------------|--------------|-------------|--------------|--------------|----------|-----------------|
| TOLD-2:P | | | | | | | |
| Picture Vocabulary subtest Items 9-22 | X | | | | 14 | R | V |
| Oral Vocabulary subtest Items 2-15 | X | | | | 14 | E | V |
| Grammatical Understanding subtest Items 11-24 | X | | | | 14 | R | S |
| Sentence Imitation subtest Items 4-10 | X | | | | 7 | E | S |
| Grammatical Completion subtest Items 6-12 | X | | | | 7 | E | S |
| PPVT-R | | | | | | | |
| Items 70, 71, and 74-79 | | X | | | 8 | R | V |
| Items 80-83 | | X | X | | 4 | R | V |
| Item 84 | | X | | | 1 | R | V |
| Item 85 | | X | X | | 1 | R | V |
| Items 86-94 | | | X | | 9 | R | V |
| Items 99-106 and Items 108-113 | | | | X | 14 | R | V |
| CREVT | | | | | | | |
| Items 2 and 3 | | X | | | 2 | E | V |
| Items 4-10 | | X | X | | 7 | E | V |
| Items 11-15 | | X | X | X | 5 | E | V |
| Items 16 and 17 | | | X | X | 2 | E | V |
| Items 18 and 19 | | | | X | 2 | E | V |
| CELF | | | | | | | |
| Word Structure subtest Items 26-32 | | X | | | 7 | E | S |
| Recalling Sentences subtest Items 6-8 | | X | | | 3 | E | S |
| Recalling Sentences subtest Items 9 and 10 | | X | X | | 2 | E | S |
| Recalling Sentences subtest Items 11 and 12 | | X | X | X | 2 | E | S |
| Recalling Sentences subtest Items 13-15 | | | X | X | 3 | E | S |
| Recalling Sentences subtest Items 16-24 | | | | X | 9 | E | S |
| Formulated Sentences subtest Items 6-12 | | | X | | 7 | E | S |
| Sentence Structure subtest Items 14-20 | | X | | | 7 | R | G |
| Concept and Directions subtest Items 17-23 | | | X | X | 7 | R | G |
| Concept and Directions subtest Items 24-30 | | X | X | X | 7 | R | G |

Note. Items were assigned to modality (receptive [R] and expressive [E]) and language domain (vocabulary [V] and sentence [S]). TOLD-2:P = Test of Oral Language Development—2: Primary; PPVT-R = Peabody Picture Vocabulary Test—Revised; CREVT = Comprehensive Receptive and Expressive Vocabulary Test; CELF-3 = Clinical Evaluation of Language Fundamentals—3.

(α) given to an examinee with an ability (θ) was obtained from the IRT analysis on the real data using BILOG (Meslevy & Bock, 1990). These parameters were obtained from the performance of all participants at each grade level, and, therefore, the same distribution of ability contributed to each item at each grade level.

These item parameters were then used to compute the expected contingency table for each pair of items using a method described by Budescu et al. (1997), which is reviewed in the Appendix. After the tetrachoric correlations between all 56 items were computed from the contingency tables, the eigenvalues for this correlation matrix were derived and compared with the eigenvalues from the real data. After this, the factor loadings of each

item onto each factor were examined so that the meaning of the factors could be discerned.

Confirmatory factor analysis provides a means of testing specific hypotheses concerning the presence of latent traits or dimensions underlying test scores. Because the hypothesized dimensions pertained to the organization of subtests, composite scores for subtests were used rather than employing the individual test items as used in the revised modified parallel analysis. The competing models were a unidimensional model treating all of the subtests as measuring one latent trait, “language,” and a model with two latent traits. The subtests could be either classified into receptive–expressive categories or vocabulary–sentence categories. For all the four grade

levels, it was found that the model goodness of fit was better with vocabulary–grammar classification than with receptive–expressive classification. Thus, the one-factor model was compared with this vocabulary–grammar two-factor model. This confirmatory factor analysis was conducted on language subtest scores using the computer program Mplus (Muthen & Muthen, 2001). There were five subtests from the kindergarten battery, six from second grade, five from fourth grade, and four from eighth grade.

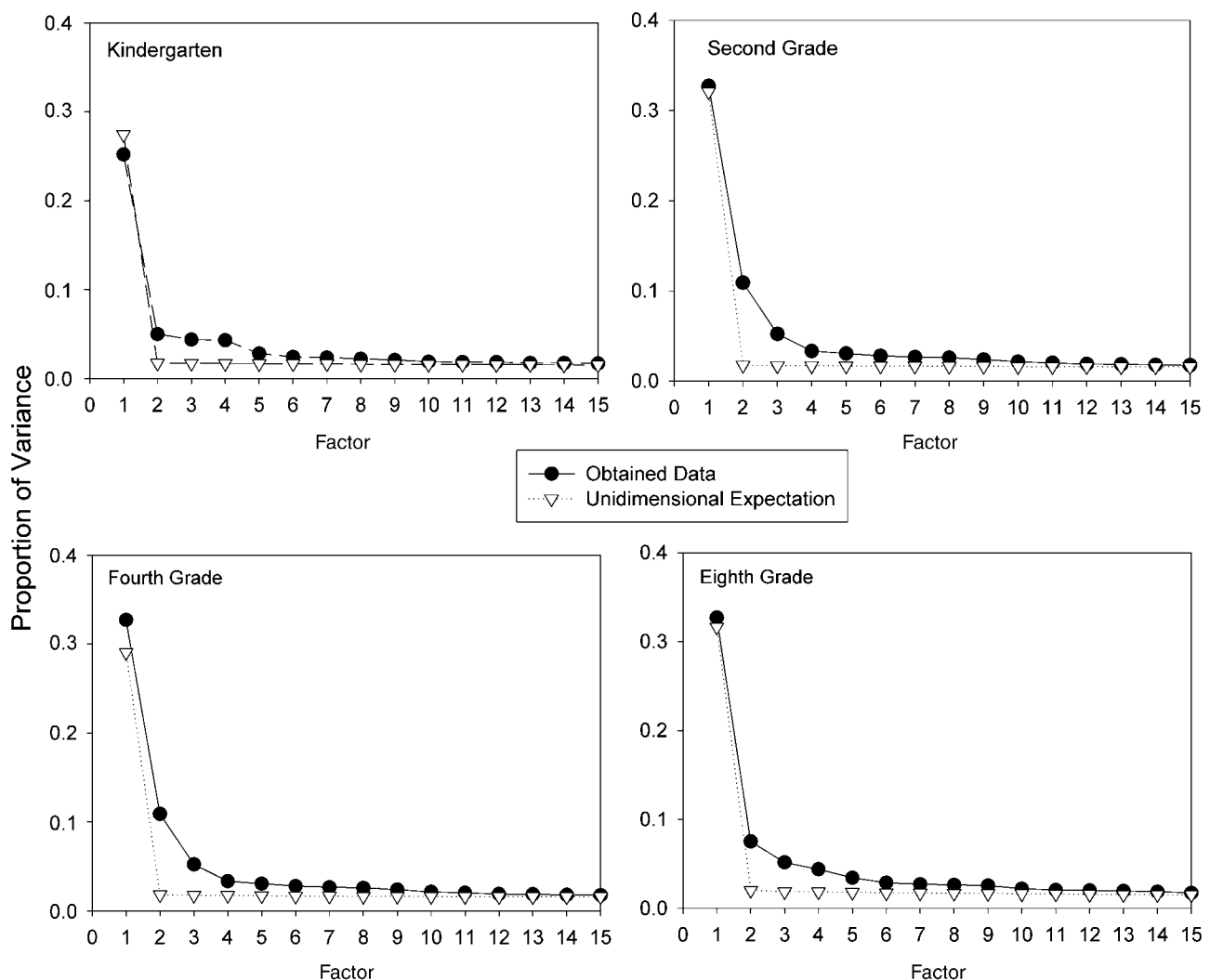
To see how the relation between the vocabulary and grammar factor changes, the correlation between the two factors was compared across grade levels. Comparison between second grade, fourth grade, and eighth grade was also conducted using scores from those subtests that were administered at all three grade levels. These subtests were the PPVT–R, CREVT, CELF–III Concepts and Directions subtest and the Recalling Sentences subtest.

