MONOGRAPHS OF THE SOCIETY FOR
RESEARCH IN CHILD DEVELOPMENT

IV. NIH TOOLBOX COGNITION BATTERY (CB): MEASURING LANGUAGE (VOCABULARY COMPREHENSION AND READING DECODING)

Richard C. Gershon, Jerry Slotkin, Jennifer J. Manly, David L. Blitz, Jennifer L. Beaumont, Deborah Schnipke, Kathleen Wallner-Allen, Roberta Michnick Golinkoff, Jean Berko Gleason, Kathy Hirsh-Pasek, Marilyn Jager Adams, and Sandra Weintraub

ABSTRACT Mastery of language skills is an important predictor of daily functioning and health. Vocabulary comprehension and reading decoding are relatively quick and easy to measure and correlate highly with overall cognitive functioning, as well as with success in school and work. New measures of vocabulary comprehension and reading decoding (in both English and Spanish) were developed for the NIH Toolbox Cognition Battery (CB). In the Toolbox Picture Vocabulary Test (TPVT), participants hear a spoken word while viewing four pictures, and then must choose the picture that best represents the word. This approach tests receptive vocabulary knowledge without the need to read or write, removing the literacy load for children who are developing literacy and for adults who struggle with reading and writing. In the Toolbox Oral Reading Recognition Test (TORRT), participants see a letter or word onscreen and must pronounce or identify it. The examiner determines whether it was pronounced correctly by comparing the response to the pronunciation guide on a separate computer screen. In this chapter, we discuss the importance of language during childhood and the relation of language and brain function. We also review the development of the TPVT and TORRT, including information about the item calibration process and results from a validation study. Finally, the strengths and weaknesses of the measures are discussed.

In this chapter, we discuss language as represented by measures of vocabulary comprehension and reading decoding (in both English and Spanish) in the Cognition Battery (CB) of the NIH Toolbox for the Assessment of Neurological and Behavioral Function (Gershon et al., 2010).

Corresponding author: Sandra Weintraub, Cognitive Neurology and Alzheimer's Disease Center, Northwestern Feinberg School of Medicine, 303 E Chicago, IL 60611, email: sweintraub@northwestern.edu

Subdomain Definition

Language is a shared symbol system that facilitates communication, categorization, and thought (Pinker, 2000). The simplest definition of language is that it is a means of communication consisting of all the words used by a community and the rules for varying and combining them. Language users can express the full range of their experience by joining words into clauses, sentences, and connected discourse (Gleason & Ratner, 2009). Language can be spoken or written, or it can be transmitted gesturally, as in sign language. Though language does not require audition and speech (as in sign language), important language abilities include auditory comprehension, speaking, naming, reading, and writing. Language is hierarchically organized, and composed of a number of subsystems. These include phonology, morphology, syntax, the lexicon and semantics, pragmatics, and discourse—components that have been linked to constituents within a large-scale neuroanatomical network primarily in the left cerebral hemisphere (Price, 2000).

Communication via spoken and written language promotes the transmission of culture, societal values, and history. In an ever more literate world, language skills are important predictors of daily functioning and health (Burton, Strauss, Hultsch, & Hunter, 2006). Language is commonly assessed through receptive vocabulary (comprehension), expressive vocabulary and production, object naming, speech fluency, reading, and writing.

For purposes of the NIH Toolbox CB, it was desirable to establish quick measures, available for researchers' use at low or no cost, that would correlate highly with overall cognitive functioning and with success in school and work (Kastner, May, & Hildman, 2001; Schmidt & Hunter, 2004). Vocabulary comprehension was chosen as the primary language measure after much deliberation and with the full recognition of the equivalent importance of grammatical proficiency for development and growth (Gleason & Ratner, 2009; Hirsh-Pasek & Golinkoff, 1996). Vocabulary knowledge is of particular interest because it has a high association with general measures of "intelligence," or the "*g*" factor (Cattell, 1987) and with success in school and work (Kastner et al., 2001; Schmidt & Hunter, 2004).

The TORRT, the second language measure, is a proxy for a broad range of cognitive, educational, and socioeconomic factors. The ability to pronounce low-frequency words with irregular orthography has also been used as an estimate of overall intelligence (Grober & Sliwinski, 1991). The TORRT measures the accuracy of pronouncing single printed words and of naming or recognizing single letters. In healthy individuals, single-word reading tasks reflect (1) level of exposure to written text/material; (2) whether one's environment provided a context in which

to develop basic and complex reading skills; (3) specific cognitive skills needed to develop decoding, such as phonological processing and working memory; and (4) general cognitive ability, since more "able" individuals are expected to be exposed to a greater volume and higher complexity of written stimuli.

Importance During Childhood

Language consists of a complex system of rules that is acquired relatively effortlessly by infants. Children across a wide range of different environments and cultures learn to understand and use language in a remarkably short period of time. Language has a biological basis. It depends both on skills specific to language (e.g., the perception of phonemes) and general cognitive skills (e.g., categorization and memory) (Kuhl & Rivera-Gaxiola, 2008). Comprehension of single words is a fundamental language skill that infants begin to acquire well before children speak (Kuhl, 2004). Infants typically have a repertoire of about 50 words they can understand before age 1 year (Fenson et al., 1994) and typically begin to produce single word utterances around their first birthdays. Typically, they begin to combine words to form brief sentences by the age of 2 years. Initially syntax is highly simplified, but over time develops to include more complex constructions. For example, reversible passive sentences like "Bart was seen by Marge" are not correctly comprehended at a 90% level until age 9 (Hirsch & Wexler, 2007). Acquisition of basic letter and word recognition skills typically begins in preschool and is typically well anchored by second grade. Over the ensuing school years, through instruction and practice in reading and writing, children's ability to read and spell words continues to grow and to become richly interconnected with their development of vocabulary and grammatical knowledge.

