Supplemental Material S1. Description of measures.

Overview

Children completed individual measures and multiple subtests from commonly used standardized measures of language to assess six nominally distinct dimensions of oral language skills. When necessary, children of different ages completed subtests from different versions of these tests that were developed and normed with specific age groups. The younger group of children (i.e., children in preschool through second-grade groups) completed subtests from the Test of Language Development–Primary: Fourth Edition (TOLD-P4; Newcomer & Hammill, 2008). The older group of children (i.e., children in third-, fourth-, and fifth-grade groups) completed subtests from the Test of Language Development–Intermediate: Fourth Edition (TOLD-I4; Hammill & Newcomer, 2008). Children completed multiple subtests from the Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4; Semel, Wiig, & Secord, 2003) and the Comprehensive Assessment of Spoken Language (CASL; Carrow-Woolfolk, 2008). With the exception that some subtests were used with children slightly above or slightly below the intended age range for the subtest, all standard administration and scoring rules were followed.

Measure Selection and Classification

Decisions concerning classification of measures were made when the measures were selected for the study (i.e., the missing-by-design assessment approach used the classification of measures). The goal of the project design was to have at least three measures for each of six possible dimensions of language, and measures were selected to fit this project specification. Measures were selected by the first author based on a review of the measures' content, psychometric information reported in the test manuals, prior use in research, and consultation with colleagues who also conduct research on language-related topics. Prior to the analyses reported in this study, classification of measures was reviewed by the second author, a speech-language pathologist. Therefore, all decisions concerning the classification of measures to the specific dimensions were done a priori. Our classification of measures was consistent with the following definitions:

- Receptive measures were those that required a response indicating comprehension of an item and either did not require a verbally produced response or required selection of a verbal response from among a limited set of provided responses. For these measures, the response options to items were provided as a part of the test administration, and, therefore, responses (typically pointing to a picture that represented a spoken stimulus [i.e., word, phrase, or sentence]) were constrained to the options provided.
- Expressive measures were those that required a verbally produced response. For these measures, response options were not provided, and children had to generate a response to a stimulus, which could include a pictured representation or a spoken stimulus.

Measures were selected and classified as vocabulary, syntax, or listening comprehension based on the primary language dimension required by the expected response, and these decisions were guided by how subtests were conceptually and empirically classified in the larger test batteries from which they were obtained or how the subtest had been classified in prior published research.

- Receptive vocabulary measures were those that required responses indicating comprehension of spoken words (e.g., pointing to the correct picture representing a spoken word or words) or identifying semantically related words.
- Expressive vocabulary measures were those that required spontaneously producing a correct word for a pictured stimulus or describing why two words were semantically related.
- Depth of vocabulary measures were those that required a demonstration of word knowledge beyond labeling an item or selecting the related words. For instance, depth of vocabulary measures included those that required producing the opposite of a word, providing definitions of words, or identifying the superordinate category to which a series of words belong.
- Receptive syntax measures were those that required a response to an orally presented sentence, such as pointing to the correct picture that represented a sentence spoken by the examiner, selecting the one word from among several presented words that would correctly complete a sentence spoken by the examiner, or identifying whether or not a spoken sentence was grammatically correct.
- Expressive syntax measures were those that required the child to produce a spontaneous verbal response that resulted in a syntactically correct sentence, such as providing a word to complete a sentence, describing a picture, or combining sentences spoken by the examiner into a single sentence.
- Listening comprehension measures were those that required the child to listen to and comprehend spoken language, typically involving comprehension of both vocabulary and syntactic form. Although these measures were similar to those selected to represent either receptive or expressive syntax measures, the measures selected to represent this construct were either formally identified as a listening comprehension measure in the test from which they were obtained or have been frequently classified as listening comprehension measures in prior research. For instance, the Oral Comprehension subtest of the Woodcock–Johnson III Test of Achievement (WJ-OC; Woodcock, McGrew, & Mather, 2001) is the companion subtest to the Passage Comprehension subtest of the Woodcock–Johnson that is used to measure children's reading comprehension and is similar in form and response requirements but requires children to read the stimulus items.