Results

Revised Modified Parallel Analysis

This analysis asked whether there was more than one latent variable underlying the set of items drawn from the various subtests in each battery. Tetrachoric correlations representing correlations between item pairs were computed for the 56 items in each of the kindergarten, second grade, and fourth grade batteries and the 51 items in the eighth grade battery. Eigenvalues were obtained from these tetrachoric correlation matrices. The 15 largest eigenvalues for each battery are shown in the form of a scree plot in Figure 1. Each eigenvalue corresponds to an independent variance component or factor that can be thought of as a possible latent variable; however, these variance components can also simply represent minor clustering of measurement error. Therefore, it is necessary to determine which components are likely to

Figure 1. Scree plot of the largest 15 eigenvalues divided by the total number of items for the real data and unidimensional simulated data.



reflect valid latent variables. Larger eigenvalues indicate that a factor accounts for a larger amount of the total variance than smaller eigenvalues. The factors with large eigenvalues are the most likely to represent valid latent variables. Examination of Figure 1 shows that there was a clear break point on the scree plot after the first factor. The eigenvalues computed based on IRT parameter expectations, assuming that all items are measuring the same ability, are also shown in Figure 1. The obtained data can be seen to be similar to the expected data across the batteries; however, some deviation from the unidimensional example can be seen for each battery, suggesting that the items in these batteries are not fully unidimensional.

Because eigenvalues represent amount of variance accounted for by a factor, the eigenvalues can be converted to indices of percentage of total variance accounted for. Across the four batteries, the first component accounted for 25.19% of the total sample variance in the kindergarten battery, 32.07% in the second grade, 32.69% in the fourth grade, and 31.60% in the eighth grade. The second principal component in all of these batteries accounted for much less: 4.98% at kindergarten, 10.89% at second grade, 10.89% at fourth grade, and 7.51% in eighth grade. These results show that across the batteries, the first factor accounted for a substantially greater amount of the variance than the next largest factor. In second, fourth, and eighth grades, the second factor accounted for more than 5% percent of the total variance; whereas in kindergarten, the second, third, and fourth factors were similar and collectively accounted for 13.61% of the variance.

Because the lower order factors accounted for a modest proportion of the total variance, the items from these test batteries were examined with respect to their loading onto the factors that accounted for more than 10% of the variance. Because the kindergarten battery had three lower order factors that collectively loaded over 10%, we examined all four factors for kindergarten. The loadings of items onto the first four factors for kindergarten are shown in Figure 2. As expected, all items generally load onto the first factor. Items loading on the higher order factors were in each case largely associated with a particular subtest. Items loading onto the second factor were largely from the Grammatical Understanding subtest, whereas the third factor predominantly comprised the Expressive Vocabulary items and the fourth factor was marked by high loadings from the Receptive Vocabulary subtest. At second and fourth grades, all items loaded onto the first factor and the second factor appeared to consist of items from the PPVT-R. In eighth grade, as in all other batteries, there was a generally uniform loading of items onto the first factor. The second factor at eighth grade comprised items from two tests each representing vocabulary skills, but in two very different modes of use (recognition vs.

expressive/metalinguistic definition construction). Items from the receptive and expressive sentence usage tasks negatively loaded onto this factor.

Collectively, these results from the revised modified parallel analysis suggest that when examined at the item level, these items drawn from different tests are all measuring predominantly a single latent variable. In some cases, these items also reflect some additional latent variables; in these cases, these seem to be associated with specific tests or, as in eighth grade, the domain of vocabulary.

