Published in final edited form as:
J Int Neuropsychol Soc. 2014 July ; 20(6): 611–619. doi:10.1017/S1355617714000460.

Measuring Episodic Memory Across the Lifespan: NIH Toolbox Picture Sequence Memory Test

Sureyya S. Dikmen[1], **Patricia J. Bauer**[2], **Sandra Weintraub**[3], **Dan Mungas**[4], **Jerry Slotkin**[5], **Jennifer L. Beaumont**[5], **Richard Gershon**[5], **Nancy R. Temkin**[6], and **Robert K. Heaton**[7]

[1]Department of Rehabilitation Medicine, University of Washington, Seattle, Washington

[2]Department of Psychology, Emory University, Atlanta, Georgia

[3]Cognitive Neurology and Alzheimer's Disease Center, Northwestern Feinberg School of Medicine, Chicago, Illinois

[4]Department of Neurology, University of California, Davis, California

[5]Department of Medical Social Sciences, Northwestern University, Chicago, Illinois

[6]Departments of Neurological Surgery and Biostatistics, University of Washington, Seattle, Washington

[7]Department of Psychiatry, University of California, San Diego, California

Abstract

Episodic memory is one of the most important cognitive domains that involves acquiring, storing and recalling new information. In this article, we describe a new measure developed for the NIH Toolbox, called the Picture Sequence Memory Test (PSMT) that is the first to examine episodic memory across the age range from 3 to 85. We describe the development of the measure and present validation data for ages 20 to 85. The PSMT involves presentation of sequences of pictured objects and activities in a fixed order on a computer screen and simultaneously verbally described, that the participant must remember and then reproduce over three learning trials. The results indicate good test–retest reliability and construct validity. Performance is strongly related to well-established "gold standard" measures of episodic memory and, as expected, much less well correlated with those of a measure of vocabulary. It shows clear decline with aging in parallel with a gold standard summary measure and relates to several other demographic factors and to self-reported general health status. The PSMT appears to be a reliable and valid test of episodic memory for adults, a finding similar to those found for the same measure with children.

Keywords

Episodic memory; Learning; Test development; NIH toolbox; Validation; Cognition

Correspondence and reprint requests to: Sureyya S. Dikmen, Department of Rehabilitation Medicine, University of Washington, 325 Ninth Avenue, Box 359612, Seattle, WA 98104. dikmen@uw.edu.

INTRODUCTION

Learning and memory for new information are essential components of cognition that undergo a protracted course of development and exhibit decline even in healthy aging. Acquiring information on the basis of one or more episodes of learning is the foundation for formation of memories ranging from the mundane such as whether milk was on the grocery list to the significant events that constitute one's personal life story. In this article, we describe the development and psychometric properties in adulthood of a new measure of episodic memory, a form of declarative memory. It is a component of the non-proprietary NIH Toolbox Cognition Battery (NIHTB-CB) and has undergone a U.S. national norming for both English and Spanish speakers on 4700 subjects ranging in age from 3 to 85 years old.

Declarative memory is specialized for rapid, even one-trial, learning of new information for later purposeful conscious recollection. There are two types of declarative memory: semantic and episodic memory (e.g., Squire, 2004). Semantic memory is specialized for storage of world knowledge not tied to a specific event, time, or place. It remains largely intact in aging and is relatively resistant to neurological insult. Episodic memory, in contrast, is specialized for storage of unique events or experiences associated with a specific time and place. Episodic memory is relatively fragile and susceptible to decay and interference over time. It also declines with normal aging and is vulnerable to many diseases and injuries that affect the brain.

Importance of Episodic Memory in Adulthood

Episodic memory undergirds the growth of knowledge in development and is the source of continuity of self-concept over the lifespan. Decline in episodic memory function is one of the most frequently lamented consequences of normal aging. Behavioral research reveals a consistent pattern in older adults of decrements in episodic learning and memory for lists of words, text, contextual details, faces, abstract visual materials, and televised news stories (Park & Gutchess, 2005). Memory decline in normal and abnormal aging causes outcomes that range from a reduction in confidence in the integrity of one's memory to restrictions in independent living. The importance of episodic memory explains why it is the most frequently measured form of memory and why it is included in the NIHTB-CB.

Evidence Linking Episodic Memory to Different Brain Networks

Episodic memory, initially linked with integrity of the hippocampus, is now known to be supported by a more extensive cortical network including not only components of the medial temporal lobe (MTL) but also of the prefrontal cortex (PFC), lateral temporal neocortex, and posterior parietal regions (see Dickerson and Eichenbaum, 2010, for review). The role of the hippocampus and other medial temporal lobe structures such as the entorhinal, perirhinal, and parahippocampal cortices has been ascribed to tagging and binding events with respect to "what," "where," and "when" features of an event. Each of the MTL and PFC components also has a role in encoding as well as retrieval, albeit with different weights depending on the task, as demonstrated by functional imaging studies.

