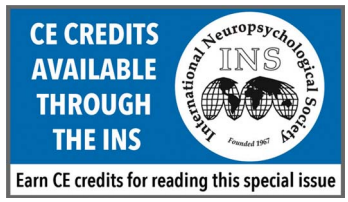
*Journal of the International Neuropsychological Society* (2014), **20**, 642–651

doi:10.1017/S1355617714000411

**SPECIAL SERIES – Guest Edited by Skye McDonald**

# Language Measures of the NIH Toolbox Cognition Battery




Richard C. Gershon,[1] Karon F. Cook,[1] Dan Mungas,[2] Jennifer J. Manly,[3] Jerry Slotkin,[1] Jennifer L. Beaumont,[1] AND Sandra Weintraub[4]

[1]Department of Medical Social Sciences, Northwestern University, Chicago, Illinois
[2]Department of Neurology, University of California, Davis, California
[3]Cognitive Neuroscience Division, Taub Institute for Research in Alzheimer's Disease and the Aging Brain, Columbia University, New York, New York
[4]Cognitive Neurology and Alzheimer's Disease Center, Northwestern Feinberg School of Medicine, Chicago, Illinois

(RECEIVED September 12, 2013; FINAL REVISION April 1, 2014; ACCEPTED April 3, 2014; FIRST PUBLISHED ONLINE June 24, 2014)

**Abstract**

Language facilitates communication and efficient encoding of thought and experience. Because of its essential role in early childhood development, in educational achievement and in subsequent life adaptation, language was included as one of the subdomains in the NIH Toolbox for the Assessment of Neurological and Behavioral Function Cognition Battery (NIHTB-CB). There are many different components of language functioning, including syntactic processing (i.e., morphology and grammar) and lexical semantics. For purposes of the NIHTB-CB, two tests of language—a picture vocabulary test and a reading recognition test—were selected by consensus based on literature reviews, iterative expert input, and a desire to assess in English and Spanish. NIHTB-CB's picture vocabulary and reading recognition tests are administered using computer adaptive testing and scored using item response theory. Data are presented from the validation of the English versions in a sample of adults ages 20–85 years (Spanish results will be presented in a future publication). Both tests demonstrated high test–retest reliability and good construct validity compared to corresponding gold-standard measures. Scores on the NIH Toolbox measures were consistent with age-related expectations, namely, growth in language during early development, with relative stabilization into late adulthood. (*JINS*, 2014, *20*, 642–651)

**Keywords:** Reading test, Vocabulary test, Neuropsychological assessment, Computer adaptive testing, Item response theory, Spanish version

## INTRODUCTION

The NIH Toolbox for the Assessment of Neurological and Behavioral Function is the result of a 6-year contract funded by the NIH Neuroscience Blueprint to create a battery of assessments of Cognition, Emotion, Motor, and Sensory function that could be used as a form of common currency across research studies. This study highlights the development of the Cognition instruments to assess Language. The NIH Toolbox Picture Vocabulary Test (TPVT) assesses vocabulary comprehension; the NIH Toolbox Oral Reading Recognition Test (TORRT) assesses reading decoding.

Correspondence and reprint requests to: Richard Gershon, Northwestern University, Suite 2700, 625 North Michigan Avenue, Chicago, IL 60611. E-mail: gershon@northwestern.edu

### Subdomain Definition

Language is a fundamental human capacity that facilitates communication and thought. In an increasingly literate world, language skills are powerful predictors of adaptive functioning and health (Burton, Strauss, Hultsch & Hunter, 2006). Language promotes the transmission of culture, values, and history. Language skills predict communicative competence and subsequent overall academic performance (Catts, Fey, Tomblin, & Zhang, 2002; Scarborough, 2001). The importance of language across the lifespan for cognitive and social development and for academic success and work achievement, as confirmed by rankings provided by surveys of hundreds of NIH-funded epidemiologists, provided impetus for its inclusion in the NIH Toolbox (Gershon et al., 2010; Nowinski, Victorson, Debb, & Gershon, 2013; Victorson et al., 2013; Weintraub et al., 2013).

There is general agreement on language milestones and the processes whereby they are acquired (Fenson et al., 1994; Golinkoff & Hirsh-Pasek, 1999; Hirsh-Pasek & Golinkoff, 1996). In childhood, all components of language competence, namely phonology, morphology, syntax and semantics, are undergoing development. In adulthood, some language skills remain substantially intact with advancing age. Vocabulary continues to develop as individuals acquire more words through experience (Salthouse, 1988). Because of the need to have the NIHTB-CB measures apply to a broad age range, word knowledge and reading were selected as particularly relevant for measurement across the lifespan.

In healthy normal individuals, tests of reading and vocabulary knowledge, can serve as useful "proxy" measures for deriving an estimate (Baumann, 2009) of overall intellectual ability since there is a high correlation between IQ scores and vocabulary (Smith, Smith, Taylor, & Hobby, 2005). Vocabulary knowledge also has been found to be an effective marker of the level of acculturation in minority groups (Deyo, Diehl, Hazuda, & Stern, 1985). Reading level, along with education, is a robust proxy measure for "cognitive reserve," defined as the brain's ability to efficiently access brain networks and alternative strategies in the face of cognitive challenge (Jefferson et al., 2011). Scores on language measures predict occupational interest and performance (Broadley, 1994), academic success, and socioeconomic status in adulthood (Ritchie & Bates, 2013). In addition, language proficiency is important for maintaining health. For example, health literacy, the capacity to use basic health information to make informed health decisions, is very strongly related to reading vocabulary and to a variety of other cognitive skills (Wolf et al., 2012).