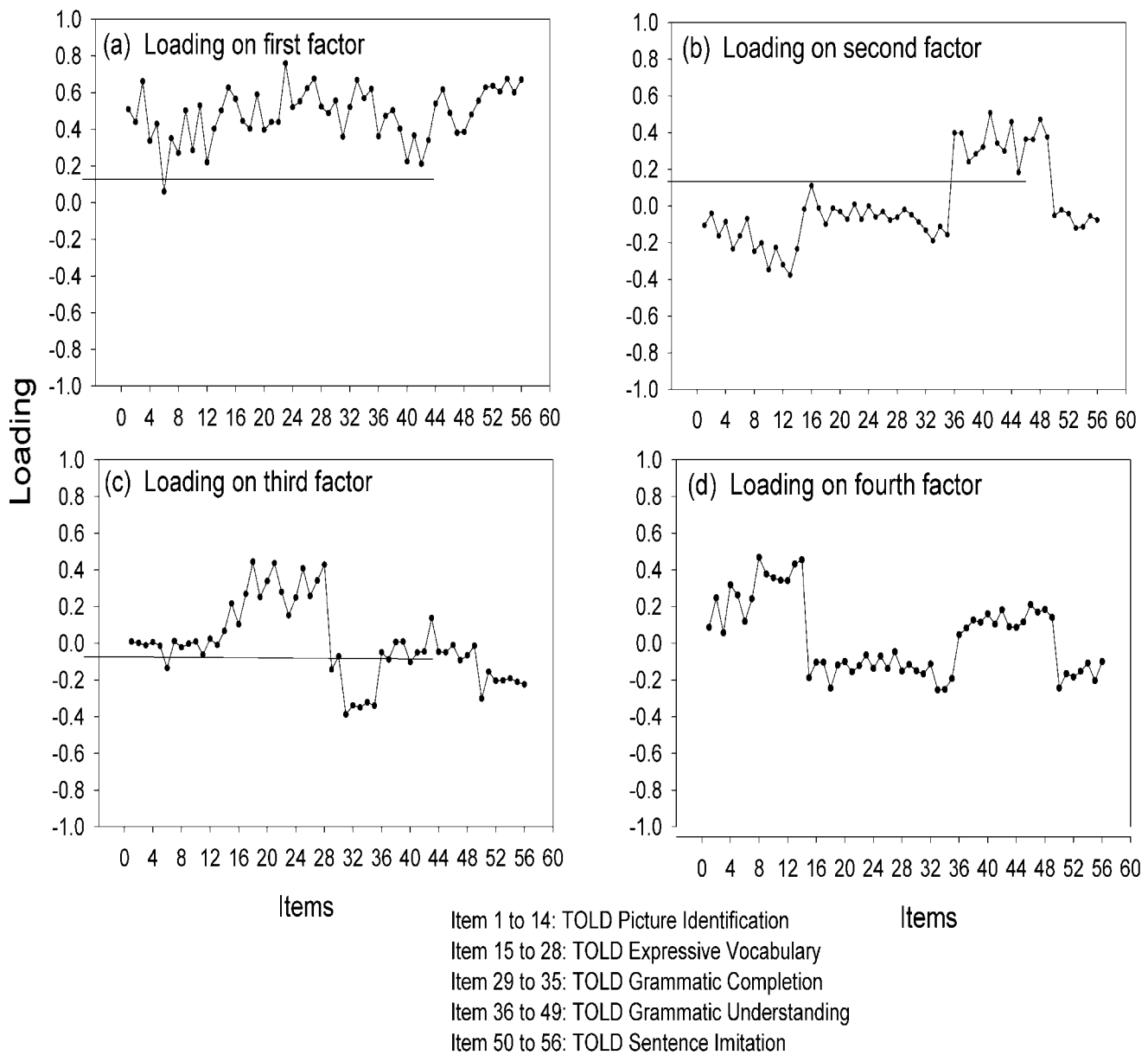
Confirmatory Factor Analysis

The parallel analysis previously performed provided a means of examining the dimensionality of the language test batteries at the item level using an exploratory approach. Exploratory factor analytic approaches are largely descriptive in nature and lack specific tests for alternative interpretations. Confirmatory factor analysis does provide such methods but requires the specification of a model. The parallel analysis largely supported a unidimensional model; however, some evidence was found for two dimensions that made up the language domains of vocabulary and grammar, particularly in the eighth grade battery. An alternate two-dimensional model consisting of receptive and expressive language was not suggested in the exploratory parallel analysis; however, this model cannot be refuted from this analysis and was therefore also tested in the confirmatory analysis.

The results of confirmatory factor analysis appear in Figures 3, 4, 5, and 6. In each figure, the correlations of each subtest with the latent factor(s) are shown. Values are also provided for the correlations between the two factors in the cases of the two-factor models. Finally, a set of measures indicating the goodness of fit of the models are provided. The principal goodness-of-fit criteria are the Comparative Fit Index (CFI) (larger values are better, and values more than .95 are considered excellent) and the Akaike Information Criterion (AIC), which rewards models for good fit but penalizes extra parameters (lower values are more desirable).

As can be seen from the figures, for all grade levels the two-factor model consisting of receptive and expressive aspects of language was not better than the single-factor model. In contrast, the comparison of the two factors of vocabulary and sentence use with the one-factor model favored the two-factor model across the batteries because the two-factor model had a smaller value of AIC. It should be noted, however, that the single-factor model fit the data nearly as well as the two-factor models for the kindergarten, second grade, and fourth grade data. In each of these cases, the CFI values were always well above .95 for the single-factor model, which is considered evidence of a good fit. In eighth grade,

Figure 2. Loading of items onto first and second factors. TOLD = Test of Oral Language Development.