Descriptions of Specific Measures Used in Study by Construct

Receptive Vocabulary

Receptive One-Word Picture Vocabulary Test (ROWPVT). The ROWPVT (Brownell, 2000) consists of 170 items that are intended to assess children's receptive vocabulary. The examiner shows the child a page with four pictures and says a word, and the child has to point to the picture that most closely matches the word spoken by the examiner. Basal—based on child age—and ceiling (i.e., eight consecutive incorrect responses) rules are used. Internal consistency (alpha) for this age group ranges from .95 to .98, and split-half coefficients, corrected for the full length of the test, range from .97 to .98. Criterion-related validity as measured by correlations with mean standard scores of the Receptive Language composite score of Clinical Evaluation of Language Fundamentals–Third Edition (CELF-3; Semel, Wiig, & Secord, 1995) is .82 and the Listening Comprehension subtest of Oral and Written Language Scales (1995) is .90.

Test of Language Development, Picture Vocabulary subtest (TOLD-PV). This measure consists of 34 items in the TOLD-P4 and 80 items in the TOLD-I4 that are used to assess children's listening vocabulary. The examiner says a series of two-word phrases one at a time (e.g., eating, utensil) and the child selects the best picture out of six that corresponds to the two-word phrase (i.e., fork). All children begin at the first item, and a ceiling is established after five consecutive incorrect responses. Internal consistency (alpha) for TOLD-P4 (ages 4–8 years) ranges from .81 to .87 and for TOLD-I4 (ages 8–10 years) ranges from .94 to .96. Test–retest reliability for this subtest of TOLD-P4 is .85 and for TOLD-I4 it is .87. Criterion-related validity of this measure based on correlations between standard score means of TOLD-P4 and Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV; Wechsler, 2004) is .53, and correlations between TOLD-I4 and Peabody Picture Vocabulary Test–Third Edition (PPVT-3; Dunn & Dunn, 1997) is .84.

Clinical Evaluation of Language Fundamentals, Word Classes—Receptive (I and II) subtest (CELF-WCR). This subtest consists of 20 items used to assess children's ability to recognize and identify words that share a semantic relationship. The examiner says three or four words (e.g., porcupine, quills, ball) and the child correctly responds by saying the two words that semantically go together (i.e., porcupine and quills). All children begin at the first item, and a ceiling is established after seven consecutive incorrect responses. Internal consistency (alpha) for ages 5–11 years range from .85 to .92. Validity, as reported in the manual according to standardized solutions on confirmatory factor analysis, indicate that the scores of children ages 5–7 years load onto the Language Content factor at .79.

Expressive Vocabulary

Expressive One-Word Picture Vocabulary Test, Third Edition (EOWPVT-3). The EOWPVT (Brownell, 2000) consists of 170 items used to assess children's ability to verbally label illustrations of objects, actions, or concepts. Basal—based on child age—and ceiling (i.e., six consecutive incorrect responses) rules are used. Internal consistency (alpha) for ages 3–10 years ranges from .95 to .97, and split-half reliability, corrected for full length of the test, ranges from .96 to .98. Criterion-related validity, as measured by correlations between mean standard scores and Peabody Picture Vocabulary Test–Revised (PPVT-R; Dunn & Dunn, 1981) is .82 and the Relational Vocabulary subtest of the Test of Language–Primary, Third Edition (TOLD-P:3; Newcomer & Hammill, 1997) is .75.