The process of new learning that initiates encoding into episodic memory begins when the elements that constitute an event register across primary sensory areas (visual, auditory, somatosensory). The primary cortical inputs are projected to unimodal association areas, where they are integrated into whole percepts of how objects look, sound, and feel. The unimodal association areas in turn project their inputs to poly-modal posterior, limbic, and prefrontal association cortices, where inputs from the different sense modalities are integrated and maintained over brief delays (seconds; e.g., Petrides, 1995). For maintenance beyond mere seconds, the inputs must be stabilized and integrated into long-term storage sites, tasks attributed to medial temporal structures, in concert with cortical areas (McGaugh, 2000). Demands on this network are especially high in tasks that require free recall *versus* recognition, and memory for temporal order information *versus* memory for items alone (e.g., Shimamura, Janowsky, & Squire, 1990).

Individuals with lesions of the neural network that supports episodic memory have impairments in learning new information. As noted earlier, normal aging also is associated with decrements in episodic memory, perhaps related to small declines in the volume of medial-temporal and prefrontal structures that support the function (as measured by MRI; reported in Raz, 2005). Substantially accelerated atrophy of these structures is noted in Alzheimer's disease (e.g., Jack et al., 2000), and the hippocampus is one structure affected early in the progression of the disease (Grady, 2005).

Normal aging and Alzheimer's disease represent conditions with more specific disorders of learning and memory and associated brain structures. Episodic memory problems are the most common and sensitive indicators of injury or disease of the brain in general. These conditions include traumatic brain injury, stroke, multiple sclerosis, Huntington's disease, and others. Learning and memory dysfunction in such conditions often occur in association with other disorders of cognitive function. Nevertheless, learning and memory problems are the most common initial complaint, with episodic memory representing one of the most sensitive cognitive constructs for the neurologic integrity of the brain.

In this article, we describe the development and validation of a new measure of episodic learning and memory for the NIHTB-CB in adults (ages 20 to 85 years). The validation of this measure for ages 3 to 15 years is published elsewhere (Bauer et al., 2013). Importantly, a mandate for the development of all NIHTB-CB measures is that they all be applicable throughout the lifespan, from ages 3 to 85 years (Weintraub et al., this issue). As such, the Picture Sequence Memory Test (PSMT) is the first measure of episodic memory that can be used for such a broad age range. Consistent with previously reported findings with children (Bauer et al., 2013) we expected the PSMT to show good test–retest reliability, and convergent and divergent validities compared to relevant gold standard measures. Although with children PSMT performance improved consistently during the course of development, with adults we expected age-related decline. Finally, we expected that lower PSMT performance would be associated with self-reported prior academic difficulties and worse general health status.

METHODS

NIH –Toolbox Measurement of Episodic Memory

There were several challenges posed by the NIH charge of developing a test of episodic learning and memory that will be useful across the wide age span of 3 to 85 years. Most measures of episodic memory rely heavily on verbal skills (requiring comprehension of complex verbal instructions, a verbal response, or both), which made them inappropriate as measures of the construct in early childhood. Accordingly, the decision was made to adapt for use with older participants a visually based task that originally had been used in experimental work with very young children (including infants). This approach uses elicited and deferred imitation (props are used to produce a specific action or sequence of actions that the infant is required to imitate either immediately, after a delay, or both; e.g., Bauer & Mandler, 1989; Bauer & Shore, 1987; Bauer, Wenner, Dropik, & Wewerka, 2000; see Bauer, 2005, 2006, 2007). The task is an accepted analogue to verbal report (Bauer, 2007). In support of this conclusion, individuals with documented lesions to medial-temporal structures involved in episodic encoding and recall show impaired performance on age-appropriate versions of the task (e.g., Adlam, Vargha-Khadem, Mishkin, & de Haan, 2005; McDonough, Mandler, McKee, & Squire, 1995). In addition, performance on imitation-based tests of memory in infancy are correlated with standardized measures of declarative memory in later childhood (Riggins, Cheatham, Stark, & Bauer, 2013).

Another decision was to use procedures commonly used in assessing episodic learning and recall in adults. This included using sequence lengths that would exceed immediate normal working memory span (i.e., “supra-span”) and using multiple learning trials to engage episodic memory and improve test–retest reliability. Task difficulty for the various age groups was designed and tested with this concept in mind. Due to significant time constraints imposed on the length of NIHTB-CB (not to exceed 30 min in total) delayed recall was not examined.