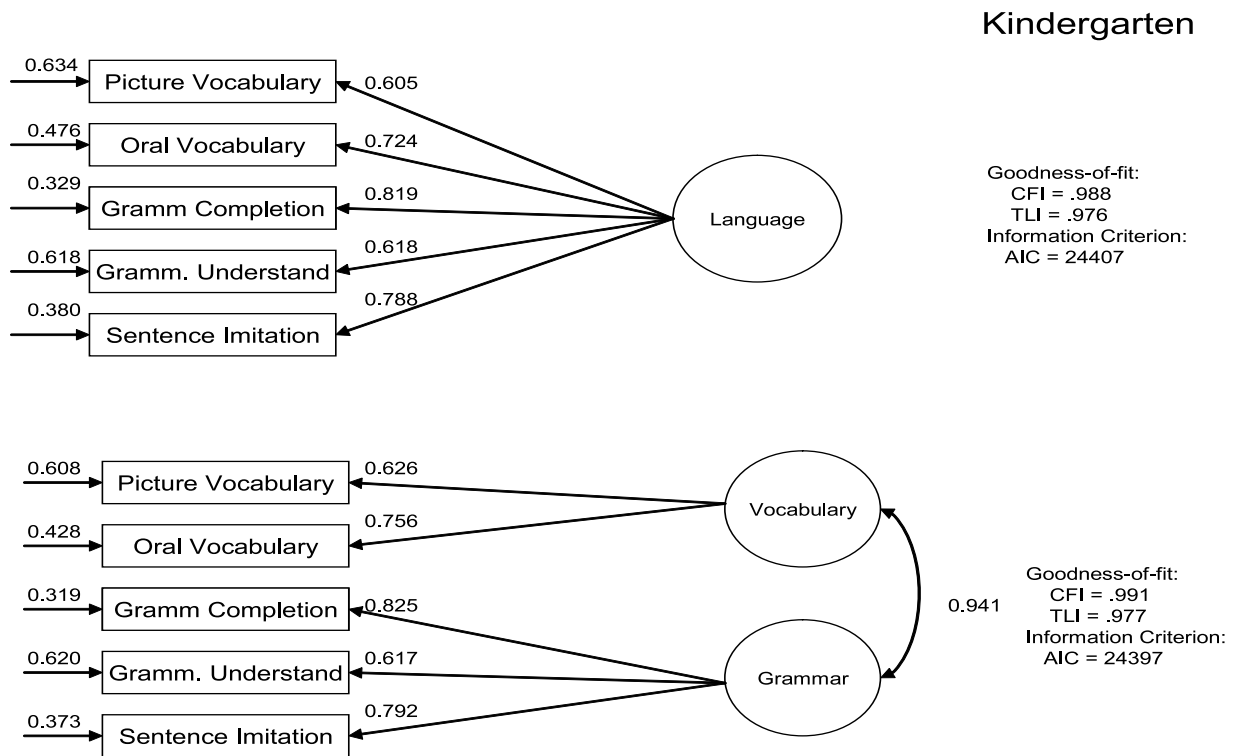


however, the two-factor model was clearly better than the one-factor model, even though the goodness of fit (CFI = .932 and TLI [Tucker–Lewis index] = .796) values for this model were still acceptable.

As shown in Figures 3–6, the correlations between the sentence use and vocabulary factors were high for all batteries; however, it can be seen that the correlations between these two domains decline across the grade levels (for kindergarten, $r = .941$; for second grade, $r = .934$; for fourth grade, $r = .902$; and for eighth grade, $r = .782$). By eighth grade, the two domains show the potential for separate traits underlying these domains.

In the previous analysis, the test batteries that were used did not use the same subtests in each case. Thus, it is possible that the factor correlation change across batteries may have resulted from the changes in subtest composition from kindergarten to eighth grade. To rule out this possibility, we compared the factor correlations between second grade, fourth grade, and eighth grade for only those tests that were administered at all the three grade levels. The results are shown in Figure 7. In these correlations, there is a clear decreasing trend from .925 at second grade to .861 at fourth grade and to .782 in eighth grade. This change showed that language

Figure 3. Confirmatory factor analysis of kindergarten subtest scores for models combining subtest scores according to modality or to the language domains of vocabulary and sentence use. CFI = Comparative Fit Index; TLI = Tucker–Lewis index; AIC = Akaike Information Criterion.



ability differentiated from mainly one factor to correlated two factors if one agrees that two factors correlated above .90 can be combined into one factor. Thus, even when the analysis was limited to data derived from the same instruments across age levels, the same type of factor structure was obtained.

Discussion

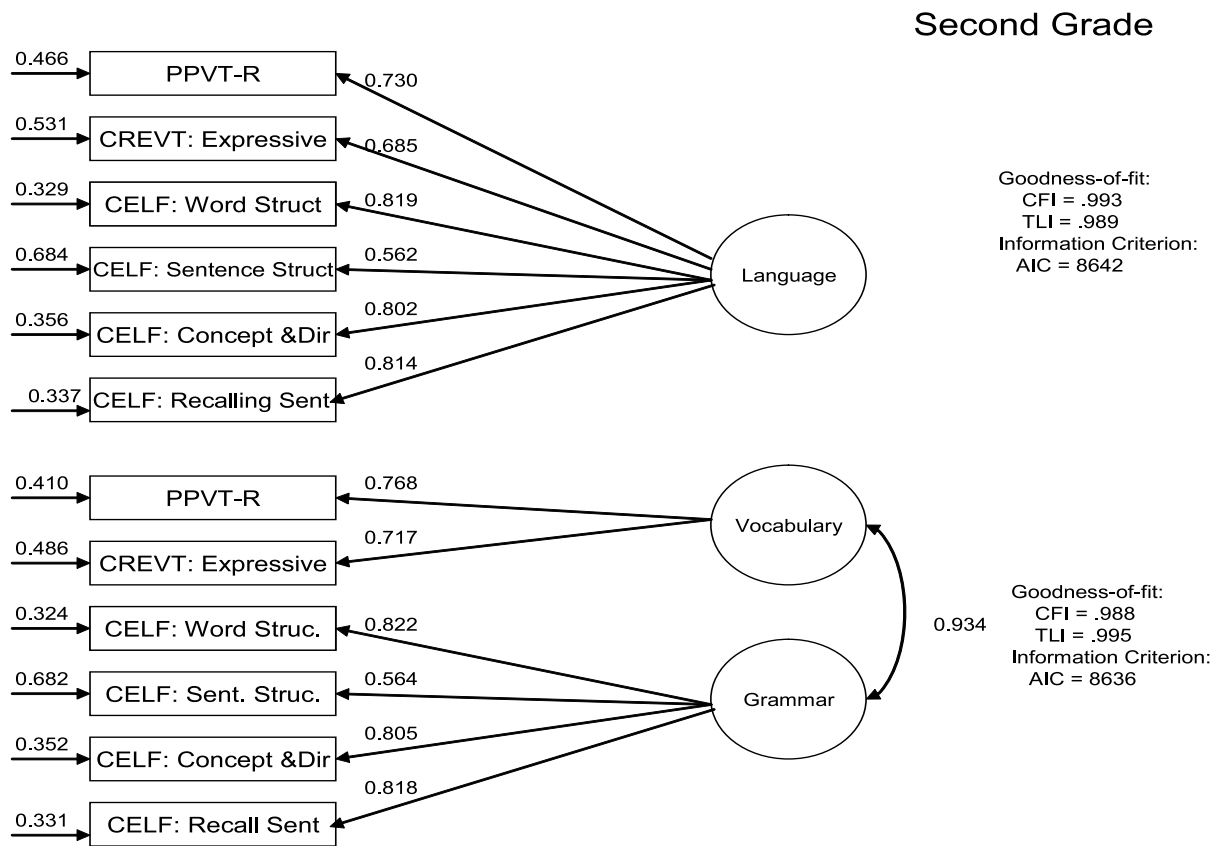
This study was concerned with the extent to which the various types of language tasks contained within standardized language assessment instruments for school-age children reflect different underlying language abilities. The literature on language development and language disorder has often assumed that multiple language skills are measured by such instruments. Most often, these skills are either characterized in terms of the modalities of expression and reception or linguistic domain in the form of lexical (vocabulary) and grammatical abilities. In this study, items reflecting each of these categories were examined at four time points across the age range of 6 years to 14 years of age using both exploratory and confirmatory analysis methods. The results of each of these methods were similar and provide a fairly coherent picture of the information provided by these language measures.