Brain injuries caused by trauma, neurodegenerative diseases, stroke, and tumors can affect neuroanatomical language networks, causing aphasia, and many of these diseases are more common among aging adults. For example, the National Aphasia Association (https://www.aphasia.org) reports that more than 80,000 individuals per year acquire aphasia as a consequence of stroke. Thus, the NIHTB language measures also reflect the integrity of the left Perisylvian language areas. The NIHTB Picture Vocabulary Test measures auditory comprehension of single words that are graded in difficulty and measured with an auditory word-picture matching paradigm. Auditory word comprehension can be disturbed in aphasia due to stroke or neurodegenerative disease. Oral reading ability can be sensitive to language disorders that interfere with word comprehension and production.

The TPVT and TORRT have several advantages over existing instruments. They were developed in both English and Spanish and target a broad age range (3 to 85 years). In addition, each instrument is based on an "item bank," a collection of items calibrated to an item response theory (IRT) model (Revicki & Cella, 1997). This approach allows for individually tailored assessment. Both instruments are administered using computer adaptive testing (CAT), a dynamic approach to testing in which the difficulty of the items administered is tailored to the ability of the participant. CAT-based assessment increases precision without increasing response burden.

A summary of the development of the TPVT and TORRT has been published as it relates to pediatric populations (Gershon et al., 2013). In the present study, we focus on the relevance of testing language in adults, the development of the TPVT and TORRT, and the reliability and validity of TPVT and TORRT scores in adult samples.

## METHOD

### General Methods for the NIHTB Cognition Battery

Once the target constructs within language were identified, several beta versions of the language tests preceded the version used in validation. The validation version represents the fifth and final version; the previous four served to improve stimulus presentation, ease of administration, and item acceptability based on cultural parameters. In addition to developing the NIHTB-CB language measures, the authors also selected "gold standard" validation instruments to evaluate convergent and discriminant validity (see below). Individuals in the validation study were administered the NIHTB-CB measures and the gold standard measures.

### Participants

Two samples of participants were recruited for this study. The calibration sample served to associate the difficulty of each item with probabilities of answering the item correctly at different levels of ability. This was accomplished using an IRT calibration. The validation sample served to evaluate reliability and convergent and discriminant validity of the final versions of the TPVT and TORRT. Although children and adult participants were included in the development of these language measures, only the adult portion of the validation study is reported in this study. There were 268 participants, 159 from 20–60 years of age and 109 from 65–85 years. There were more individuals with high school education or lower than those with a college education. Fifty-five percent of the sample was white, with 28% black and 17% Hispanic. The study was approved by the Institutional Review Board at Northwestern University and informed consent was obtained from all participants. A more complete description of this sample is included in the article by Weintraub et al. in this special edition (Weintraub et al., manuscript submitted).

### Measures

#### *NIH Toolbox Picture Vocabulary Test (TPVT)*

After much consideration of many different types of language measures, picture vocabulary (receptive vocabulary) was chosen as the primary language measure for the NIH Toolbox. Although grammatical proficiency is critical

https://doi.org/10.1017/S1355617714000411 Published online by Cambridge University Press

for development (Gleason & Ratner, 2009; Hirsh-Pasek & Golinkoff, 1996), we ultimately selected vocabulary knowledge because of its high association with success in school and work (Kastner, May, & Hildman, 2001; Schmidt & Hunter, 2004) and its high correlation with general measures of "intelligence", i.e., the "*g*" factor hypothesized by Cattell (Cattell, 1987). Furthermore, vocabulary is a more stable construct across languages and less subject to the complexities of equating syntax or morphology across languages.

### *Development of the TPVT*

*Step 1*. Identify an initial pool of candidate words we used three sources to develop a pool of candidate words for the TPVT: The Johnson-O'Connor Research Foundation database of previously field-tested vocabulary words (Gershon, 1988), The Living Word Vocabulary (Dale & O'Rourke, 1976) and the Children's Writer's Word Book (Mogilner, 1992). Words were selected from these sources based on their perceived difficulty, their appropriateness for different age levels, and their frequency. This selection process yielded a total of 739 words in the candidate item pool.

*Step 2*. Review and modify candidate item pool the 739 candidate words were reviewed against two standards: (1) translatability into visual images ("imageability") and (2) appropriateness for and relevance to the measurement of vocabulary. For each candidate word, we obtained imageability and frequency rankings from the University of Western Australia MRC Psycholinguistic Database (University of Western Australia School of Psychology, 2011). A formal imageability criterion was not used because the purpose was to identify words (and potential distractors) that could be suitably captured with a photograph, not to discriminate among levels of imageability; however, words with ratings below 350 (approximately 1 *SD* below the reported mean) were flagged for further review by content experts to determine if they should be retained for the photograph-searching phase. Frequency was examined to ensure that words with both low and high frequency were included. Retained words were reviewed by pediatric and geriatric professionals with expertise in language. Based on their feedback, words were dropped from the candidate list.

*Step 3*. Development of images for correct choices and distractors for every candidate word (correct choice), three plausible alternatives (distractors) were developed. The distractor choices were developed in a way that considered the specific word and the general developmental level at which it was targeted (e.g., simpler distractors would be used for words targeted at preschool children than for those targeted for high performers). Some general patterns for developing distractors included the use of words that were phonologically similar, semantically similar (without also being a potential correct answer) or antithetical (antonym), visually similar, and/or represented common misconceptions of a word's meaning. A narrative description of how each candidate word and distractor would be imaged was reviewed by both a senior content expert and a team of language experts who evaluated them for content coverage and sensitivity. After this review, 4–10 suggested images for each word were provided by Getty images staff selected from the Getty Images library.