Writing is the last major language skill to emerge in early childhood. In young children, measures of language function need to capture proficiency in comprehension, naming, and generating and interpreting simple sentences. As the fundamental skills become more established in late childhood and early adolescence, vocabulary increases and language becomes the primary medium for establishing and accessing "semantic memory"—our storehouse of information and facts. In young adulthood and into old age, vocabulary and semantic memory are referred to as "crystallized capacities" that are relatively resistant to the effects of aging and neurological disease (see Heaton et al., Chapter 8, this volume; National Research Council Committee on the Prevention of Reading Difficulties in Young Children, Snow, Burns, Griffin, & National Research Council Commission on Behavioral Social Sciences Education, 2002; Sternberg, 2004). Reading encompasses phonological, orthographic, and semantic processing, and several models have been proposed to account for reading ability (National Early Literacy Panel & National Center

for Family Literacy, 2008; National Research Council Committee on the Prevention of Reading Difficulties in Young Children et al., 2002).

To build rapid and functional word recognition skills, beginning readers must first develop basic language and decoding skills (National Reading Panel & National Institute of Child Health and Human Development, 2000). Reading also demands a modicum of world experience and vocabulary knowledge so that children can begin to use reading to learn, which typically takes place at the third grade level (Dickinson, Golinkoff, & Hirsh-Pasek, 2010).

The process of sounding out words results in neural associations between letters or graphemes and the phonemes they represent. As readers repeatedly encounter common sequences of letters, these associations become extended and differentiated, linking larger spelling patterns with larger phonological units and resulting in what is known as *decoding automaticity*—the ability to pronounce even new and less familiar words quickly, easily, and accurately provided that they are regularly spelled. Irregular words (e.g., "one," "two," "colonel," and "island"), whose spelling-sound correspondences do not conform to the norms of the language, become set off and learned as wholes. Repeated experience reading and decoding sequences of letters that correspond to the same "phoneme blend" allow the reader to build a knowledge base that can be applied to correctly pronounce words. However, irregular words, whose pronunciations do not conform to the rules, require different lexical routes for correct pronunciation. The ability to read regularly versus irregularly spelled words is differentially affected among people with acquired dyslexia due to brain trauma (Rapcsak, Henry, Teague, Carnahan, & Beeson, 2007; Ziegler et al., 2008).

Developmental disorders of language and communication (e.g., autism, dyslexia) and limited opportunities to acquire literacy in childhood have a significant impact on academic achievement and life adaptation in developed countries. Scores on language measures can predict occupational attainment and performance (Schmidt & Hunter, 2004). Many acquired conditions can affect language in adulthood, including stroke and Alzheimer's disease (Kastner et al., 2001).

There is evidence that reading disability may be under-identified in children if measures of reading fluency, such as naming speed, are not included (Meisinger, Bloom, & Hynd, 2010). Single-word reading recognition tasks are strong predictors of health and cognition outcomes across the lifespan. Poor health literacy (literacy skills related to health information, such as reading prescription bottles, appointment slips, or medical education brochures) is one critical factor in health outcome, especially in older adults (Wolf, Gazmararian, & Baker, 2005). Performance on single-word reading recognition tasks is also useful as a general estimate of reading level and quality of education (Manly, Byrd, Touradji, & Stern, 2004; Manly et al., 1999).

Relation of Subdomain With Brain Function

From a clinical perspective, language capabilities are sometimes divided into two broad categories: receptive language and expressive language. Receptive language involves the comprehension of language. Expressive language involves the production of language and includes skills such as naming, speaking, and repeating. Although this is a convenient way to divide language functions for the clinician, on a cognitive systems level, language is not represented in that manner. Instead, the psycholinguistically supported subcomponents of language are its phonology (or basic sound system), morphology (structure of words and their modifiers), lexicon (the dictionary of all words in any given language), syntax (the rules of grammar that link words together), and semantics (meaning) (Gleason, 1997). Speakers who have communicative competence must also be aware of discourse rules that govern the way that utterances may be combined, as well as pragmatic rules for appropriate language in social settings. It is these components of language that are represented in the brain in the context of a language system, rather than merely the dichotomy of input and output capabilities.

Early evidence supporting that the left cerebral hemisphere of the brain is the major contributor to language functions came from the study of patients who had suffered strokes in various regions of the left perisylvian area. Classical models of aphasiology were based on this type of evidence. Different aphasia subtypes correspond to the loss of one or more components of language. Thus, patients with strokes can be agrammatic, having difficulty comprehending and producing small grammatical features of language while others can produce normal grammar but have difficulty accessing nouns and verbs. In more recent years, studies of nonbrain-injured individuals using functional neuroimaging have affirmed the relative modularity of language components (Caplan & Hildebrandt, 1988; Friederici, Rüschemeyer, Hahne, & Fiebach, 2003; Price, 1998).

To assess the language subdomain, we developed the TPVT and the TORRT. To develop these new measures and assess their psychometric properties, we adhered to stringent development processes and utilized state-of-the-art psychometric approaches, in an effort to assess whether researchers will obtain stable and valid scores when using these measures. A detailed description of these processes is provided below.

METHOD

Participants

Two samples of participants were used in the preliminary item calibration for the TPVT. The goal of the calibration sample was to calculate item

response theory parameters for the item bank. Matching participant vocabulary ability with item difficulty was of primary importance in calibration accrual. The first sample contained 4,703 participants ranging in age from 3 to 69 ($n = 3{,}190$ children ages 3–17, Mean = 9.41, Female = 48.1%; $n = 1{,}513$ adults ages 18–69, Mean = 25.76, Female = 61.1%), with education for adults spread relatively evenly from completion of 10th grade through graduate/doctorate level. Participants were recruited via an online panel company (a company that specializes in procuring subjects for online surveys and test administrations), and participants were paid to take the test online. Parents of children under age 7 years were given specific instructions about how to administer the test to their children, what their children would be asked to do, and how to help children maintain attention and complete the tests without providing material assistance. Participants were administered one of 21 test forms that were believed to closely match their likely ability level (based on age for those 17 years and under and based on level of education for participants 18 years and above).

Unlike the TPVT, online calibration testing for the TORRT was not an option, given the requirement for one-on-one administration (after the participant reads the item on the screen the test administrator scores the item right or wrong). Instead, 146 participants for the initial item calibrations were recruited from the general population from four geographic locations to test at five sites associated with some of our academic collaborators: West Orange, NJ; Minneapolis, MN; Atlanta, GA; Evanston, IL; and Chicago, IL. The data from these participants were used in the initial item calibrations. Data from the validation study (see Weintraub et al., Chapter 1, and Table 3, this volume, for sample composition) were combined with those from the data collection described above. This merged sample was then used to recalibrate items prior to use in the norming study.