Evidence for Dimensionality

The exploratory analysis using modified parallel analysis permitted an examination of the dimensionality of the language measures at an item level. In this respect, no preexisting assumptions were made about what the items were measuring. At each of the four grade levels, there was evidence of a dominant single factor that all items loaded onto. In modified parallel analysis, the evidence for dimensionality is based on a comparison of the eigenvalues of the second and third factors for the obtained data and the data expected under a unidimensional IRT model. In each of the four item sets, there was some evidence of multidimensionality. The nature of these lower order factors was revealed by plotting the factor loading for each item, as shown in Figure 1.

For those items in the kindergarten battery using the TOLD-2:P, three lower order factors collectively accounting for more than 10% of the variance were found. Each of these factors could be attributed to one of the subtests of the TOLD-2:P. In this respect, there is no strong evidence for the language performance of these children showing systematic variance across a modality or a domain of language, but rather these lower level factors appeared to reflect task-specific features. In

Figure 4. Confirmatory factor analysis of second-grade subtest scores for models combining subtest scores according to modality or to the language domains of vocabulary and sentence use. PPVT-R = Peabody Picture Vocabulary Test—Revised; CREVT = Comprehensive Receptive and Expressive Vocabulary Test; CELF = Clinical Evaluation of Language Fundamentals.



second, fourth, and eighth grades, only one lower order factor accounted for an appreciable amount of the variance. The item loadings in second and fourth grades onto this second factor were all items from the PPVT-R, and all other items including the expressive vocabulary measure formed the contrasting factor. Thus, as in kindergarten, a coherent lower order factor that transcended a single test was not apparent, again suggesting task-specific variance as the basis of the lower order dimensionality. In eighth grade, however, the items loading onto the second factor continued to include the PPVT-R items, but the expressive vocabulary items from the CREVT loaded onto this factor as well. Thus, by eighth grade, the items concerning vocabulary, regardless of the manner of assessment, seem to be tapping into a common aspect of language and, in contrast, the items that involved sentence processing organized in a separate factor. Nowhere in the exploratory analysis was there support for a receptive-expressive dimension.

The exploratory analysis performed at the item level also showed that the items contained in any given subtest across the batteries were quite homogeneous. Thus, there

was no evidence that these items should be reorganized to form subtests other than the ones they have been assigned to by the test developers. Given that the batteries for second grade through eighth grade comprised items from three different tests, these findings cannot be attributed to prior item analyses in the construction of the tests.

The pattern of an increasing presence of multidimensionality of the items in these test batteries around a vocabulary-sentence contrast was supported by the confirmatory factor analysis. Although across all of the batteries the two-factor model involving a vocabulary-grammar contrast yielded a slightly better fit to the data, it was only in eighth grade that this two-factor model was noticeably better than the one-dimensional model. The trend toward a multidimensionality of language contrasting vocabulary and sentence use was shown in the systematic decline in the correlations between these two factors across grade levels. These results could be viewed as providing support for there being some dimensionality of language with regard to the contrast of vocabulary versus sentence use, particularly among

Figure 5. Confirmatory factor analysis of fourth-grade subtest scores for models combining subtest scores according to modality or to the language domains of vocabulary and sentence use.

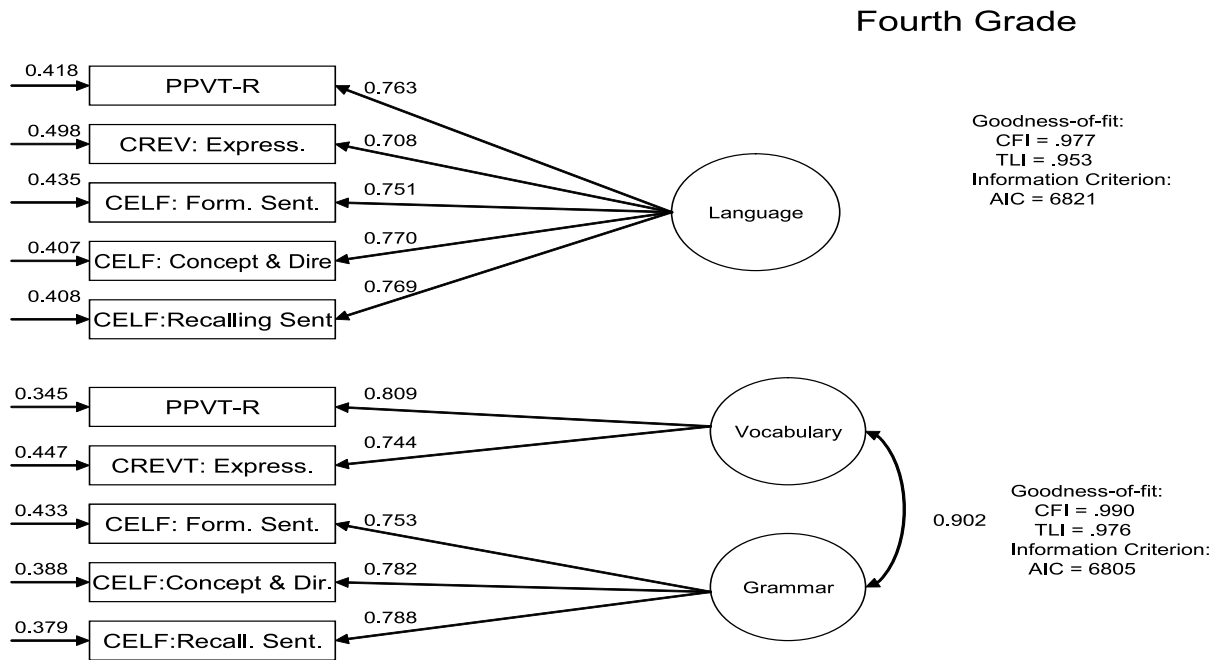


Figure 6. Confirmatory factor analysis of eighth-grade subtest scores for models combining subtest scores according to modality or to the language domains of vocabulary and sentence use.