In the next round of reviews, a panel comprised of senior content experts, educator-consultants, and experts in multicultural issues reviewed the words and corresponding images. This review included examination for appropriateness, cultural

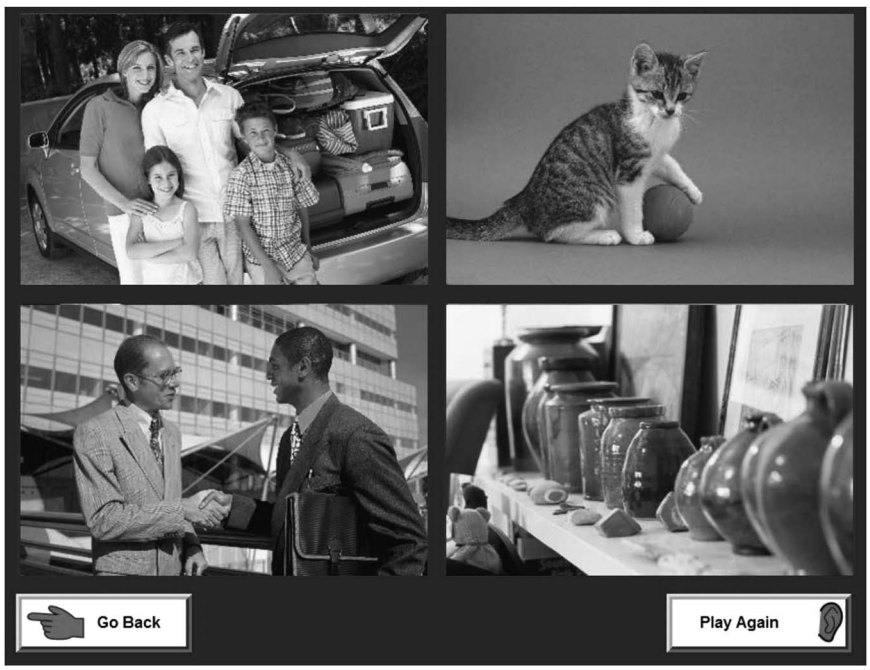



**Fig. 1.** Toolbox Picture Vocabulary Test Sample Item ("Kin"). © 2010–2014 National Institutes of Health and Northwestern University. (Correct answer is in upper left-hand corner.)

https://doi.org/10.1017/S1355617714000411 Published online by Cambridge University Press

fairness and sensitivity. Based on the panel's feedback, items were edited or dropped. Once satisfactory images were identified, they were edited professionally to create consistency in background and orientation and to remove extraneous information. The resulting candidate item bank contained 625 items. A sample TPVT item is shown in Figure 1.

### *Development of an IRT-Calibrated Item Bank and CAT for TPVT*

*Step 1.* Calibrate items to an IRT Model calibration to an IRT model requires that responses to test items be collected from a sample of individuals. It would have been impractical to administer all 625 items to all respondents; therefore, 21 forms were generated that each included 40–60 items, insuring that at least 200 unique responses were obtained for each item. Only one form was administered to each participant. Half of the items of each form overlapped with items of another form to allow for joint calibration of items. The initial calibration sample included 4703 paid participants accrued by an online panel company ($N = 1513$; adults ages, 18–69 years; Mean = 25.76 years; Female = 59.9%), with education for adults spread relatively evenly from completion of 10th grade through graduate/doctorate level. Items were scored and calibrated, using the Winsteps software program (Linacre, 2005), to the one-parameter/Rasch Item Response Theory (IRT) model (Rasch, 1980). Twenty-three items were removed with pt-biserials $< .2$, INFIT $> 1.2$ or OUTFIT $> 3$. The resulting 602 word item bank was used as the basis for CAT administration of the items during validation.

*Step 2.* Develop TPVT CAT every CAT has a "stopping rule." Most stopping rules are based on reaching a pre-specified level of precision (variable CAT), administering a given number of items (fixed-length CAT), or some combination of these two rules. For the validation study, we elected to use a fixed-length 25-item CAT primarily to increase the number of items for which data were collected, increasing the potential for future refinement of item calibrations.

### *Description of the TPVT*

When the TPVT is administered, participants are positioned in front of a computer screen. Instructions and images are presented by computer. A recorded voice says,

> You are going to be asked the meaning of some words. For each item, you will hear a word and see four pictures on the screen. Click on the picture that you think best matches the meaning of the word that was said. If you are not sure, make your best guess. If you need to hear the word again, click on the button that has a picture of an *EAR*, also called the *PLAY AGAIN* button. After you make your choice and click on a picture, the computer will automatically go to the next word and pictures. You will keep hearing words and clicking on pictures until you are done. (National Institutes of Health & Northwestern University, 2006–2012a).

Testing begins with practice items. Participants receive feedback on accuracy only for these trials. Participants are allowed to change their answers to the previous item if they wish. During the validation study, participants used a touchscreen with instructions adjusted accordingly, but the final release version of the Toolbox is administered using a standard screen and mouse.

After the participant responds to the first item, the computer "selects" the second item based on whether or not the first was answered correctly. Successive items are based on a continually updated estimate of the respondent's ability. Test difficulty is controlled such that the participant has a 50% likelihood of answering each item correctly. Testing continues until the standard error (SE) of performance is less than 0.3. A maximum of 25 items are administered, which typically takes five minutes or less. The .3 SE cutoff was originally selected as it the accuracy level obtained by more than 95% of subjects in less than 5 min—a time limit imposed by the NIH Toolbox design team. Scores are calculated by computer and require no additional work by the examiner.