The sample from which the validation results discussed in this chapter were derived is described in Weintraub et al. (Chapter 1, this volume; Table 2).

Measures

Toolbox Picture Vocabulary Test (TPVT)

For the TPVT, single words are presented via an audio file, paired simultaneously with four images of objects, actions, and/or depictions of concepts (e.g., ball, running, friendship; see Figure 6). The participant is asked to select the picture whose meaning most closely corresponds with the spoken word. Participants are permitted as much time as necessary to complete their responses. Because the test does not require reading or writing, the test design removes the literacy load for children and for those who struggle with literacy skills. Further, the test does not require a spoken response, making it particularly suitable for children.

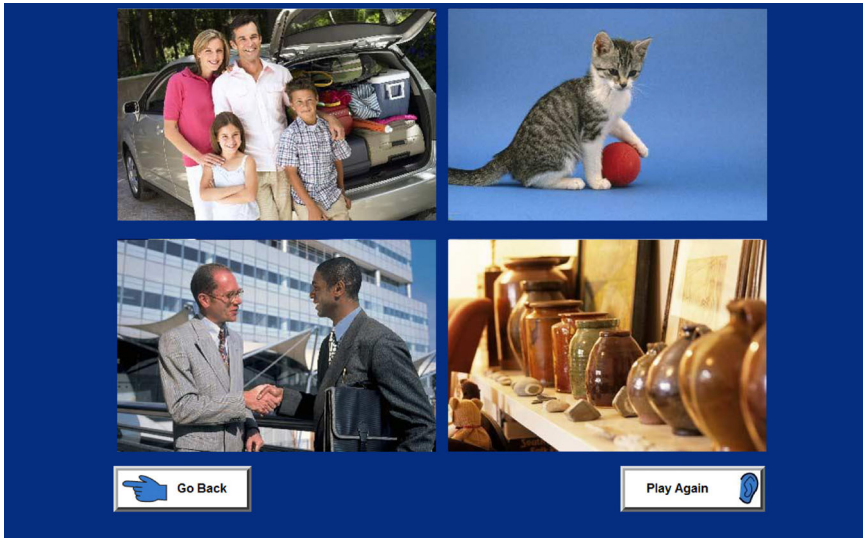

FIGURE 6.—Sample item from the Toolbox Picture Vocabulary Test ("Kin").
Note. Photo Credits (clockwise from upper left): Flying Colours Ltd/Photodisc/Getty Images; David De Lossy/Photodisc/Getty Images; Stockbyte/Getty Images; Andy Sotiriou/Photodisc/Getty Images.

To select words, initial candidate words were based on difficulty, from previously field-tested and calibrated items made available by the Johnson-O'Connor Research Foundation (Gershon, 1988). Additional items were drawn from the *Living Word Vocabulary* (Dale & O'Rourke, 1976) and *Children's Writer's Word Book* (Mogilner, 1992) based on the need for age, difficulty level, and frequency. Candidate words were then reviewed against the University of Western Australia MRC Psycholinguistic Database (UWASP, 2011) to evaluate how well they could potentially be translated into a photograph. Words with low imageability were dropped from the list of candidates. The list of potential words was then reviewed by experts (a diverse group of pediatric and geriatric professionals with "language expertise") and pruned accordingly based upon their feedback.

To create distractors for each item (word), four answer options were written, and these functioned as requirements for photographs that were eventually selected. The distractors all needed to be plausible but incorrect and had to have high imageability as described above, so that photos could eventually be selected as the corresponding answer choices. A senior internal content team then reviewed all items; modifications were made resulting in

the deletion of a small number of additional items. Items were then re-reviewed by the language experts for content and sensitivity, with specific instruction that these descriptions would be the basis for photo selection and then subsequent reviews. Items and distractors were further modified or dropped based on expert feedback, and prepared for the photo selection process.

Color photographs for the four options for each item of the TPVT (one correct answer and three distractors) were selected from the Getty Images library of millions of high-quality, photographic images. Initially, Getty staff provided 4–10 suggested images for each of the item options. The senior content team at Northwestern University, together with the language experts, were then trained in photo selection, and an extensive review process was implemented to ensure appropriate photos were selected. The selection process included a review by a multicultural group of experts empaneled to view all items for fairness and sensitivity. Items were edited or dropped based on the group's feedback. In many cases, the reviewers went back to the database and searched for additional photographs that would better meet the needs for any given item. Photos selected were then edited professionally where needed, primarily to make photos more consistent in background and orientation within a given item and to remove extraneous information in selected photos.

Item Calibration

In the preliminary item calibration for TPVT, each participant was administered 40–60 items from the pool of 625 items (children under age 8 years were generally administered the shorter, 40-item forms). Forms utilized common items (each form shared 50% of the items with its adjacent form) to allow for successful equating across forms. Each item received approximately 200 unique administrations to participants. Items were scored right/wrong and were calibrated using the one-parameter/Rasch Item Response Theory model (Rasch, 1960) as analyzed using Winsteps (Linacre, 2005).

Based on the initial analysis, items were reordered by difficulty and misfitting items were removed, leaving 602 items from which to construct an initial item bank to enable initial computer adaptive testing (CAT) for use in the validation study. A fixed-length 25-item CAT was constructed. A fixed-length strategy was used over a typical variable length CAT with a standard-error cutoff. This strategy "forced" participants to take a total of 25 items (versus fewer, which might be expected with the variable length test) in order to oversample items, allowing for the accumulation of additional data to refine the item calibrations. The difficulty of each successive item presented is based on the current estimate of the participant's ability level, as estimated by their responses to the previously administered items on the test. Items were

administered to match each participant's ability with item difficulty, with the consequence that each participant was correct on approximately 50% of the items. (Final target percentage correct will be adjusted based on norming to enable younger participants to have a higher "success" experience, to improve motivation.) The average administration time was about 5 min.