Development Phases of PSMT

The PSMT involves sequences of pictures of objects and activities that are presented in a fixed order, with the content of each picture simultaneously orally described, which the subject is required to reproduce after each presentation. Four waves of pretesting were carried out to determine appropriate task difficulty for the different age groups. Task difficulty was manipulated and examined by two methods. One method involved the level of connectivity of the picture sequences, which refers to the degree to which the order of the pictured objects or activities is logical or meaningful (e.g., bake a cake, before applying icing). A sequence with high connectivity would be easier to remember because there is a logical order to the sequence of pictures. In contrast a series with low connectivity would be more difficult to remember because there is no inherent constraint on the order of the pictures that the subject needs to remember and reproduce.

The other method of manipulating task difficulty involved determining appropriate sequence length (number of items within a sequence) for the various age groups to exceed immediate working memory span but also did not overwhelm the youngest and oldest participants.

Finally, to achieve acceptable psychometric properties, we examined the effect of number of exposures (single *vs*. multiple), reproducibility (test–retest correlations), and association with gold standard validation measures. Additional dimensions investigated included the scoring method (e.g., sum of correct adjacent pairs reproduced *vs*. percent correct adjacent pairs reproduced of total possible).

Findings of the Development Phases

Four Beta versions were created in stages and tested in groups of subjects to create the final versions of the NIHTB-CB measures including the PSMT (Weintraub et al., 2014) On the basis of the findings of the Development Phases, three forms of the task were developed as alternate versions of PSMT to reduce practice effects. The forms selected were those with low connectivity, making them more difficult so as to allow greater variability in performance among subjects. The general themes of the three forms are “Working on the Farm,” “Playing in the Park,” and “Going to the Fair.” Level of task difficulty for the various age groups was controlled by varying the number of pictures in a sequence to avoid floor and ceiling effect. Based on the results (including those with children presented elsewhere; Bauer et al., 2013), the following sequence lengths were selected for the different age groups: For ages 3–4 years: 6 pictures; 5–7 years: 9 pictures; 8 years: 12 pictures; 9–60 years: 15 pictures, and 61–85 years: 9 pictures. This article focuses on ages 20 to 85. To reduce the likelihood of ceiling effects, the sequence length of 15 for ages 20 to 60 was increased to 18 on the 2nd and 3rd trials if the subject obtained a score of 14 correct “adjacent pairs” (a maximum score) on the first trial. Similarly for ages 65 to 85 years, sequence length was increased to 12 from 9 for the second and 3rd trials if the subject received a ceiling score of 8 correct adjacent pairs on the first trial. Additional decisions included using multiple trials (i.e., three trials) to improve test–retest reliability, and using raw scores based on the sum of adjacent pairs correctly reproduced over three repeated trials, regardless of the initial sequence length or increase to the sequence length because of attaining the maximum score on the first trial. On the basis of the results of the development phases, the protocol for PSMT was developed and subjected to psychometric analyses (“Validation Study”) described below.

VALIDATION STUDY

Participants

The participants of the validation study are described in detail in the first study of this series by Weintraub et al. (2014). Briefly, the sample included a total of 268 normal adults. There were 119 males and 149 females. The sample included 148 non-Hispanic Caucasians, 75 African Americans, and 45 Hispanics. These subjects were recruited from four sites and proper consents were obtained in accordance with the requirements of the relevant Institutional Review Boards. Eighty-nine (approximately 1/3) randomly selected subjects were re-tested to examine test–retest reliability (M = 15.5 days; SD = 4.8 days; Range = 7–26 days). Within the total group of adults, 159 were between the ages of 20 and 60 years and received the 15-item sequence length, while 109 between 65 and 85 years and were administered the 9-item sequence length. Four percent of those 20 to 60 years, and only one subject in the age range of 65–85 years had a ceiling score on the first trial. With both age

ranges sequence length on the 2nd and 3rd trials was increased if the subjects scored at ceiling as described above.

Test Procedure

Measures of episodic memory require new learning in the context of the test. To perform on the PSMT, new information about the order of the pictures must be learned and then reproduced immediately after each of three consecutive exposures. The amount of time across trials and the supra-span amount of information to be learned exceeds those of short-term or working memory, emphasizing episodic memory components of encoding. Although highly desirable, we were unable to measure delayed recall under the time constraints of the total NIHTB-CB administration time (i.e., not to exceed 30 min).