Clinical Evaluation of Language Fundamentals, Expressive Vocabulary subtest (CELF-EV). This subtest consists of 27 items used to assess the children's ability to label illustrations of people, objects, and actions. The examiner shows the child a picture in the stimulus book and asks the child a question to label the item or action (e.g., "What is this?" or "What is he/she doing?"). All children begin at the first item designated for their age group and a ceiling is established after seven consecutive incorrect responses. Internal consistency (alpha) for ages 5–10 years ranges from .80 to .85, and test–retest reliability for ages 6–10 years ranges from .87 to .91. Validity, as reported in the manual according to standardized solutions on confirmatory factor analysis, indicates that scores of children ages 5–7 years load onto the Language Content factor at .78 and at age 8 years of age at .38 and on the Receptive language Factor at .35, and at 9 years of age on the Language Content factor at .80.

Clinical Evaluation of Language Fundamentals, Word Classes--Expressive (I and II) subtest (CELF-WCE). The CELF-WCE subtest includes 21 (ages 5–7 years) to 24 (ages 8–10 years) items used to assess children's ability to recognize and identify words that share a semantic relationship. The examiner says three or four words (e.g., porcupine, quills, ball) and the child correctly responds by saying how two of the words are related (i.e., porcupines have quills). All

children begin at the first item, and a ceiling is established after seven consecutive incorrect responses. Internal consistency (alpha) for ages 5–11 years ranges from.85 to .92. Validity, as reported in the manual according to standardized solutions on confirmatory factor analysis, indicates that the Word Classes I scores of children 5–7 years of age load onto the Language Content factor at .79, Word Classes II scores of children 8 years of age load onto Language Content at .35 and Receptive Language at .35, and Word Classes II scores of children 9 and 10 years of age load onto Expressive Language and Language Content at .37 (for both factors) and .43 (for both factors), respectively.

Depth of Vocabulary

Comprehensive Assessment of Spoken Language, Antonyms subtest (*CASL-A).* The CASL-A (Carrow-Woolfolk, 2008) consists of 55 items used to assess children's word knowledge, retrieval, and production of antonyms. The examiner asks the child to "Tell me a word that means the opposite of ____." Basal—based on child age—and ceiling (i.e., five consecutive incorrect responses) rules are used. Internal reliability, using Rasch split-half method (odd/even), for 5–10 years of age ranges from .82 to .90, and test–retest reliability is .80 for children 5–7 years and .95 for children 8–11 years. Validity, as reported by the manual and as measured by correlation with the Word Classes and Relations subtest of the Test for Auditory Comprehension of Language–Revised (TACL-R; Carrow-Woolfolk, 1985), is .71.

Clinical Evaluation of Language Fundamentals, Word Definitions subtest (CELF-WD). This measure consists of 24 items used to measure the older children's ability to define words based on specific use in a given sentence. The examiner says a word (e.g., cafeteria) and then a sentence (i.e., My mother said, "Let's go to the theatre"), followed by a request for the child to define the word (i.e., theatre). All children begin at the first item and ceiling is established after seven consecutive incorrect responses. Internal consistency (alpha) for ages 10–11 years is .87, and test–retest reliability for ages 10–11 years is .86. Validity, as reported in the manual according to standardized solutions on confirmatory factor analysis, indicates that the scores of children ages 10–12 years load onto the Language Content factor at .84.

Test of Language Development, Relational Vocabulary subtest (TOLD-RV). The TOLD-RV consists of 34 items in the TOLD-P4 and 80 in the TOLD-I4 used to assess children's semantic knowledge in an associative task. For example, the examiner says three words (e.g., daisy, lily, rose) and the child identifies the semantic category to which the words belong (i.e., flowers). All children begin at the first item, and a ceiling is established after five consecutive incorrect responses for TOLD-P4 and three consecutive incorrect responses for TOLD-I4. Internal consistency (alpha) for TOLD-P4 (ages 4–8 years) ranges from .88 to .92, and test–retest coefficient is .82. Criterion-related validity, based on correlations between standard score means of TOLD-P4 and the Verbal Comprehension Index of WISC-IV, is .86. Internal consistency (alpha) for TOLD-I4 (ages 8–10 years) ranges from .91 to .92, and test–retest reliability is .80. Criterion-related validity, based on correlations between standard score means of TOLD-I4 and Verbal Comprehension Index of WISC-IV, is .78.