## NIH Toolbox Oral Reading Recognition Test

The second language measure is the NIH Toolbox Oral Reading Recognition Test (TORRT). It measures ability to pronounce single printed letters or words, including words that occur infrequently and have irregular orthography. This ability has been used as a proxy for educational, cognitive, and socioeconomic factors and as an estimate for general intelligence (Grober & Sliwinski, 1991).

### *Development of the TORRT*

To obtain a set of candidate words for the TORRT, we searched the University of Western Australia MRC Psycholinguistic Database (University of Western Australia School of Psychology, 2011) considering (1) frequency of occurrence in the target language, (2) complexity of letter-sound relations, (3) degree of orthographical typicality, (4) rating of age of acquisition, (5) number of syllables, and (6) number of phonemes. Single letters of the alphabet were included to test those with low literacy levels.

Because the candidate item pool of words and letters was large, it was reduced by applying criteria that maximized the range of, and minimized redundancy in, the difficulty of items. Selected words varied in the number of letters (2–14), frequency, and whether they had irregular orthography to phonology matches. Words with multiple acceptable pronunciations were removed.

After this initial reduction in the candidate item pool, a panel including geriatric and pediatric language experts reviewed the items and provided feedback on what items to retain and what content was not covered by current items. This process netted a pool of 298 candidate items, which included 277 words and 21 letters and "pre-reading" items with a broad range of difficulty. "Pre-reading" items are administered as multiple-choice items in which the respondent is asked to identify the correct letter

https://doi.org/10.1017/S1355617714000411 Published online by Cambridge University Press

from among other letters and/or non-letter symbols. After pilot testing, a one-item-per-screen format was chosen for presenting the TORRT items. This format proved less visually cluttered, was easier for the examiner to score, and took no more time for respondents to complete than a format in which 5–6 items were presented per screen.

### *Development of an IRT-Calibrated Item Bank for TORRT*

The TORRT items require that the test administrator score the oral response of the participant. A total of 146 participants were recruited from five test sites and four geographic locations (West Orange, NJ; Minneapolis, MN; Atlanta, GA; Evanston, IL; Chicago, IL). The data from these participants were used for initial item calibrations. For the validation study, four pilot TORRT forms with graduated difficulty, ranging from 70 to 120 items each (Forms 1–4) were developed. These long forms facilitated data collection to improve item calibrations. Subjects were routed to one of these forms following the administration of a nine-item pretest.

### *Description of the TORRT (administered as a CAT)*

Participants and examiners are seated in front of different computer screens. The examiner first keys in the subject's educational level to set the starting point for the test. The examiner then tells the participant:

> Now, I'm going to show you some letters and some words. I want you to read each letter or word out loud. Read each one loud enough so that I can hear you. Some will be easy and some will be hard. Don't worry if you don't know the word or its meaning—just read it out loud the best you can. Let's begin. (National Institutes of Health & Northwestern University, 2006-2012b)

The examiner views the accepted pronunciation on his/her screen and codes the response as correct or incorrect. For those with low literacy levels, letters and other multiple-choice "pre-reading" items are presented. Toolbox examiners are trained on correct word pronunciation before administering the TORRT, using audio recordings and a written list of the correct pronunciations, as well as a brief proficiency test. Figure 2 is a screen shot for a sample item as viewed by the participant. Figure 3 is the corresponding examiner screen that provides the correct pronunciation of the word. Testing continues until a .3 standard error level of accuracy is obtained or 25 items are administered, with a median of 20 items administered in 4 minutes.

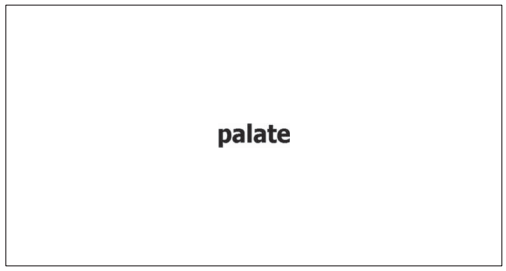



**Fig. 2.** NIH Toolbox Oral Reading Recognition Examinee Screen (National Institutes of Health & Northwestern University, 2006–2012b). © 2010–2014National Institutes of Health and Northwestern University.

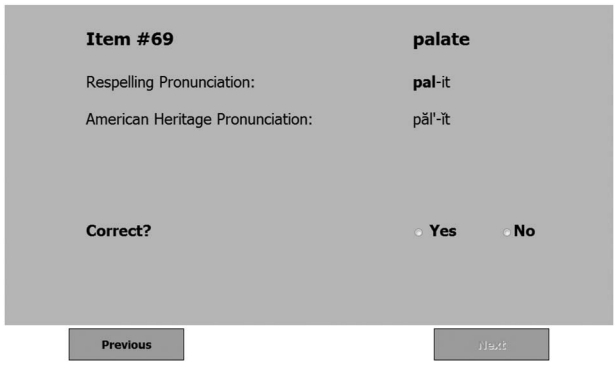



**Fig. 3.** NIH Toolbox Oral Reading Recognition Test Examiner Screen (National Institutes of Health & Northwestern University, 2006–2012b). © 2010–2014National Institutes of Health and Northwestern University.

## Reliability and Validation Assessment

### *Validation measures*

Four "gold standard" measures were selected to assess the validity of the TPVT and the TORRT, two for convergent validity and two for discriminant validity.