A computer touch screen was used for the administration of the task in the validation study. To orient participants to the PSMT requirements, two to three practice sequences were administered first (see Figure 1). The purpose of the practice sequences was to inform the subject of task requirements, and to provide experience moving the pictures on the computer screen into the correct position in the sequence. A trained examiner was present throughout the testing session to make sure that the subjects understood the requirements of the task and put forth good effort.

The PSMT involved color-illustrated sequence of pictures which appear one at a time in the center of the computer screen in a fixed order. As each picture appears, a recording briefly describes its content. Once described, the picture reduces in size and is moved to its fixed position in the sequence, making way for the next picture. This is followed by the next picture in the sequence without delay. This continues until all pictures in a sequence have been displayed and placed in their proper positions which the subject observes. Once all the pictures in the sequence are displayed the pictures then are placed in a random spatial array at the center of the screen. The task of the subject is to move each picture from the center to its correct location to replicate the correct sequence. Exposure to each picture, its description and being placed in its proper position in the sequence was approximately 5 s. Thus, for the 15-item sequence, task exposure/description time per trial is approximately 1.25 min. How long the subject took to perform the task varied. Performance was measured by the number of correct adjacent pairs reproduced, not by the time it took to complete the task. For the study described here, three trials were administered to improve test score variability and test–retest reliability (Strauss, Sherman, & Spreen, 2006).

The three forms (“Working on the Farm,” “Playing in the Park,” and “Going to the Fair”) were randomly assigned to subjects. As mentioned previously, ages 20 to 60 years were administered 15-picture sequences and those 65 to 85 years were given nine-picture sequences of the same forms, with additional pictures added to the end of the sequence on the 2nd and 3rd trials in case of ceiling score on the first trial. Three trials were administered with recall after each exposure The participant’s score on the PSMT was the cumulative number of adjacent pairs of pictures remembered correctly over the 3 learning trials regardless of the number of pictures in the sequence. Adjacent pairs are two adjacent pictures placed in the correct consecutive, ascending order. Thus, pictures placed in the orders 3–4 and 5–6 would receive credit, whereas pictures placed in the orders 1–5 and 3–9

would not receive credit. For each trial, the possible number of adjacent pairs is one less than the number of pictures in the sequence. The total possible number of adjacent pairs is the sum of the adjacent pairs scores across three trials with a maximum raw score of 48 for ages 20–60 years and 30 for ages 65–85 years. There were no significant differences in the difficulty level among the three forms ($F(2,262) = .97$; $p = .380$). Total administration time for the age range 20–60 years was a mean of 10 min ($SD = 1.7$) and for the age range 65–85 years it was 8.1 min ($SD = 2.6$ min).

Measures of Convergent Validity

Rey Auditory Verbal Learning Test (RAVLT)—The RAVLT is one of the most widely studied measure of memory, has good psychometric properties, extensive normative data, and has been used in different languages, cultures and ethnic groups (Lezak, 1983; Strauss et al., 2006). The task requires learning a list of 15 unrelated words over five presentation trials and recalling them again after a delay. For the NIHTB validation, only three trials were administered and there was no delayed recall trial, making it comparable to the administration of the PSMT. The score used for validation was the sum of the words recalled over three trials.

Brief Visuospatial Memory Test-Revised (BVMT-R)—This is a widely used measure of visual memory with very good psychometric properties (Benedict, 1997). There are three learning trials in which six geometric figures are viewed and then immediately reproduced from memory. It is scored both for the accuracy of the designs, as well as their location on the page. The score used for the validation study was the sum of scores over three trials.

Measure of Discriminant Validity

Peabody Picture Vocabulary Test-4th edition (PPVT-4)—This is an individually administered, untimed test of receptive vocabulary that uses a multiple-choice non-verbal response format. The participant must select one among four pictures that best represents an orally presented stimulus word. This is one of the oldest and most commonly used standardized tests of vocabulary (Dunn & Dunn, 2007).

Data Analysis

As noted above, the scores for PSMT, RAVLT, and for BVMT-R were the sum of correct scores over three learning trials. For all of these tests scaled scores were created by first ranking the raw scores for all participants between the ages of 20 and 85 years, 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. These normalized and unadjusted scaled scores were used in all analyses. When examining the effects of various demographic factors and other indices of ecological validity, these unadjusted scaled scores were adjusted using regression as described below. Pearson correlation coefficients between age and test performances were calculated to assess the ability of both the PSMT and the gold standard measures to detect cognitive decline during adulthood. Intraclass correlation coefficients (ICC) and Pearson correlation with 95% confidence intervals were calculated to evaluate test–retest reliability. The practice effect was evaluated using paired *t* tests and the effect size estimated by dividing the mean change

by the baseline standard deviation. Convergent validity was assessed with correlations between PSMT scores and the scores derived from the gold standard measures of the same construct, namely BVMT-R and RAVLT. A third score was calculated from the average of the BVMT-R and RAVLT scaled scores and was also compared with the PSMT score. Discriminant validity was assessed with correlations with the PPVT-4. Other demographic comparisons in relation to test performance were then performed using linear regression to examine associations with performance, adjusted for age, gender, education, and race/ethnicity except for the one being examined. Effect sizes are reported as Cohen's d, with .20, .50, and .80 exemplifying small, medium, and large effects, respectively.