Toolbox Oral Reading Recognition Test

For the TORRT, a word or letter is presented on the computer screen, and the participant is asked to read it aloud. Participants are permitted as much time as necessary to complete their responses. Responses are recorded as correct or incorrect by the examiner, who views accepted pronunciations on a separate computer screen. A sample TORRT item is shown in Figure 7, with the participant screen shown in Panel (a) and the examiner screen shown in Panel (b) (Toolbox examiners must be trained on correct word pronunciation prior to administering this measure). For "prereaders" and those with low literacy levels, letters and other multiple-choice "prereading" items are presented, making the test as accessible as possible for young children. "Ceiling" rules were also implemented to minimize frustration, especially for early and prereaders.

Initial candidate words were drawn from the University of Western Australia MRC Psycholinguistic Database (UWASP, 2011). A variety of search criteria were applied, including frequency in the language, complexity of letter-sound relations, orthographical typicality, age of acquisition rating, number of syllables, and number of phonemes. Individual letters of the alphabet were later added to this list to enable assessment of emerging reading ability.

The Kucera and Francis rating (Kucera & Francis, 1967), which is closely correlated with Brown Verbal Frequency (Brown, 1984), was the frequency statistic that was present most often in the database. Many words had no Kucera and Francis frequency information, however, which indicates that they were not present in that corpus (low frequency). The letters and words were selected using the following guidelines: (1) letters could be roughly matched in relative frequency with another letter in the alphabet; (2) words had between 2 and 14 letters; (3) within words with 2–4 letters, emphasis was placed on including frequent words (for words with 5 or more letters, a few common words were included but this was not emphasized); (4) among words with 5 or more letters, a sample of words with low Kucera and Francis frequency was selected; (5) among words with 4 or more letters, a sample of words with an irregular orthography to phonology match, regardless of frequency, was selected; (6) words that appeared to be technical terms (e.g., medical or zoological terms) were eschewed; (7) words with many different acceptable pronunciations were avoided for ease of scoring.

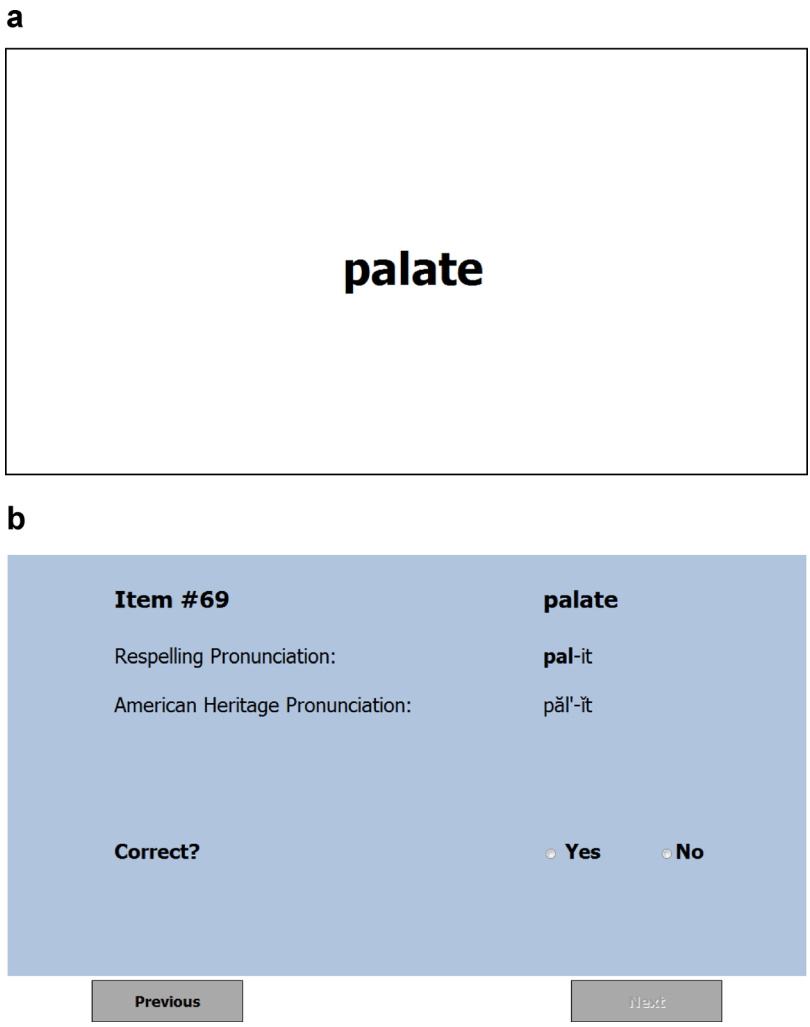

FIGURE 7.—Sample item from the Toolbox Oral Reading Recognition Test. The participant's screen is shown in Panel a, and the examiner's screen is shown in Panel b.

National experts in reading as well as a diverse group with expertise in language and pediatrics or geriatrics reviewed the initial word list. Words were eliminated or added based on experts' feedback, resulting in an initial item pool of 268 words and 21 letters and prereading items. The prereading items were developed in a multiple-choice format in which the respondent is asked, for example, to identify the letter when shown three nonletter symbols and a letter, or to identify a specific letter (e.g., "B") when three other letters (or nonletter symbols) are shown.

A subset of items with an expected broad range of difficulty was pilot tested to determine what format to use for the test. Two 50-item forms designed to be parallel in length and frequency of words were created; one form was administered one item per screen, and the other had 5–6 items per screen. Each form took 5 min or less for participants. Given that both forms took a similar length of time, the format with one item per screen (easier for the examiner to score and less cluttered for the participant), was selected for the test.

Item Calibration

For the item calibration, a 9-item screener was used to determine which test form the respondent would receive. Four test forms were created from 280 items, with the following numbers of items per form: Form 1 (70 items); Form 2 (101 items); Form 3 (120 items); and Form 4 (125 items). Based on rough preliminary information, it was expected that Form 1 would be easier than Form 2, which was expected to be easier than Form 3, etc. Each form had approximately one-third common items to allow for successful calibration.