*Peabody Picture Vocabulary Test-4th edition (PPVT-4).* (Dunn & Dunn, 2007) The PPVT-4 is a norm-referenced test of receptive oral vocabulary. There are two parallel forms, each with 228 items. The PPVT-4 is not a timed test, but usually takes 10–15 min to complete. Testing materials include a spiral-bound set of pages, each of which has four colored images. The examiner says a word and presents a page with the four images, asking the participant to identify the picture for that word. The PPVT-4 averages 65% percent correct, but individual subjects experience 7.05–99.29% correct (Pae, Greenberg, & Williams, 2012). In the current study, PPVT-4 scores were compared to TPVT scores to assess convergent validity.

*Wide Range Achievement Test 4 – reading subtest (WRAT-4).* (Wilkinson & Robertson, 2006) The WRAT-4 reading subtest 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. WRAT-4 subjects average a 38% correct experience, with individual experience ranging from 0 to 77% correct (Chiappe, Siegel, & Gottardo, 2002). The WRAT-4 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

https://doi.org/10.1017/S1355617714000411 Published online by Cambridge University Press

measure visuospatial memory. Test takers view six geometric figures and are asked in each of three successive learning trials to draw from memory as many of the figures in their correct spatial position as possible, after the figures are removed from view. The BVMT-R correlates most strongly with other tests of visual memory and less strongly with tests of verbal memory (Benedict, 1997; Benedict, Schretlen, Groninger, Dobraski, & Shpritz, 1996). For college students, the BMVT-R scores average 89.58%, but individual scores vary from 67.42% to 100% correct (Kontos, 2007). The BVMT-R served as an assessment of discriminant validity for both Toolbox Language tests.

*Rey Auditory Verbal Learning Test (RAVLT).* (Rey, 1958) On the RAVLT, a list of 15 words is read aloud by the examiner and the test-taker's task is to repeat as many words as possible, in any order. Three such learning trials with immediate recall were presented. RAVLT scores average 52.33% correct, but with a range of 45.13–66.6% (Boone, Lu, & Wen, 2005). The RAVLT was included as a second measure of discriminant validity for the TPVT.

### *Reliability assessment*

A full description of the test–retest sample has been published (Weintraub et al., 2014). Briefly, the TPVT and TORRT were re-administered to 89 participants, 7 to 21 days (Mean = 15.5; *SD* = 4.8) following the initial assessment.

## Analyses

Key psychometric properties of both IRT- and classical statistics-based measure are test reliability and validity.

### *Score conversion*

Scores on all measures were scaled by first ranking the raw scores by magnitude and then applying a normative transformation to the ranks to create a standard normal distribution. This distribution was then linearly transformed to have a mean of 10 (*SD* = 3). This placed the scores on a common metric, facilitating analyses of reliability and validity. Scores were not adjusted for age.

### *Test–retest reliability*

Intra-class correlation coefficients (ICC) with 95% confidence intervals were calculated to evaluate test–retest reliability.

### *Validity analysis*

The validity of scores on Toolbox measures was evaluated by correlating them with scores on well-established "gold standard" measures. Evidence of convergent validity was defined as obtaining relatively higher correlations with measures of the same construct; evidence of discriminant validity was defined as obtaining relatively lower correlations with measures of *different* cognitive constructs. Additional validity evidence was evaluated based on group comparisons using general linear models that examined associations between scores and subgroup membership. These analyses were adjusted for age, gender, and education, where appropriate. Effect sizes are reported as Cohen's d, with cutoffs of .20, .50, and .80 indicating small, medium, and large effects, respectively. These analyses are the same whether an instrument is based upon IRT or classical test theory.

## RESULTS

Of the 268 adult participants in the study, 265 completed the TORRT ($n$ = 158; age, 20–60 years; and $n$ = 107; age, 65–85 years) and 263 completed the TPVT ($n$ = 157; age, 20–60 years; and $n$ = 106; age, 65–85 years).

## Test–Retest Reliability

A total of 89 of the Time 1 participants volunteered to take the TPVT and TORRT twice, on average 15.5 days apart (*SD* = 4.8; range = 7 to 26 days). The test–retest intraclass correlation (ICC) between Time 1 and Time 2 for the TPVT was 0.80 (95% CI = 0.71–0.86). This coefficient was notably lower than for the PPVT-4 with the same sample. The test–retest ICC for the PPVT-4 was 0.92 (95% CI = 0.88 to 0.95). TPVT scores increased a mean of 0.24 points (*SD* = 1.88; ES = 0.13; $p$ = .232), indicating a negligible practice effect.

The retest ICCs were higher for the TORRT. The test–retest ICC for the TORRT was 0.90 (95% CI = 0.85–0.93). This compared favorably with the WRAT-4, which had a test–retest ICC of 0.84 (95% CI = 0.77 to 0.89). TORRT scores increased a mean of 0.10 points (*SD* = 1.41; ES = 0.07; $p$ = .505), indicating no practice effect.

## Validity Assessment

### *Convergent and discriminant validity*

The convergent validity of the TPVT was evaluated by correlating its scores with scores from the PPVT-4. To evaluate the TPVT's discriminant validity, its scores were correlated with scores of the BVMT-R Total Recall, the RAVLT, and a combined BVMT-R Total Recall and RAVLT scores (BVMT/RAVLT). The results supported the validity of the TPVT, with correlations of 0.80 ($p < .001$) with PPVT-4 scores, 0.11 ($p$ = .068) with BVMT-R Total Recall scores, 0.09 ($p$ = .087) with RAVLT scores, and 0.10 ($p$ = .105) with BVMT/RAVLT.