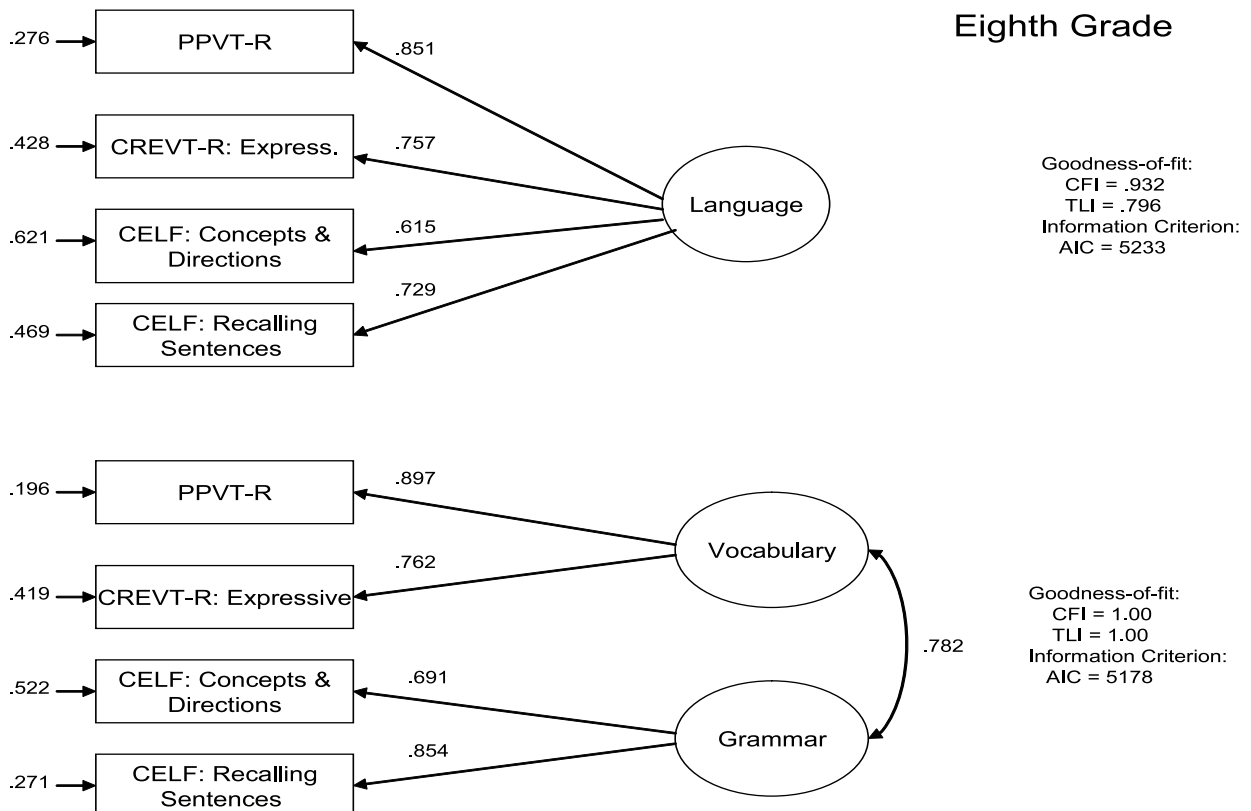
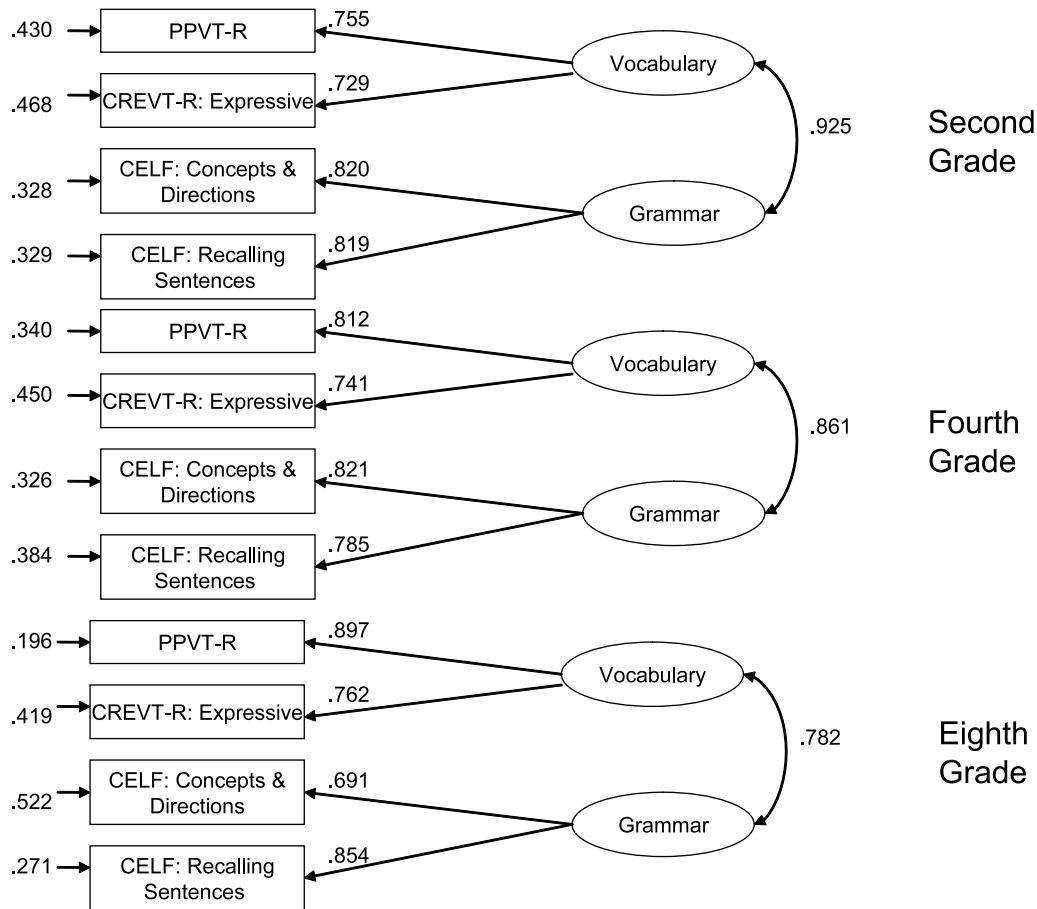