Receptive Syntax/Grammar

Comprehensive Assessment of Spoken Language, Grammaticality Judgment subtest (CASL-G). The CASL-G subtest consists of 60 items designed to assess children's ability to make judgements of the grammaticality of spoken sentences and to correct sentences that contain grammatical errors, such as errors in noun–verb agreement, number, tense (e.g., "She go to school"). Basal—based on child age—and ceiling (i.e., five consecutive incorrect responses) rules are used. Internal consistency, using Rasch split-half method (odd/even), for 7–10 years of

age ranges from .88 to .94. Test–retest reliability is .91 for children 8–11 years. Validity, as reported by the manual based on standard score correlations with the Oral Expression subtest of the Oral and Written Language Skills (OWLS; Carrow-Woolfolk, 1995), is .80.

Clinical Evaluation of Language Fundamentals, Sentence Structure subtest (CELF-SS). The CELF-SS subtest consists of 26 items designed to assess the children's ability to comprehend spoken sentences of increasing length and syntactic complexity. Children listen to a sentence spoken by the examiner and select one picture out of four that illustrates the referential meaning of the sentence. All children begin at the first item and a ceiling is established after five consecutive incorrect responses. Internal consistency (alpha) for ages 5;0–8;11 (years;months) years of age ranges from .64 to .76. Test–retest reliability for ages 6;0–7:11 ranges from .67 to .74. Validity, as reported in the manual according to standardized solutions on confirmatory factor analysis, indicates that the scores of children ages 5–7 years load onto the Language Content factor at .78 and at 8 years at .34.

Test of Language Development, Syntactic Understanding subtest (TOLD-SU). The TOLD-SU subtest consists of 30 items used to assess the younger children's comprehension of the meaning of sentences that have increasing syntactic complexity across items. The examiner reads a sentence, and the child points to one of three pictures in the stimulus manual that illustrates the meaning of the sentence (e.g., "He arrived at the store"). All children begin at the first item, and a ceiling is established after five consecutive incorrect responses. Internal consistency (alpha) for ages 4;0–7;11 years ranges from .80 to .90. Test–retest reliability is .80. Criterion-related validity, based on correlations between standard scores and the Verbal Comprehension Index of WISC-IV, is .86.

Test of Language Development, Morphological Comprehension subtest (TOLD-MComp). The TOLD-MComp subtest consists of 50 items used to assess the older children's ability to listen to sentences and judge them as having correct or incorrect grammar. Items include both syntactic and morphological errors such as noun–verb agreement, pronouns, and comparative/superlative affixes (e.g., "She were here yesterday"). The first 10 items are administered to all children regardless of whether responses are correct or incorrect, and a ceiling is established after that when three of five consecutive items are responded to with incorrect answers. Internal consistency (alpha) ranges from .96 to .97, and test–retest reliability is .89. Criterion-related validity, based on correlations between standard scores and the Verbal Comprehension Index of WISC-IV, is .72.

Morphological Syntax Awareness task (MSA). The MSA task (Connor & Lonigan, 2010) consists of 11 passages used to assess children's ability to perform morphological-choice tasks. Each passage includes between 8 and 11 cloze items (103 cloze items total) that require children to select the correct word. The examiner verbally presents each sentence and the possible choices to the child, and the child says the correct choice (e.g., Examiner: "Fresh water [*freeze freezes freezing*] at 32 degrees Fahrenheit." Child: "freezes"). Basal—based on child grade—and ceiling (i.e., six items in any one story answered incorrectly) rules are used. Internal consistency (alpha) is .98 for children in preschool to second grade and .98 for children in third to fifth grade.

Expressive Syntax/Grammar

Comprehensive Assessment of Spoken Language, Syntax Construction subtest (CASL-SC). The CASL-SC subtest consists of 56 items used to assess the children's ability to formulate and express sentences using a variety of morphosyntactic rules. While showing the child a picture in the stimulus manual, the examiner either reads one to three sentences with a phrase missing from the final sentence (e.g., "This is a small fish. This is a ___" [big fish]) or gives the child a directive to say something related to the picture (e.g., "The man found the hat. Where was it?"