A similar approach was used to evaluate the validity of the TORRT scores. TORRT scores correlated 0.86 ($p < .001$) with WRAT-4 scores, 0.23 ($p < .001$) with BVMT-R Total Recall scores, 0.25 ($p < .001$) with RAVLT scores, and 0.26 ($p < .001$) with BVMT/RAVLT scores. Even though all TORRT correlations are significant, discriminant validity is fully demonstrated due to the much stronger relationship with the WRAT-4 (convergent validity measure).

https://doi.org/10.1017/S1355617714000411 Published online by Cambridge University Press

**Table 1.** Effect sizes (ES) for comparisons of scores between groups

| | TORRT | WRAT-4 | TPVT | PPVT-4 |
|---|---|---|---|---|
| **ES (Male vs Female)**[1] | **−0.23** | **−0.13** | **0.04** | **0.31** |
| p[1] | 0.037 | 0.25 | 0.74 | 0.004 |
| **ES (Black vs White)**[2] | **−0.60** | **−0.58** | **−0.68** | **−0.69** |
| ES (Hispanic vs White)[2] | **−0.52** | **−0.54** | **−0.59** | **−0.52** |
| p[2] | < 0.001 | < 0.001 | < 0.001 | < 0.001 |
| **ES (College vs < HS)**[3] | **1.06** | **0.96** | **0.98** | **1.05** |
| **ES (HS Grad vs < HS)**[3] | **0.40** | **0.51** | **0.44** | **0.27** |
| **ES (College vs HS grad)**[3] | **0.67** | **0.45** | **0.53** | **0.78** |
| p[3] | < 0.001 | < 0.001 | < 0.001 | < 0.001 |
| **ES (Excellent health vs very good)**[3] | **0.13** | **0.16** | **0.05** | **0.05** |
| **ES (Excellent health vs good)**[3] | **0.59** | **0.55** | **0.37** | **0.36** |
| **ES (Excellent health vs fair-poor)**[3] | **0.55** | **0.46** | **0.59** | **0.60** |
| p[3] | 0.001 | 0.006 | 0.012 | 0.013 |
| **ES (Employed vs Out of Work)**[3] | **0.41** | **0.36** | **0.23** | **0.30** |
| **ES (Employed vs Other)**[3] | **0.44** | **0.42** | **0.29** | **0.45** |
| **ES (Employed vs Retired)**[3] | **−0.05** | **−0.10** | **−0.01** | **−0.09** |
| p[3] | 0.028 | 0.039 | 0.307 | 0.038 |

[1]Adjusted for age and education
[2]Adjusted for gender, age and education
[3]Adjusted for age

### *Comparison of scores to demographic characteristics*

Validity was further evaluated by estimating associations between scores on the language measures and demographic variables (age, education, and gender). Age correlations were calculated both before and after partialing out years of education. Covariates were used to calculate demographic-adjusted means.

Age was not highly associated with vocabulary or reading in this adult sample as measured by the TPVT and TORRT. Age correlated with TPVT scores at 0.26 both before and after partialing out years of education. Age correlated with TORRT scores at 0.11 before partialing out years of education and 0.03 afterward.

Effect size (ES) comparisons were made for TPVT, TORRT, and gold standard scores (Table 1). The *p*-values in the table are for the omnibus null hypothesis of no differences in adjusted means among subgroups. Very large ESs were found in the comparison of TPVT and TORRT scores between those who had college degrees compared to those who had not graduated high school (1.06 and 0.98, respectively). These were similar to the results for the gold-standard measures of the same constructs. The PPVT-4 and WRAT-4 had effect sizes in this comparison of 1.05 and 0.96, respectively.

ESs for comparison of TPVT and TORRT scores by race/ethnicity were moderate to high. Black and Hispanic subgroups did less well on both the TPVT and TORRT, a pattern also present in the WRAT-4 and PPVT-4. General health also was an effective predictor of performance on the Toolbox language measures, and, as with race/ethnicity, the pattern was similar for the gold standard measures of the same constructs.

The smallest ESs in our comparisons were obtained for gender. Males did somewhat poorer than females on the TORRT (ES = −0.23), but did equally well on the TPVT (ES = 0.04). ESs for the WRAT-4 and PPVT-4 were −0.13 and 0.31.

## DISCUSSION

The NIH Toolbox language measures are high-quality tools developed using innovative psychometric methods. They are suitable for a broad range of populations, and tailored to the abilities of individual examinees. Though each NIH Toolbox language instrument takes 5 minutes or less to administer, the reliability and validity of the scores are similar to that of longer, "gold-standard" measures of the same construct. By including a large corpus of items that spanned the language continuum from pre-emerging language through very high language proficiency, we were able to avoid the ceiling and floor effects that often accompany measures used across a wide range of ability. Both the TPVT and TORRT take into account the proficiency of the examinee in the selection of items that are administered using a computer adaptive administration of items.

A related advantage of the NIH Toolbox language measures over standard measures is the use of an IRT model in calibrating items and computing scores. Not only does IRT allow the adaptive administration of items, but also the score metric approximates an equal interval scale, a distinct

https://doi.org/10.1017/S1355617714000411 Published online by Cambridge University Press

advantage when calculating statistics that make this assumption about the scores. Additionally, reliability of the assessments can be estimated for each individual participant. In traditional assessments, the reliability of a measure is "averaged" across the entire sample, obscuring the fact that instruments typically assess different levels of ability with varying levels of precision.

Another advantage of the NIH Toolbox language measures is the degree to which the administration has been standardized and the stimuli pruned of extraneous content. The TPVT has no reading component and is prompted by listening to a professionally recorded voice. The photographic prompts for the vocabulary items are not only contemporary and appealing, but also have been licensed for research use in perpetuity, including upgrades to higher resolutions as standard monitor resolutions continue to improve.