The 9-item screener was administered to participants aged 8 years and older; participants aged 3–7 years did not receive the screener and immediately proceeded to Form 1. Based on performance on the routing form, older children (and adults) received one of the four forms. Prior to administration, brief instructions were read to the participant. For the screener and all forms, items were presented in dual-screen mode, whereby the participant was presented the word on one screen and the examiner was presented with a scoring template and phonetic key on the other screen. Participants attempted items until they either finished the prescribed number of items for their form or they mispronounced 10 words in a row (the discontinue rule used for calibration).

For the validation study, each participant was again administered one of four forms as described above, so that a fuller calibration of the 289 items could be achieved. Results were combined with the previous data set for calibration, and were analyzed separately for the purposes of assessing convergent and discriminant validity. The average administration time was 6 min.

Validation Measures

Peabody Picture Test-4th Edition (PPVT-IV) (Dunn & Dunn, 2007)

The PPVT-IV is a test of receptive vocabulary that is individually administered and provides an estimate of verbal ability or scholastic aptitude. The test is given verbally and takes 10–15 min to administer. For its administration, the examiner presents a series of pictures (four to a page) to

the test taker. Stating a word describing one of the pictures, the examiner asks the participant to point to or say the number of the picture they feel best corresponds to the word. The total score can be converted to a percentile rank, mental age, or a standard deviation IQ score. The test is available in two parallel forms of 228 items each. Internal consistency coefficients across ages are .94 for each alternate form; test–retest reliability is .93. The PPVT-IV was used as a measure of convergent validity for the TPVT.

Wide Range Achievement Test Version 4-Reading Subtest (WRAT-IV) (Wilkinson & Robertson, 2006)

The WRAT-IV is an individually administered test in which test takers are asked to name letters and read aloud words out of context. The words are listed in order of decreasing familiarity and increasing phonological complexity. Median internal consistency coefficients across ages for each of the alternate forms used individually range from .87 to .96. Alternate-form immediate retest reliability coefficients range from .78 to .89 for an age-based sample. Validity evidence for the WRAT-IV is derived from the content and structure of the test battery, studies with special groups, and correlations with other widely used achievement and cognitive ability measures. Standard scores, percentiles, stanines, normal curve equivalents, and Rasch scaled scores are provided for the WRAT-IV. Note that although the WRAT-IV is not ordinarily administered below age 5, we did so for comparison purposes and correlated the raw scores for each measure. The WRAT-IV was included primarily to serve as a measure of convergent validity for the TORRT.

Brief Visuospatial Memory Test-Revised (BVMT-R Total Recall) (Benedict, 1997)

The BVMT-R is designed to measure visuospatial memory. Participants view six geometric figures on a page and are asked to draw as many of the figures as possible from memory in their correct location, after the figures are removed from view. Reliability coefficients range from .96 to .97 for the three Learning trials, .97 for Total Recall, and .97 for Delayed Recall. Test–retest reliability coefficients range from .60 for Trial 1 to .84 for Trial 3. The BVMT-R correlates most strongly with other tests of visual memory and less strongly with tests of verbal memory. The BVMT-R was included to serve as an assessment of discriminant validity for both CB Language tests and was administered only to ages 8 and up.

Rey Auditory Verbal Learning Test (RAVLT) (Rey, 1958)

The RAVLT starts with a list of 15 words, read aloud by the examiner at the rate of one word per second. The participant's task is to repeat as many words as possible, in any order. This procedure was carried out a total of three times in comparison with the usual five trials conducted in the standard administration. The RAVLT was also included as a measure of discriminant validity for the TPVT and was administered to ages 8 years and up.

Analyses

Normalized scaled scores were used for all analyses. These scores were created by first ranking the test scores, next applying a normative transformation to the ranks to create a standard normal distribution, and finally rescaling the distribution to have a mean of 10 and a standard deviation of 3. Pearson correlation coefficients between age and test performance were calculated to assess the ability of the NIH Toolbox language tests to detect cognitive developmental growth during childhood. Intraclass correlation coefficients (*ICC*) with 95% confidence intervals were calculated to evaluate test–retest reliability. Convergent validity was assessed with correlations between each CB measure and an established measure of the same construct (PPVT-IV for Vocabulary and WRAT-IV for Reading). Evidence of discriminant validity consisted of lower correlations with selected measures of a *different* cognitive construct (BVMT-R and RAVLT) for both TPVT and TORRT.

RESULTS (TPVT)

Eight children did not successfully complete the TPVT for reasons such as lack of attention or alertness or general noncompliance.

Age Effects

Age was significantly correlated with the TPVT score ($n = 200$; $r = .81$; $p < .001$), as well as the PPVT-IV ($n = 201$; $r = .88$; $p < .001$). A quadratic model provided the best fit of the data, with $R^2 = .67$. Pairwise comparisons between age groups are reported in Appendix. In the subset of participants age 3–6 years, the correlation between TPVT and age was .42 ($n = 112$, $p < .001$). In participants age 8 to 15 years, the correlation was .57 ($n = 88$, $p < .001$). Figure 8 shows TPVT scores as a function of age. It should be noted that the TPVT and PPVT-IV scores closely mimic each other at every age level.

Test–Retest Reliability

The test–retest reliability of the TPVT was $ICC = .81$ ($n = 66$; 95% confidence interval: .71, .88). Reliability of the PPVT-IV in our sample was $ICC = .96$ ($n = 65$; 95% confidence interval: .94, .98).