RESULTS

Equivalence of Alternate Forms

Participants were randomly assigned to one of the three forms of the PSMT: the Fair, Farm, or Park. The groups did not differ with respect to age, education, race/ethnicity, or gender. There was no significant difference in performance among the three alternate forms ($p = .380$).

Test–Retest Reliability and Practice effects

Approximately one-third of the subjects ($N = 89$) were retested (mean test–retest interval 15.5 days, $SD = 4.8$ days; range 7–26 days). The subjects were independently assigned by a computer generated program to one of the three forms of the PSMT, with 30 receiving the same form on retest. The RAVLT and BVMT-R forms were the same on both evaluations. The test–retest reliability of the PSMT, pooled over the three forms, was 0.84 as measured by Pearson correlation and 0.77 as measured by ICC, both of which are excellent for measures of episodic memory. Similar values were obtained for RAVLT (0.85 Pearson, 0.75 ICC) and BVMT-R (0.82 Pearson, 0.75 ICC). Reliability estimates for individual forms of the PSMT were not calculated due to small sample sizes. There were significant practice effects on both the PSMT and the gold standard measures. The effect size (mean change/baseline *SD*) was 1.24/2.9 = 0.42 for the PSMT. For the combination of RAVLT and BVMTR, it was 1.55/3.2 = .49.

Construct Validity

The PSMT was strongly correlated with the RAVLT, the BVMT-R, and their combination (see Table 1, r = .64 to .72), reflecting good convergent validity. The close relationship with both the RAVLT and the BVMT-R likely reflect the combined, visual and verbal aspects of PSMT. In contrast, there was no significant correlation between PSMT and PPVT-4 scores (see Table 1, r = .04), representing good discriminant validity.

EFFECTS OF DEMOGRAPHIC FACTORS

Age

Table 2 provides adjusted means and standard errors for nine age groups and shows the effect size comparing each older age group to those 20–24 years old. Scores decrease consistently with age. The age groups differed significantly whether all nine groups are

compared or only the five subgroups with ages between 20 and 60 years that had the same number of test items (each $p < .001$). Effect sizes are large (−.78 to −2.29) for comparisons of groups 40–49 or older to those 20–24. The associations between age and test scores for both the PSMT and gold standard average of BVMT-R and RAVLT are shown in Figure 2. Both the PSMT and the gold standard measures show strikingly similar patterns and negative associated correlation coefficients with increasing age (r = −.63 for PSMT, r = − 54 for RAVLT, r = −.59 for BVMT-R, and r = −.64 for the combination of the two gold standard memory measures).

Effects of Other Demographic Factors and General Health Status and PSMT

Table 3 shows means, standard errors and effect sizes for other demographic factors likely to influence cognitive performance. A significant difference in PSMT and average gold standards score was found between males and females, with small to medium effect sizes of 0.26–0.35. Females scored better than males on both sets of measures. Similarly, college graduates scored better than those with high school education or less on both measures. The effect sizes comparing college graduates to those who had completed high school or less than high school were 0.32–0.39 (small to medium effect size), values comparable to those of the averaged gold standard measures. Average BVMT-R and RAVLT scores differed somewhat according to race/ethnicity whereas these differences for the PSMT were not significant.

A significant difference in PSMT and a trend in average of gold standard measures was found for self-reported performance problems in school with small effect sizes (.19–.24). Those reporting any (i.e., performing below average, failing a grade, repeating a grade, or receiving special classes or tutoring) scored lower than those without any problems.

Better overall health as reported by subject was associated with better PSMT and gold standard scores, with small effect sizes (.17–.26) comparing excellent health with good health, and medium effect sizes (.48–.49) comparing excellent health to fair/poor health (see Table 3).