[on the table]). Basal—based on child age—and ceiling (i.e., five consecutive incorrect responses) rules are used. Internal consistency (alpha) for ages 5–10 years ranges from .82 to .90, and test–retest reliability for ages 5–7 years is .79 and for ages 8–11 years is .74. Validity, as measured by correlations with the Word Classes and Relations subtest of the TACL-R, is .65.

Clinical Evaluation of Language Fundamentals, Formulated Sentences subtest (CELF-FS). The CELF-FS subtest consists of 28 items used to assess the children's ability to formulate complete, semantically, and grammatically correct sentences of increasing complexity using given word(s) and a specific context based on an illustration in the stimulus manual. The examiner says a word and shows the child a picture and asks the child to make a sentence about the picture using the word provided. Basal—based on grade level—and ceiling (i.e., five consecutive incorrect responses) rules are used. Internal consistency (alpha) for ages 5–11 years ranges from .76 to .86, and test–retest reliability for ages 6–10 years ranges from .74 to .81. Validity, as reported in the manual according to standardized solutions on confirmatory factor analysis, indicates that the scores of children ages 5–7 years load onto the Language Content factor at .78 and at 8 years at .34.

Test of Language Development, Morphological Completion subtest (TOLD-MCmpl). The TOLD-MCmpl subtest consists of 38 items used to assess the younger children's ability to recognize, understand, and use common English morphemes in a cloze task. The examiner reads sentences (e.g., "Michael saw a fox. Adam saw a fox. They saw two ___"), and the child says the word to complete the final sentence using the correct morphological form (i.e., "foxes"). All children begin at the first item, and a ceiling is established after five consecutive incorrect responses. Internal consistency (alpha) for ages 4–8 years ranges from .90 to .94, and test–retest reliability is .82. Criterion-related validity, based on correlations between standard score means of TOLD-P4 and the Verbal Comprehension Index of WISC-IV, is .55.

Test of Language Development, Sentence Combining subtest (TOLD-SC). The TOLD-SC subtest consists of 30 items used to assess the older children's spoken syntactic ability by having them combine two or more simple sentences into one compound or otherwise syntactically complex sentence. The examiner reads the sentences (e.g., "Amanda likes singing"; "Amanda likes dancing"), and then asks the child to combine the sentences into one sentence (i.e., "Amanda likes singing and dancing"). All children begin at the first item, and a ceiling is established after three consecutive incorrect responses. Internal consistency (alpha) for ages 8–10 years ranges from .91 to .93, and test-retest reliability is .88. Content-related validity between this subtest and standard score means of TOLD-P4 is .53.

Spoken Morphological Awareness Task (SMA). The SMA task (Apel, Diehm, & Apel, 2013) consists of 40 items used to assess the children's ability to apply correct morphology (i.e., affixes) in an orally administered cloze task. The examiner says a prompt word followed by a sentence with a missing word (e.g., "cookie": "Please give each child two ____"), and the child is asked to say the missing word using the correct morphological form (i.e., "cookies"). All children begin at the first item, and a ceiling is established after three consecutive incorrect responses. Internal consistency (alpha) is .85.

Listening Comprehension

Clinical Evaluation of Language Fundamentals, Concepts and Following Directions subtest (CELF-CFD). The CELF-CFD subtest consists of 54 items used to assess children's abilities to interpret and follow directions of increasing length and complexity (e.g., names, characteristics, order of mention) using logical operations (e.g., before/after, tallest). Basal—based on child age—and ceiling (i.e., seven consecutive incorrect responses) rules are used. Internal consistency (alpha) ranges from .73 to .92, and test-retest reliability ranges from .67 to .88. Validity, as

reported in the manual according to standardized solutions on confirmatory factor analysis indicate that the scores of children ages 5–7 years load onto the Language Content factor at .80 and at 8 years at .54 and Language Memory at age 9 years at .34 and at age 10 years at .55.