Effect of Repeated Testing

Practice effects were computed as the difference between test and retest normalized scaled scores, with significance of the effect being tested with t

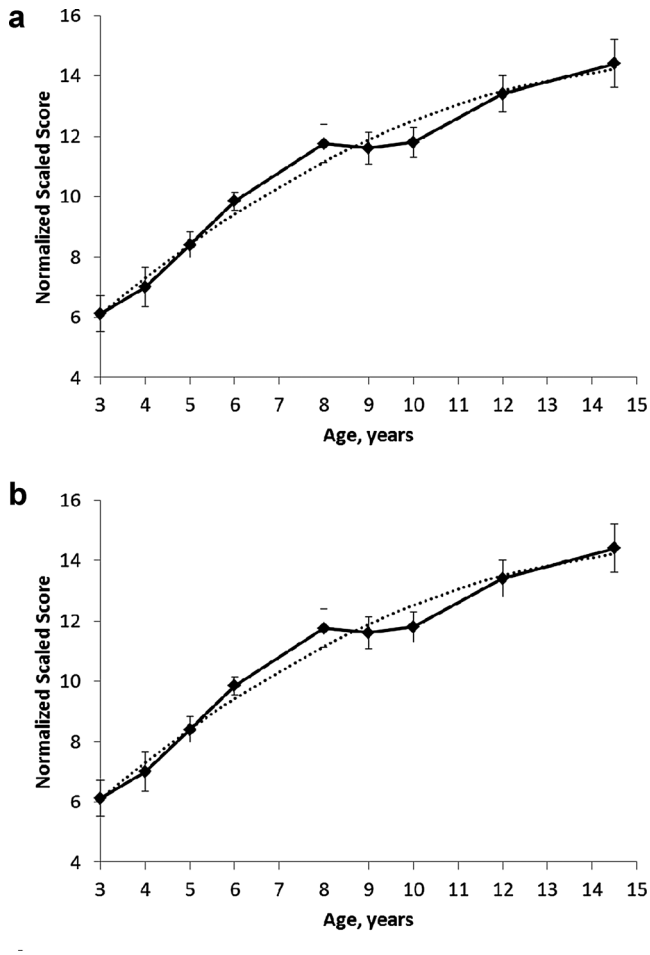

FIGURE 8.—Normalized scaled scores on the Toolbox Picture Vocabulary Test (A) and the Toolbox Oral Reading Recognition Test (B) across age groups. Error bars are ±2 standard errors. Best-fitting polynomial curves are also shown (see text).

tests for dependent means. For the total child group (ages 3–15 years, $n = 66$), the TPVT showed no practice effect over an average 2-week test–retest interval: *mean practice effect* = .1, $SD = 1.79$, $t(65) = .43$, $p = .67$.

Criterion Validity

Convergent and Discriminant Validity

Table 8 shows the correlations with the validation measures for ages 3–15 years. The TPVT scores correlated well with the PPVT-IV, which taps the same construct, thus providing evidence of excellent convergent validity. The TPVT

TABLE 8

PEARSON CORRELATIONS BETWEEN TOOLBOX VOCABULARY COMPREHENSION SCORES AND VALIDATION MEASURES

	N	*r*	*p*-Value
PPVT-IV	198	.90	<.001
BVMT-R Total Recall	87	.46	<.001
RAVLT	85	.42	<.001
Average of BVMT-R Total Recall and RAVLT	87	.53	<.001

Note. WRAT-IV, Wide Range Achievement Test-4th Edition; PPVT-IV, Peabody Picture Vocabulary Test-4th Edition; BVMT-R, Brief Visuospatial Memory Test-Revised; RAVLT, Rey Auditory Verbal Learning Test.

score correlated weakly with measures that tap different constructs (the BVMT-R Total Recall, RAVLT, and the average of BVMT-R Total Recall and RAVLT), providing evidence of discriminant validity. The discriminant correlations were significantly lower than the convergent correlations ($p < .003$).

RESULTS (TORRT)

Four children did not successfully complete the TORRT for reasons such as lack of attention or alertness or general noncompliance.

Age Effects

Age was significantly correlated with the TORRT score ($n = 204$; $r = .86$; $p < .001$), as well as the WRAT-IV ($n = 203$; $r = .88$; $p < .001$). A quadratic model provided the best fit of the data, with $R^2 = .78$. Pairwise comparisons between age groups are reported in Appendix A. In the subset of participants age 3–6 the correlation between TORRT and age was .73 ($n = 117$, $p < .001$). In participants age 8–15 the correlation was .64 ($n = 87$, $p < .001$). Figure 8 shows TORRT scores as a function of age. It should be noted that TORRT scores and WRAT scores almost mirror each other at every age.

Test–Retest Reliability

The test–retest reliability for the TORRT was $ICC = .97$ ($n = 65$; 95% confidence interval: .95, .98). Reliability for the WRAT-IV was $ICC = .96$ ($n = 65$; 95% confidence interval: .94, .98).

Effect of Repeated Testing

Practice effects were computed as the difference between test and retest normalized scaled scores, with significance of the effect being tested with t

tests for dependent means. For the total child group (ages 3–15 years, $n = 66$), the TORRT showed no practice effect over an average 2-week test–retest interval: *mean practice effect* $= -.05$, $SD = .80$, $t(64) = -.51$, $p = .61$.

Criterion Validity

Convergent and Discriminant Validity

Table 9 shows the correlations with the validation measures for ages 3–15 years. The TORRT scores correlated well with WRAT-IV, which taps the same construct, providing evidence of excellent convergent validity, and weakly with the measure that taps a different construct (the BVMT-R Total Recall, which as previously noted was only administered to ages 8 and up), providing evidence of discriminant validity. The discriminant correlations were significantly lower than the convergent correlations ($p < .001$). The correlation between the TORRT scores and PPVT-IV scores was moderate, confirming the known relation between reading and vocabulary, but also providing evidence of the independence of the two constructs.

DISCUSSION

Development of the NIH Toolbox Picture Vocabulary Test and the NIH Toolbox Oral Reading Recognition Test represents an unprecedented effort to create high-quality language assessments using cutting edge psychometric theory and computer-based test administration. We have demonstrated that precise assessments of each of these constructs can be obtained in 5 min with a level of accuracy not seen in any other short assessment of this kind. Ceiling and floor effects, common to most measures covering a wide range of ability,

TABLE 9

PEARSON CORRELATIONS BETWEEN TOOLBOX ORAL READING RECOGNITION SCORES AND VALIDATION MEASURES

	N	*r*	*p*-Value
WRAT-IV	202	.96	<.0001
PPVT-IV	200	.87	<.0001
BVMT-R Total Recall	86	.41	<.0001
RAVLT	84	.45	<.0001
Average of BVMT-R Total Recall and RAVLT	86	.53	<.0001

Note. WRAT-IV, Wide Range Achievement Test-4th Edition; PPVT-IV, Peabody Picture Vocabulary Test-4th Edition; BVMT-R, Brief Visuospatial Memory Test-Revised; RAVLT, Rey Auditory Verbal Learning Test.

have been removed through the inclusion of a large corpus of items, spanning the complete continuum of difficulty, from preemerging language through PhD-level materials. An advantage of all computer adaptive measures is that the reliability can be estimated for each individual participant and not just as an "average" across the total sample (the typical measure of reliability cited for fixed-length instruments). This enables the researcher to individually assess the accuracy of the measure obtained.