DISCUSSION

The results of this study support the conclusion that the NIHTB-CB PSMT is a reliable and valid test of episodic memory for adults ages 20 to 85 years. All psychometric properties that were assessed for the PSMT were quite comparable to those of the RAVLT and BVMT-R, considered gold standard measures of episodic memory in the field. The test–retest reliability over 2 weeks is .84, which is considered high for a measure of episodic memory (Strauss et al., 2006). Its construct validity is supported by strong correlations with the gold standard measures of episodic memory (.64–.72) and an almost zero relation to the PPVT-4, a test of vocabulary (semantic memory). The three alternate forms of the PSMT appear to be of comparable difficulty.

Construct validity of the PSMT as a measure of episodic memory was tested in a broader context in a companion article in this series by Mungas et al. (2014). That study included a confirmatory factor analysis of Toolbox and gold standard measures from all domains. The

PSMT, RAVLT, and BVMT-R together defined a common episodic memory factor, with loadings above 0.80 for the PSMT and BVMT and greater than 0.75 for RAVLT. The PSMT did not have any significant cross-loadings on other factors, including working memory or executive function factors (Mungas et al., 2014).

Consistent with more than two decades of literature regarding the effects of aging on episodic memory, PSMT scores show the expected gradual and fairly linear decline with increasing age, supporting its criterion validity. Declining scores were noted in the age decade of the 30s with a consistent decline over time for both the Toolbox measure and the gold standard measures.

It is important to note that the sequence length for those from 20 to 60 was 15 pictures (18 items on the 2nd and 3rd if the subject ceilinged on the 1st trial), while those from 65 to 85 had 9 pictures (12 items on the 2nd and 3rd trials in the event of ceiling on the 1st trial). This complicates the examination of age effects over the entire range, but a strong decrease in performance with increasing age is seen even within the 20 to 60 age range that had the same sequence length. Perhaps the strongest evidence that different sequence lengths did not substantially bias age effects is the general concordance of age effects for the PSMT and the two gold standard tests. Sequence length differences will be addressed in a more formal manner in the final version of the Toolbox where item response theory scoring methods will be used to adjust scores for sequence length effects.

Practice effects with re-testing were moderate and could not be closely examined due to the small sample size that was re-tested and the number of forms involved. Both difficulty level of the three forms and practice effects with the same and alternate forms will be more fully examined in the large national norming sample.

Other than age, the effects of various demographic factors on PSMT performance were relatively small, generally consistent with those of the gold standard measures used, and with the literature (Strauss et al., 2006). This was true for gender, with females performing better than males. With respect to education, a significant overall effect was seen for both the PSMT and the gold standard measures. After controlling for relevant demographic variables (gender, age, and education), there was no significant race/ethnicity effect for PSMT, but small to medium effect sizes were observed for the gold standard measures). Finally, in terms of ecological validity, participants reporting performance problems in school and/or worse general health status did less well than those that did not report problems on the PSMT and showed a similar trend on the gold standard measures.

Whereas race/ethnicity effects for the gold standard measures are consistent with the literature, those for the PSMT using the same subject pool are not. The suggested race/ethnicity neutrality is important and highly desirable and will be explored further in the national norming sample. A potential limitation of the PSMT is not having delayed recall due to time constraints of the entire cognition battery. However, it is important to note that the correlation between initial learning and delayed recall in normals as well as individuals with neurologic conditions is reported to be high (Heaton, Taylor, & Manly, 2003; Wechsler, 2009). Furthermore, learning trials and delayed recall have been shown to load

heavily on a single factor in both normals and patients with neurologic conditions as illustrated on the California Verbal Learning Test (Delis, Kramer, Freeland, & Kaplan, 1988).

The PSMT uses stimulus materials and responses relatively familiar to U.S. examinees at all ages and with a broad range of cultural, educational, and linguistic backgrounds. However, the pictorial stimuli of the PSMT are not universal and their appropriateness for cross-cultural work in other areas of the world (e.g., developing economies) is uncertain, at best. Nevertheless, the general method of assessing picture sequence learning in a computer assisted format could be rather easily adapted for use with other populations.

In conclusion, the PSMT appears to be a reliable and valid measure of episodic memory in adults, and in the pediatric population (Bauer et al., 2013). The NIH Toolbox for the Assessment of Neurological and Behavioral Function, which includes Sensory, Motor, and Emotional health batteries as well as the Cognition Battery has undergone norming with a sample size of 4700 subjects, and includes both English and Spanish speakers. Individually and as part of the four batteries, the PSMT has several advantages. First, it can be used with ages 3 to 85. To our knowledge, this is the only episodic memory measure that covers such an age range with norming data. Second, it has three alternate forms that appear to be of comparable difficulty and, with further validation based on the norming sample, can be used to reduce the magnitude of practice effects in longitudinal studies. Third, large and comprehensive demographically corrected norming data will be available to improve the sensitivity of all the Toolbox measures to health-related changes in functioning. Fourth, all four batteries of the Toolbox and the measures within them are being co-normed on the same sample, allowing comparisons among domains and among measures of different constructs in studies of various health conditions. Finally, and very importantly, the goal of the NIH Toolbox is to encourage the use of the measures in epidemiological studies and clinical trials within and across different health conditions, to allow comparisons of results and accumulation of knowledge, removing differences in measures as an explanation for differences in findings.