Oral and Written Language Scales, Listening Comprehension subtest (OWLS-LC). The OWLS-LC subtest (Carrow-Woolfolk, 1995) consists of 111 items used to assess the children's ability to listen to and comprehend spoken language. Items are presented verbally and the child responds by pointing at one of four illustrations in the stimulus manual (e.g., "Show me the picture of the man standing between the boy and the dog"). Basal—based on child grade—and ceiling (i.e., five of seven consecutive items are answered incorrectly) rules are used. Internal consistency (alpha) for ages 4–10 years ranges from .75 to .89, and test–retest reliability for children ages 4–6 years is .80 and for children ages 8–11 years is .73. Criterion-related validity, as measured by correlations with Peabody Picture Vocabulary Test–Revised (PPVT-R; Dunn & Dunn, 1981) is .61 and with the Total Language Composite Score of CELF–Revised (CELF-R; Semel, Wiig, & Secord, 1987) is .84.

Woodcock-Johnson III Test of Achievement, Oral Comprehension subtest (WJ-OC). The WJ-OC subtest (Woodcock et al., 2001) consists of 34 items that assess children's ability to comprehend short, audio-recorded or orally presented passages to complete an oral cloze task. The child hears a sentence followed by two beeps (for audio-recorded presentation), at which time the child provides a single word to finish the sentence. Basal—based on child grade—and ceiling (i.e., six consecutive items are responded to incorrectly) rules are used. Median internal consistency for children ages 5–19 years is .80. One-year, test–retest reliability for children ages 4–7 years of age is .82 and for children ages 8–10 years of age is .74.

References

Apel, K., Diehm, E., & Apel, L. (2013). Using multiple measures of morphological awareness to assess its relation to reading. *Topics in Language Disorders, 33*(1), 42–56. https://doi.org/10.1097/TLD.0b013e318280f57b

Brownell, R. (2000a). *Receptive One-Word Picture Vocabulary Test*. Novato, CA: Academic Therapy Publications, Inc.

Carrow-Woolfolk, E. (1985). *Test for Auditory Comprehension of Language–Revised*. Allen, TX: DLM Teaching Resources.

Carrow-Woolfolk, E. (1995). *Oral and Written Language Scales*. Circle Pines, MN: American Guidance Services.

Carrow-Woolfolk, E. (2008). *Comprehensive Assessment of Spoken Language*. Los Angeles, CA: Western Psych.

Connor, C. M., & Lonigan, C. J. (2010). *Morphological Syntax Maze Task*. Tallahassee, FL: Florida State University, Reading for Understanding Network.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test–Revised*. Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test–Third Edition*. Circle Pines, MN: American Guidance Service.

Hammill, D. D., & Newcomer, P. L. (2008). *Test of Language Development–Intermediate: Fourth Edition*. Austin, TX: Pro-Ed.

Newcomer, P. L., & Hammill, D. D. (1997). *Test of Language Development–Primary: Third Edition*. Austin, TX: Pro-Ed.

Newcomer, P. L., & Hammill, D. D. (2008). *Test of Language Development–Primary: Fourth Edition*. Austin, TX: Pro-Ed.

Semel, E. M., Wiig, E. H., & Secord, W. A. (1987). *Clinical Evaluation of Language Fundamentals–Revised*. San Antonio, TX: The Psychological Corporation.

Semel, E., Wiig, E., & Secord, W. (1995). *Clinical Evaluation of Language Fundamentals–Third Edition (CELF-3)*. San Antonio, TX: The Psychological Corporation.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4)*. San Antonio, TX: Pearson.

Wechsler, D. (2004). *Wechsler Intelligence Scale for Children–Fourth Edition: Integrated (WISC®-IV Integrated)*. San Antonio, TX: Pearson.