The results of the current study support the reliability and validity of the language measures in an adult sample. Test–retest correlations were particularly strong for the TORRT and marginally higher than those for the WRAT-4. Test–retest reliability for TPVT was also good, though retest values for the PPVT-4 were stronger.

Both the TPVT and TORRT exhibited appropriate convergent and discriminant validity. In addition, expected associations with demographics such as education, race, ethnicity and general health were observed. The validity of the measures was comparable to that of gold standard criterion measures.

Future research should explore the extent to which differences in reliability are based upon potential ceiling effects in the PPVT compared to the TPVT, which has more items available at the higher end of the scale (an instrument with a ceiling may still exhibit strong test–retest reliability). Additional work could also examine how reliability and validity is related to the tailored nature of the administration of Toolbox measures (using computer adaptive testing), versus the fixed from administration used by the validation measures.

The validity of the Toolbox language measures should continue to be evaluated, including testing hypotheses regarding language acquisition and the other domains examined by the NIHTB. Vocabulary and reading as measured by the TPVT and TORRT should be explored in relation to emotional health and sensory functioning. To date, there have been no studies of the language measures in clinical populations and their utility in patients with language disorders should be examined.

Spanish versions of the NIH Toolbox language measures were developed, calibrated, and normed separately from the English versions due to significant differences in performance (e.g., Spanish reading fluency is significantly easier overall than is English reading fluency). These data should be used in the future to evaluate the ability of the tests to assess bilingual language proficiency.

Work on the development of the TPVT and TORRT has continued. In late 2011, the latest English versions of the TPVT and TORRT were administered in a large national norming study. These norming data were used to refine IRT item calibrations of all banks and, as needed, to prune weaker items from the item bank.

The NIH Toolbox Picture Vocabulary Test and NIH Toolbox Oral Reading Recognition Test were released by the National Institutes of Health for royalty-free use by health and education researchers in late 2012 (see www.nihtoolbox.org for current information).

## ACKNOWLEDGMENTS

We thank Abigail Sivan and Edmond Bedjeti (Northwestern University) for their valuable assistance in the validation phase of testing. We also wish to acknowledge the following individuals for their helpful consultation during the development of the language measures of the NIH Toolbox Cognition Battery: Jean Berko Gleason (Boston University), Roberta Golinkoff (University of Delaware), Kathy Hirsh-Pasek (Temple University), and Marilyn Jager Adams (Brown University). *Disclosures:* This study is funded in whole or in part with Federal funds from the Blueprint for Neuroscience Research, National Institutes of Health, under Contract No. HHS-N-260-2006-00007-C. *Dr. Gershon* has received personal compensation for activities as a speaker and consultant with Sylvan Learning and the American Board of Podiatric Surgery. He is currently funded by several grants awarded by the NIH: N01-AG-6-0007, HHSN260200600007, 1U01DK082342-01, HD05469, 1RC2AG036498-01; NIDRR: H133B090024. *Dr. Cook* has received personal compensation as a consultant for Focus on Therapeutic Outcomes and Evidera. She is funded on grants from NIH (5U54AR05943-04, 5U54AR057951-04, 3U01AR052177-06S1, National MS Society (321-SUB//HC0145), and Veteran's Affairs Research and Development (679-13-1-1310-0006). *Dr. Mungas* is funded by research grants from the National Institute on Aging and a grant from the California Department of Public Health California Alzheimer's Disease Centers program. *Dr. Manly* is funded by NIH grants R01AG028786, R01AG037212; she had received funding previously from NIH grant R01AG016206 and a grant from the Alzheimer's Association (IIRG 05-14236). She is a consulting editor for the Journal of the International Neuropsychological Society. She serves on the Medical and Scientific Advisory Board of the Alzheimer's Association, and as a member of the Advisory Council on Alzheimer's Research, Care, and Services. *Dr. Slotkin* is funded by NIH grants 1U54AR057943-01 and HHSN275201200007I. *Ms. Beaumont* served as a consultant for NorthShore University HealthSystem, FACIT.org, and Georgia Gastroenterology Group PC. She received funding for travel as an invited speaker at the North American Neuroendocrine Tumor Symposium. *Dr. Weintraub* is funded by NIH grants # R01DC008552, P30AG013854, and the Ken and Ruth Davee Foundation and conducts clinical neuropsychological evaluations (35% effort) for which her academic-based practice clinic bills. She serves on the editorial board of *Dementia & Neuropsychologia* and advisory boards of the *Turkish Journal of Neurology* and *Alzheimer's and Dementia*.

## REFERENCES

Baumann, J.F. (2009). Intensity in vocabulary instruction and effects on reading comprehension. *Topics in Language Disorders*, *29*(4), 312–328. 310.1097/TLD.1090b1013e3181c1029e1022.

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

Benedict, R.H.B., Schretlen, D., Groninger, L., Dobraski, M., & Shpritz, B. (1996). Revision of the Brief Visuospatial Memory Test: Studies of normal performance, reliability, and validity. *Psychological Assessment*, *8*(2), 145–153.

Boone, K.B., Lu, P., & Wen, J. (2005). Comparison of various RAVLT scores in the detection of noncredible memory performance. *Archives of Clinical Neuropsychology*, *20*(3), 301–319.

Broadley, M.E. (1994). *Your natural gifts: How to recognize and develop them for success and self-fulfillment*. McLean, VA: EPM Publications.

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.