Acknowledgments

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. *Disclaimer:* The views and opinions expressed in this article are those of the authors and do not necessarily represent the views of the National Institute on Drug Abuse, the National Institutes of Health, or any other governmental agency. *Disclosures: Dr. Dikmen* receives research grant funding from NIH R01 NS058302 and R01HD061400, NIDRR H133A080035, NIDRR H133G090022, and NIDRR, H133A980023, and DoD W81XWH-0802-0159. *Dr. Bauer* serves as a member of the editorial board for the journal Journal of Experimental Child Psychology, as Associate Editor for the journals Developmental Review and Memory, and as Editor of the Monographs of the Society for Research in Child Development, for which she receives a stipend. She has received royalties from the publication of Memory in Infancy and Beyond (2007, Erlbaum), and Advances in Child Development and Behavior (Volumes 37 and 38, 2009 and 2010, respectively; Elsevier); and is funded by NIH grants HD067359, HD074724, and HD071845. *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. 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. *Dr. Slotkin* reports no disclosures. *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. Gershon* has received personal

compensation for activities as a speaker and consultant with Sylvan Learning, Rockman, and the American Board of Podiatric Surgery. He has several grants awarded by NIH: N01-AG-6-0007, 1U5AR057943-01, HHSN260200600007, 1U01DK082342-01, AG-260-06-01, HD05469, NINDS: U01 NS 056 975 02, NHLBI K23: K23HL085766 NIA; 1RC2AG036498-01; NIDRR: H133B090024, OppNet: N01-AG-6-0007. *Dr. Temkin* is funded by grants from NIH, CDC, NIDRR, DOD, VA, and MS Society. She serves on Data and Safety Monitoring Committees for several pharmaceutical companies. *Dr. Heaton* is funded by NIH grants # P30MH062512, HHSN271201000036C, R01MH92225, R01MH094160, and P50DA026306. He is on the editorial board of the Journal of the International Neuropsychological Society and The Clinical Neuropsychologist.

References

Adlam AL, Vargha-Khadem F, Mishkin M, de Haan M. Deferred imitation of action sequences in developmental amnesia. Journal of Cognitive Neuroscience. 2005; 17:240–248. [PubMed: 15811236]

Bauer, PJ. New developments in the study of infant memory. In: Teti, DM., editor. Blackwell handbook of research methods in developmental science. Oxford, United Kingdom: Blackwell Publishing; 2005. p. 467-488.

Bauer PJ. Constructing a past in infancy: A neuro-developmental account. Trends in Cognitive Sciences. 2006; 10:175–181. [PubMed: 16537115]

Bauer, PJ. Remembering the times of our lives: Memory in infancy and beyond. Mahwah, NJ: Erlbaum; 2007.

Bauer PJ, Dikmen S, Heaton R, Mungas D, Slotkin J, Beaumont JL. III. NIH Toolbox Cognition Battery (CB): Measuring episodic memory. Monographs of the Society for Research in Child Development. 2013; 78:34–48. [PubMed: 23952201]

Bauer PJ, Mandler JM. One thing follows another: Effects of temporal structure in one- to two-year-olds' recall of events. Developmental Psychology. 1989; 25:197–206.

Bauer PJ, Shore CM. Making a memorable event: Effects of familiarity and organization on young children's recall of action sequences. Cognitive Development. 1987; 2:327–338.

Bauer PJ, Wenner JA, Dropik PL, Wewerka SS. Parameters of remembering and forgetting in the transition from infancy to early childhood. Monographs of the Society for Research in Child Development. 2000; 65:1–204.

Benedict, R. Brief Visuospatial Memory Test-Revised professional manual. Odessa, FL: Psychological Assessment Resources, Inc; 1997.

Delis DC, Kramer JH, Freeland J, Kaplan E. Integrating clinical assessment with cognitive neuroscience: Construct validation of the California Verbal Learning Test. J Consulting and Clinical Psychology. 1988; 56(1):123–130.

Dickerson BC, Eichenbaum H. The episodic memory system: neurocircuitry and disorders. Neuropsychopharmacology. 2010; 35:86–104. [PubMed: 19776728]

Dunn, DM.; Dunn, LM. PPVT-4: Peabody Picture Vocabulary Test. 4. Minneapolis, MN: Pearson; 2007.