Each measure has been reduced to as pure a form as possible. The TPVT has no reading component and is prompted by listening to a professionally recorded voice. The TORRT presents simple letter or word prompts on a clear field background with no distractions. TPVT has several advantages over the PPVT, including the increased sensitivity that results from having more words at every level, particularly at higher ability levels.

The photographic prompts for the vocabulary items are both contemporary and appealing. These professional images have been licensed for research use in perpetuity. Licensing for higher resolutions was also acquired, insuring continued use with evolving technology. As common monitor resolutions continue to improve (e.g., yesterday's VGA standard versus today's high definition), the NIH Toolbox items can be re-released in higher resolution formats.

The test–retest correlation as well as convergent and discriminant validity results obtained for both CB language measures were strong. The relation of each measure to participant age was as expected. Test–retest reliability for the PPVT-IV was noted to be stronger than that obtained with the TPVT, implying that the accuracy of the CB scores obtained were marginally weaker. This may be attributable to the fact that the CB measures were designed to be administered in 5 min, as compared to the PPVT-IV, which has administration times reported to fall in the 10–15 min range. Generally, a longer, well-developed test will always outperform a shorter one. Given the goal to create a "brief" measure of language proficiency for use in the NIH Toolbox, the newly created vocabulary comprehension measure performs admirably. For researchers who require increased reliability (as might be the case when examining individual ability at a clinical level), the CAT algorithm can be adjusted to administer a longer test. Clinical level accuracy can similarly be obtained for the TORRT through adjustment of the reading CAT algorithm. Additionally, TPVT responding during validation was through the use of a touch screen—a modality judged to be poor for the youngest children. During the NIH Toolbox norming phase, young children will be directed to point to the correct answer or use a mouse.

The vocabulary measure within CB is largely patterned after vocabulary measures that have been used in the past to infer more general linguistic attainment. These tests, such as the PPVT-IV or the Picture Vocabulary measure on the Woodcock–Johnson-III, are weighted toward nouns and

object words, and test vocabulary knowledge (vs. grammatical competence). As a practical matter, we were constrained to use vocabulary as the primary index of language. Time and delivery method precluded development of an assessment that measured multiple language skills. We also wrestled with these constraints, knowing that full language competence rests on more than mere noun learning, and requires mastery of grammatical constructs such as verb agreement (e.g., "The boy smiles at the man"), pluralization, and the use of passive sentence structure ("The car was driven by the woman"). Some research suggests that this fuller examination of language is a better predictor not only of future language, but also of future reading outcomes (NICHD, 2005).

We are also aware that language is characterized not only by the *products* of learning or the outcomes, but also by the *processes* of learning (Fisher, 1996; Hirsh-Pasek & Golinkoff, 1996; Hirsh-Pasek, Kochanoff, Newcombe, & De Villiers, 2005; Marcus, Vijayan, Bandi Rao, & Vishton, 1999; Saffran, Aslin, & Newport, 1996; Seymour, Roeper, & de Villiers, 2004). Processes, such as *fast mapping*, that help children connect a word and a referent with few exposures are hallmarks of language processing that are amenable to assessment. Indeed, there is also growing evidence in the literature that it is these early processes that predict later observable language milestones (Hurtado, Marchman, & Fernald, 2008; Tsao, Liu, & Kuhl, 2004). Process indicators might also be better predictors of success in learning than observable milestones because they tend to be less culturally and linguistically biased and less influenced by environmental variables.

Following norming, IRT item calibrations will again be recalculated, and any weaker items permanently removed from the item bank. Spanish versions of these instruments have also been developed and will be validated as part of the national norming study.

Data from the norming study will enable numerous examinations of the performance of these new instruments, as well as the assessment of hypotheses regarding the role of language acquisition relative to the other 45 constructs examined by the NIH Toolbox. We expect to find that reliability is poorest for emerging readers whose language acquisition appears to be the most inconsistent. In an attempt to further examine the relation between language attainment of children and their parents' education, we hope 1 day to assess the vocabulary of parent–child dyads. We would obtain language scores from children and their parents, as well as their respective levels of education. Vocabulary comprehension and reading decoding could be explored in relation not only to the other measures of cognition, but also to emotional health and sensory functioning as well. (Note: As of the date of publication the norming of the NIH Toolbox is complete. Norming results are available at www.nihtoolbox.org.)

REFERENCES

Benedict, R. (1997). *Brief Visuospatial Memory Test—Revised: Professional Manual*. Odessa, FL: Psychological Assessment Resources.

Brown, G. D. A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavior Research Methods, Instruments, & Computers*, **16**, 502–532.

Burton, C. L., Strauss, E., Hultsch, D. F., & Hunter, M. A. (2006). Cognitive functioning and everyday problem solving in older adults. *Clinical Neuropsychologist*, **20**(3), 432–452.

Caplan, D., & Hildebrandt, N. (1988). *Disorders of syntactic comprehension*. Cambridge, MA: MIT Press.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. Amsterdam: Elsevier.

Dale, E., & O'Rourke, J. (1976). *The living word vocabulary: The words we know: A national vocabulary inventory*. Elgin, IL: Field Enterprises Educational Corp.; distributed exclusively by Dome.

Dickinson, D. K., Golinkoff, R. M., & Hirsh-Pasek, K. (2010). Speaking out for language: Why language is central to reading development. *Educational Researcher*, **39**(4), 305–310.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody picture vocabulary test* (4th ed.). San Antonio, TX: Pearson.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, **59**(5), 1–173.

Fisher, C. (1996). Structural limits on verb mapping: The role of analogy in children's interpretations of sentences. *Cognitive Psychology*, **31**(1), 41–81.