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

Catts, H.W., Fey, M.E., Tomblin, J.B., & Zhang, X. (2002). A longitudinal investigation of reading outcomes in children with language impairments. *Journal of Speech, Language, and Hearing Research*, *45*(6), 1142–1157.

Chiappe, P., Siegel, L.S., & Gottardo, A. (2002). Reading-related skills of kindergartners from diverse linguistic backgrounds. *Applied Psycholinguistics*, *23*(01), 95–116.

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

Deyo, R.A., Diehl, A.K., Hazuda, H., & Stern, M.P. (1985). A simple language-based acculturation scale for Mexican Americans: Validation and application to health care research. *American Journal of Public Health*, *75*(1), 51–55.

Dunn, D.M., & Dunn, L.M. (2007). *PPVT-4: Peabody picture vocabulary test* (4th ed.). Minneapolis: 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.

Gershon, R.C. (1988). *Index of words in the Johnson O'Connor Research Foundation, Inc. Vocabulary Item Bank* (Technical report 1988-3). New York, NY: 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.

Gershon, R.C., Slotkin, J., Manly, J.J., Blitz, D.L., Beaumont, J.L., Schnipke, D., … Weintraub, S. (2013). NIH Toolbox Cognition Battery (CB): Measuring language (vocabulary comprehension and reading decoding). Chapter IV, *Monographs of the Society for Research in Child Development*, *78*(4), 49–69.

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

Golinkoff, R.M., & Hirsh-Pasek, K. (1999). *How babies talk: The magic and mystery of language in the first three years of life*. New York: Dutton.

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.

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

Jefferson, A.L., Gibbons, L.E., Rentz, D.M., Carvalho, J.O., Manly, J., Bennett, D.A., … Jones, R.N. (2011). A life course model of cognitive activities, socioeconomic status, education, reading ability, and cognition. *Journal of the American Geriatrics Society*, *59*(8), 1403–1411.

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.

Kontos, D.L. (2007). *Investigation of validity, reliability, and practice effects of the Immediate Postconcussion Assessment and Cognitive Test (ImPACT) and Traditional Paper-Pencil Neuropsychological Tests* 2014. Retrieved from https://cdr.lib.unc.edu/indexablecontent/uuid:3bbcd1dd-b3d0-4aa2-99cb-1ce42792445f

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

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

National Institutes of Health, & Northwestern University. (2006-2012a). *NIH Toolbox Picture Vocabulary Test. Used with permission*. Retrieved from http://www.nihtoolbox.org/WhatAndWhy/Cognition/Language/Pages/NIH-Toolbox-Picture-Vocabulary-Test.aspx

National Institutes of Health, & Northwestern University (2006-2012b). *NIH Toolbox Reading Recognition Test. Used with permission*. Retrieved from http://www.nihtoolbox.org/WhatAndWhy/Cognition/Language/Pages/NIH-Toolbox-Oral-Reading-Recognition-Test.aspx

Nowinski, C.J., Victorson, D., Debb, S.M., & Gershon, R. (2013). Input on NIH Toolbox criteria: Surveying the end user research community. *Neurology*, *80*(11 Suppl. 3), S7–S12.

Pae, H.K., Greenberg, D., & Williams, R.S. (2012). An analysis of differential response patterns on the Peabody Picture Vocabulary Test-IIIB in struggling adult readers and third-grade children. *Reading and Writing*, *25*(6), 1239–1258.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Revicki, D.A., & Cella, D.F. (1997). Health status assessment for the twenty-first century: Item response theory, item banking and computer adaptive testing. *Quality of Life Research*, *6*(6), 595–600.

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

Ritchie, S.J., & Bates, T.C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science*, *24*(7), 1301–1308.

Salthouse, T.A. (1988). Effects of aging on verbal abilities: Examination of the psychometric literature. In L.L. Light & D.M. Burke (Eds.), *Language, memory, and aging* (pp 17–35). New York: Cambridge University Press.

Scarborough, H.S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S.B. Neuman & D.K. Dickinson (Eds.), *Handbook of early literacy research* (pp 97–110). New York: Guilford Press.

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.

Smith, B.I., Smith, T.D., Taylor, L., & Hobby, M. (2005). Relationship between intelligence and vocabulary. *Perceptual and Motor Skills*, *100*(1), 101–108.

University of Western Australia School of Psychology. (2011). *MRC Psycholinguistic Database*. Retrieved from http://www.psych.rl.ac.uk

Victorson, D., Manly, J., Wallner-Allen, K., Fox, N., Purnell, C., Hendrie, H.C., … Gershon, R.C. (2013). Using the NIH Toolbox in special populations: Considerations for the assessment of pediatric, geriatric, culturally diverse, non-English speaking and disabled individuals. *Neurology*, *80*(11 Suppl. 3), S13–S19.

Weintraub, S., Dikmen, S.S., Heaton, R.K., Tulsky, D.S., Zelazo, P. D., Bauer, P.J., … Gershon, R.C. (2013). Cognition assessment using the NIH Toolbox. *Neurology*, *80*(11 Suppl. 3), S54–S64.

Weintraub, S., Dikmen, S.S., Heaton, R.K., Tulsky, D.S., Zelazo, P.D., Slotkin, J., … Gershon, R. (2014). The cognition battery of the NIH Toolbox for assessment of neurological and behavioral function: Validation in an adult sample. *Journal of the International Neuropsychological Society*.

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

Wolf, M.S., Curtis, L.M., Wilson, E.A., Revelle, W., Waite, K.R., Smith, S.G., … Baker, D.W. (2012). Literacy, cognitive function, and health: Results of the LitCog study. *Journal of General Internal Medicine*, *27*(10), 1300–1307.