Figure 7. Confirmatory factor analysis of subtest scores at second, fourth, and eighth grades for models combining subtest scores according to modality or to the language domains of vocabulary and sentence use when only subtests common to all grades were used.



older children. Such a contrast provides some support for those who argue for a distinction between vocabulary and grammar (Pinker, 1998; van der Lely, Rosen, & McClelland, 1998). However, if these data are to be used to support such a position, it will be necessary to explain why this dimensionality is weak initially and only appears at older age levels.

This increase in dimensionality concerned with vocabulary and sentence use with increments in participants' age can be explained in several ways. One explanation draws on the obvious requirement of the sentences in these tasks to use lexical items. As Bates and Goodman (1999) have noted, it is not possible to have a task involving grammatical performance without also employing the lexicon (Bates & Goodman, 1999). To minimize the influence of lexical ability on grammatical ability, measures of sentence use should use lexical items that are familiar to the examinee. The technical manual of the CELF-III does not discuss the criteria for selection of lexical items of the sentence material. It is possible that young examinees would be less familiar with some of the words used in the sentence tasks than the older

examinees. If this were so, then the confounding of vocabulary with sentence use would diminish with age and greater dimensionality would become apparent. This explanation allows for the processing and representational systems involved at the word level and the sentence level skills to be inherently separate at least to some degree throughout this developmental period, and the high level of association between these in the early years is explained by a measurement confound.

Alternatively, the increasing dimensionality of these systems with age could be viewed as support for those who view sentence level systems such as syntax and morphology as emergent (Bates & Goodman, 1999). Such accounts predict that a grammatical system that is at least partially independent of lexical abilities would become apparent as the language user approached maturity as a language user.

A final explanation for these results also needs to be considered. The grammatical system that may account for much of the unique knowledge to be learned at the sentence level is typically thought of as a finite symbolic system (Chomsky, 1965) that is mastered by middle-to-late

childhood. Thus, the learning trajectory for much of the component skills involved in sentence use is likely to be nonlinear with an asymptotic feature, as has been shown for tense marking by Rice and colleagues (Rice, Wexler, & Hershberger, 1998). Knowledge of vocabulary is much less developmentally constrained, and growth often proceeds well into adulthood. The appearance of greater independence between measures of vocabulary and sentence use in early adolescence may reflect the different growth trajectories of these systems as grammatical development asymptotes and lexical development continues to progress.

These three explanations for the systematic decrease in the association between vocabulary and sentence use across development cannot be distinguished with the data at hand. At this point, further study of these data, and preferably data based on stimuli and tasks that incorporate controls such as lexical confounds in sentence tasks, should permit better tests of these different accounts.

Implications for Clinical and Research Applications

Standardized tests of language for the measurement of children's spoken language abilities are used widely in both research and clinical settings. In most instances, the users of these tests are interested in aggregating the individual item performance of children in some way to reach conclusions about the language trait status of the children receiving the test. The results of this study would support such an approach. In both the exploratory and confirmatory analyses, there was strong evidence for a general language trait upon which all items and subtests loaded rather well. In this regard, language, as measured and represented in these various tests, can be thought of predominantly as a single trait, particularly in the early school years.

In some research or clinical situations, these measures are used to characterize profiles of strengths and weaknesses and, in some instances, the subtype of language impairment. The results of this study show that these measures only provide information for certain kinds of contrasts, and even then only for certain ages. The current findings indicate that these measures are not likely to be able to reflect reliable differences within individuals with respect to receptive and expressive modalities. When such differences are found, these results would suggest that the differences are more likely spurious and unreliable, which may explain the findings of Conti-Ramsden and Botting (1999) wherein subtypes that were partially or wholly reflective of modality dimensions were found to be unstable. Such conclusions stand in contrast with common clinical and research practice in which children with language disorders are subtyped according to modality (see, for instance, Conti-Ramsden, Crutchley, & Botting, 1997; Rapin, 1996; Wilson & Risucci, 1986). It

remains possible that these subtypes do exist; however, to identify these subtypes it will be necessary to develop new measures that can be shown to measure receptive and expressive language independently.

In contrast with the receptive-expressive dimension, it does seem that separate measures of vocabulary and sentence use may be informative, but only among older school-age children. Even in this case, there is good evidence that these are not independent traits and that the correlation between these aspects of language is high. Thus, a single composite score does not result in much loss of information and probably results in better reliability for clinical decision making. Taken collectively, these results would suggest that in clinical or research settings it is not necessary to employ as many different language assessment tasks as are often provided in commercial test batteries. Simply because a test or subtest involves different content or different tasks does not mean that the latent trait measured by the measure is different from those measured in other subtests. Instead, this is an empirical question that should be provided with each test.

Generality of Findings

Several years ago, Clark (1974) pointed out that research in language must contend with the problem of generalization of results beyond the specific language items used in research. These concerns certainly apply to the findings from this study as well. If one has randomly sampled items from a universe of items, then it is reasonable to generalize the results back to that universe. However, we must exercise caution in cases such as this study, where the manner in which the items were sampled is not known and, in fact, it is not even clear what universe these items were drawn from. Thus, a conservative conclusion would be that the dimensionality of the items and subtests of the particular tests included in this study is low, consisting of one major dimension and one emerging minor dimension. The measures used in this study did comprise some of the most commonly used instruments for testing language in children, and these instruments used the most typical methods of testing language among children in this age range. Thus, it is at least reasonable to argue that these conclusions are probably applicable to other similar standardized tests of language. The measures in this study did not extend to levels of language often contained in the domains of discourse, text, and pragmatics. Therefore, it would be inappropriate to conclude that the findings of this study can be generalized to these domains of language use.