Friederici, A. D., Rüschemeyer, S. A., Hahne, A., & Fiebach, C. J. (2003). The role of left inferior frontal and superior temporal cortex in sentence comprehension: Localizing syntactic and semantic processes. *Cerebral Cortex*, **13**(2), 170–177.

Gershon, R. C. (1988). *Index of words in the Johnson O'Connor Research Foundation, Inc. Vocabulary Item Bank*. New York: Johnson O'Connor Research Foundation Human Engineering Laboratory.

Gershon, R. C., Cella, D., Fox, N. A., Havlik, R. J., Hendrie, H. C., & Wagster, M. V. (2010). Assessment of neurological and behavioural function: The NIH Toolbox. *Lancet Neurology*, **9**(2), 138–139.

Gleason, J. B. (1997). *The development of language*. Boston: Allyn and Bacon.

Gleason, J. B., & Ratner, N. B. (2009). *The development of language* (7th ed.). Boston: Pearson/Allyn and Bacon.

Grober, E., & Sliwinski, M. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology*, **13**(6), 933–949.

Hirsch, C., & Wexler, K. (2007). The late acquisition of raising: What children seem to think about seem. In W. D. Davies & S. Dubinsky (Eds.), *New horizons in the analysis of control and raising* (pp. 35–70). New York: Springer.

Hirsh-Pasek, K., & Golinkoff, R. M. (1996). *The origins of grammar: Evidence from early language comprehension*. Cambridge, MA: MIT Press.

Hirsh-Pasek, K., Kochanoff, A. T., Newcombe, N., & De Villiers, J. (2005). Using scientific knowledge to inform preschool assessment: Making the case for empirical validity. *Society for Research in Child Development Social Policy Report*, **19**(1), 3–19.

Hurtado, N., Marchman, V. A., & Fernald, A. (2008). Does input influence uptake? Links between maternal talk, processing speed and vocabulary size in spanish-learning children. *Developmental Science*, **11**(6), 31.

Kastner, J. W., May, W., & Hildman, L. (2001). Relationship between language skills and academic achievement in first grade. *Perceptual and Motor Skills*, **92**(2), 381–390.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Kuhl, P., & Rivera-Gaxiola, M. (2008). Neural substrates of language acquisition. *Annual Review of Neuroscience*, **31**, 511–534.

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, **5**(11), 831–843.

Linacre, J. M. (2005). *A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs*. Chicago, IL: Winsteps.

Manly, J. J., Byrd, D. A., Touradji, P., & Stern, Y. (2004). Acculturation, reading level, and neuropsychological test performance among african american elders. *Applied Neuropsychology*, **11**(1), 37–46.

Manly, J. J., Jacobs, D. M., Sano, M., Bell, K., Merchant, C. A., Small, S. A., et al. (1999). Effect of literacy on neuropsychological test performance in nondemented, education-matched elders. *Journal of the International Neuropsychological Society*, **5**(3), 191–202.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, **283**(5398), 77–80.

Meisinger, E. B., Bloom, J. S., & Hynd, G. W. (2010). Reading fluency: Implications for the assessment of children with reading disabilities. *Annals of Dyslexia*, **60**(1), 1–17.

Mogilner, A. (1992). *Children's writer's word book*. Cincinnati, OH: Writer's Digest Books.

National Early Literacy Panel, & National Center for Family Literacy. (2008). Developing early literacy report of the National Early Literacy Panel, from http://purl.access.gpo.gov/GPO/LPS108121

National Reading Panel, & National Institute of Child Health and Human Development. (2000). *National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. Washington, DC: National Institute of Child Health and Human Development, National Institutes of Health.

National Research Council Committee on the Prevention of Reading Difficulties in Young Children, Snow, C. E., Burns, M. S., Griffin, P., & National Research Council Commission on Behavioral Social Sciences Education. (2002). *Preventing reading difficulties in young children: Intellectual property in the information age* (8th ed.). Washington, DC: National Academies Press.

NICHD Early Child Care Research Network. (2005). Pathways to reading: The role of oral language in the transition to reading. *Developmental Psychology*, **41**(2), 428–442.

Pinker, S. (2000). *The language instinct: How the mind creates language*. New York: Perennial Classics.

Price, C. J. (1998). The functional anatomy of word comprehension and production. *Trends in Cognitive Sciences*, **2**(8), 281–287.

Price, C. J. (2000). The anatomy of language: Contributions from functional neuroimaging. *Journal of Anatomy*, **197**(3), 335–359.

Rapcsak, S. Z., Henry, M. L., Teague, S. L., Carnahan, S. D., & Beeson, P. M. (2007). Do dual-route models accurately predict reading and spelling performance in individuals with acquired alexia and agraphia? *Neuropsychologia*, **45**(11), 2519–2524.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Dansmarks Paedagogiske Institut.

Rey, A. (1958). *L'Examen clinique en psychologie.* Paris: Press Universitaire de France.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, **274**(5294), 1926–1928.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, **86**(1), 162–173. doi: 10.1037/0022-3514.86.1.162

Seymour, H. N., Roeper, T., & de Villiers, J. G. (2004). Conclusions, future directions, and implications for remediation. *Seminars in Speech and Language*, **25**(1), 113–115.

Sternberg, R. J. (2004). Intelligence in humans. In S. Charles (Ed.), *Encyclopedia of Applied Psychology* (pp. 321–328). New York: Elsevier.

Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2004). Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child Development*, **75**(4), 1067–1084.

University of Western Australia School of Psychology. (2011). MRC Psycholinguistic Database Retrieved May 5, 2011, from http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm

Wilkinson, G. S., & Robertson, G. J. (2006). *WRAT 4: Wide range achievement test professional manual.* Lutz, FL: Psychological Assessment Resources.

Wolf, M. S., Gazmararian, J. A., & Baker, D. W. (2005). Health literacy and functional health status among older adults. *Archives of Internal Medicine*, **165**(17), 1946–1952.

Ziegler, J. C., Castel, C., Pech-Georgel, C., George, F., Alario, F. X., & Perry, C. (2008). Developmental dyslexia and the dual route model of reading: Simulating individual differences and subtypes. *Cognition*, **107**(1), 151–178.