Acknowledgment

This study was supported by National Institutes of Health Contract DC-19-90 from the National Institute on Deafness

and Other Communication Disorders and by a clinical research center grant (P0-DC-02748), which is also from the National Institute on Deafness and Other Communication Disorders. The conduct of this study was aided considerably by a valuable research team comprising Marlea O'Brien, Paula Buckwalter, Juanita Limas, Connie Ferguson, Jodi Schwirtz, Amy Schminke, and Marsha St. Clair.

References

- Bates, E., & Goodman, J.** (2001). On the inseparability of grammar and the lexicon: Evidence from acquisition. In M. Tomasello & E. Bates (Eds.), *Language development: The essential readings* (pp. 134–162). Oxford, England: Blackwell.
- Bates, E., & Goodman, J. C.** (1999). On the emergence of grammar from the lexicon. In B. MacWhinney (Ed.), *The emergence of language* (pp. 29–79). Mahwah, NJ: Erlbaum.
- Budescu, D. V., Cohen, Y., & BenSimon, A.** (1997). A revised modified parallel analysis for the construction of unidimensional item pools. *Applied Psychological Measurement, 21*, 233–252.
- Chomsky, N.** (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
- Clark, H.** (1974). The language-as-fixed effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 355–359.
- Conti-Ramsden, G., & Botting, N.** (1999). Classification of children with specific language impairment: Longitudinal considerations. *Journal of Speech, Language, and Hearing Research, 42*, 1195–1204.
- Conti-Ramsden, G., Crutchley, A., & Botting, N.** (1997). The extent to which psychometric tests differentiate subgroups of children with SLI. *Journal of Speech, Language, and Hearing Research, 40*, 765–777.
- Crocker, L., & Algina, J.** (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Holt, Rinehart and Winston.
- de Villers, J.** (2003). Defining SLI: A linguistic perspective. In Y. Levy & J. Schaefer (Eds.), *Language competence across populations: Toward a definition of specific language impairment* (pp. 425–447). Mahwah, NJ: Erlbaum.
- Dunn, L. M., & Dunn, L. M.** (1981). *Peabody Picture Vocabulary Test—Revised*. Circle Pines, MN: AGS.
- Meslevy, R. J., & Bock, R. D.** (1990). *BILOG 3: Item analysis and testing scoring with binary logistic models*. Mooresville, IN: Scientific Software.
- Miller, J.** (1981). *Assessing language production in children*. Needham-Heights, MA: Allyn & Bacon.
- Muthen, L. K., & Muthen, B. O.** (2001). *Mplus user's guide*. Los Angeles: Author.
- Newcomer, P., & Hammill, D.** (1988). *Test of Oral Language Development—2: Primary*. Austin, TX: Pro-Ed.
- Paul, R.** (2001). *Language disorders from infancy through adolescence*. St. Louis, MO: Mosby.
- Pinker, S.** (1997). Words and rules in the human brain. *Nature, 387*, 547–548.
- Pinker, S.** (1998). Words and rules. *Lingua, 106*, 219–242.
- Rapin, I.** (1996). Practitioner review: Developmental language disorders: A clinical update. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 37*, 643–655.
- Rice, M. L., Wexler, K., & Hershberger, S.** (1998). Tense over time: The longitudinal course of tense acquisition in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 41*, 1412–1431.
- SAS Institute.** (2000). *SAS OnlineDoc* (Version 8). Cary, NC: Author.
- Schatschneider, C., Francis, D. J., Foorman, B. R., Fletcher, J. M., & Mehta, P.** (1999). The dimensionality of phonological awareness: An application of item response theory. *Journal of Educational Psychology, 91*, 439–449.
- Semel, E., Wiig, E., & Secord, W.** (1995). *Clinical Evaluation of Language Fundamentals—III*. San Antonio, TX: The Psychological Corporation.
- Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M.** (1997). The prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40*, 1245–1260.
- Tomblin, J. B., Records, N. L., & Zhang, X.** (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research, 39*, 1284–1294.
- van der Lely, H. K., Rosen, S., & McClelland, A.** (1998). Evidence for a grammar-specific deficit in children. *Current Biology, 8*, 1253–1258.
- Wallace, G., & Hammill, D.** (1994). *Comprehensive Receptive and Expressive Vocabulary Test*. Austin, TX: Pro-Ed.
- Wechsler, D.** (1989a). *Wechsler Intelligence Scale for Children* (3rd ed.). San Antonio, TX: The Psychological Corporation.
- Wechsler, D.** (1989b). *Wechsler Preschool and Primary Scale of Intelligence—Revised*. San Antonio, TX: The Psychological Corporation.
- Wilson, B. C., & Risucci, D.** (1986). A model for clinical-quantitative classification. Generation I: Application to language-disordered preschool children. *Brain and Language, 27*, 281–309.

Received April 21, 2005

Accepted March 9, 2006

DOI: 10.1044/1092-4388(2006/086)

Contact author: J. Bruce Tomblin, Department of Speech Pathology and Audiology, University of Iowa, Iowa City, IA 52242. E-mail: j-tomblin@uiowa.edu.

Appendix. Revised modified parallel analysis.

Revised modified parallel analysis differs from modified parallel analysis in that the parallel unidimensional data set against which the obtained data are compared is derived from computational formulas rather than generated via simulation. The derivation of the unidimensional data begins by employing item parameters based on item response theory (IRT) that have the same parameter values as those in the obtained data set. These IRT item parameters for a two-parameter model are α_i (the item discrimination parameter) and δ_i (the item difficulty parameter). The probability of a correct response for item i by an examinee with a single latent trait θ_i ($U_{ij} = 1$) is

$$p_{ij} = Pr(U_{ij} = 1/\theta_i) = 1/\{1 + \exp[-\alpha_i (\theta_i - \delta_i)]\} \text{ (two-parameter solution).}$$

The expectation of the unidimensional hypothesis implies that the number of persons in a sample of K individuals who would pass item i , item j , and both items i and j are respectively,

$$M_i = \sum_{k=1}^K p_{ik}, \quad M_j = \sum_{k=1}^K p_{jk} \quad \text{and} \quad A = \sum_{k=1}^K p_{ik}p_{jk}.$$

With these expected frequencies (M_i , M_j , A), a 2×2 contingency table could be constructed as shown in Table 2. With this contingency table, a tetrachoric correlation between items i and j was computed. After the tetrachoric correlation between all 56 items were computed, the eigenvalues for this correlation matrix were derived and compared with the eigenvalues from the obtained data.