Grady, CL. Functional connectivity during memory tasks in healthy aging and dementia. In: Cabeza, R.; Nyberg, L.; Park, D., editors. Cognitive neuroscience of aging: Linking cognitive and cerebral aging. New York: Oxford University Press; 2005. p. 286-308.

Heaton, RK.; Taylor, MJ.; Manly, J. Demographic effects and use of demographically corrected norms with the WAIS-III and WMS-III. In: Tulsky, D.; Saklofske, D.; Heaton, RK.; Chelune, G.; Ivnik, R.; Bornstein, RA.; Ledbetter, M., editors. Clinical interpretation of the WAIS-III and WMS-III. San Diego: Academic Press; 2003. p. 183-210.

Jack CR, Petersen RC, Xu Y, O'Brien PS, Smith GE, Ivnik RJ, Kokmen E. Rates of hippocampal atrophy correlates with change in clinical status in aging and AD. Neurology. 2000; 55:484–489. [PubMed: 10953178]

Lezak, MD. Neuropsychological Assessment. 2. New York: Oxford University Press; 1983.

McDonough L, Mandler JM, McKee RD, Squire LR. The deferred imitation task as a nonverbal measure of declarative memory. Proceedings of the National Academy of Sciences of the United States of America. 1995; 92:7580–7584. [PubMed: 7638234]

McGaugh JL. Memory - A century of consolidation. Science. 2000; 287:248–251. [PubMed: 10634773]

Mungas D, Heaton RK, Tulsky D, Zelazo P, Slotkin J, Blitz D, Gershon R. Factor structure, convergent validity, and discriminant validity of the NIH toolbox cognitive health battery (NIHTB-CHB) in adults. Journal of the International Neuropsychological Society. 2014 In this issue. 10.1017/S1355617714000307

Park, DC.; Gutchess, AH. Long-term memory and aging: A cognitive neuroscience perspective. In: Cabeza, R.; Nyberg, L.; Park, D., editors. Cognitive neuroscience of aging: Linking cognitive and cerebral aging. New York, NY: Oxford University Press; 2005. p. 218-245.

Petrides M. Impairments on nonspatial self-ordered and externally ordered working memory tasks after lesions of the mid-dorsal part of the lateral frontal cortex in monkeys. The Journal of Neuroscience. 1995; 15:359–375. [PubMed: 7823141]

Raz, N. The aging brain observed in vivo: Differential changes and their modifiers. In: Cabeza, R.; Nyberg, L.; Park, D., editors. Cognitive neuroscience of aging: Linking cognitive and cerebral aging. New York, NY: Oxford University Press; 2005. p. 19-57.

Riggins T, Cheatham C, Stark E, Bauer PJ. Elicited imitation performance at 20 months predicts memory abilities in school age children. Journal of Cognition and Development. 2013; 14:593–606. [PubMed: 24436638]

Shimamura AP, Janowsky JS, Squire LR. Memory for the temporal order of events in patients with frontal lobe lesions and amnesic patients. Neuropsychologia. 1990; 28:803–813. [PubMed: 2247207]

Squire LR. Memory systems of the brain: A brief history and current perspective. Neurobiology of Learning and Memory. 2004; 82:171–177. [PubMed: 15464402]

Strauss, E.; Sherman, EMS.; Spreen, O. A compendium of neuropsychological tests: Administration, norms, and commentary. Oxford; New York: Oxford University Press; 2006.

Wechsler, D. WMS-IV Technical and Interpretative Manual. San Antonio, TX: Pearson; 2009.

Weintraub S, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Slotkin J, Gershon R. 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. 2014 In this issue. 10.1017/S1355617714000307

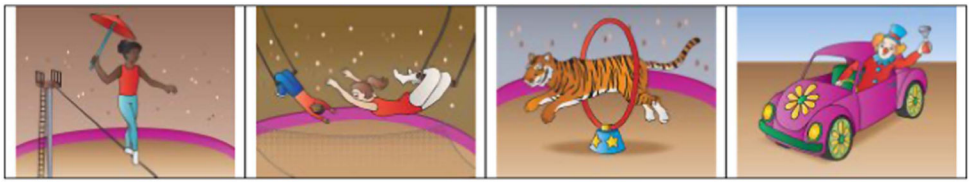

Fig. 1.

Four-step practice sequence with “Circus” theme: walk a tightrope, swing on the trapeze, jump through the hoop, and drive the funny car. (Used with permission © 2012 National Institutes of Health and Northwestern University)

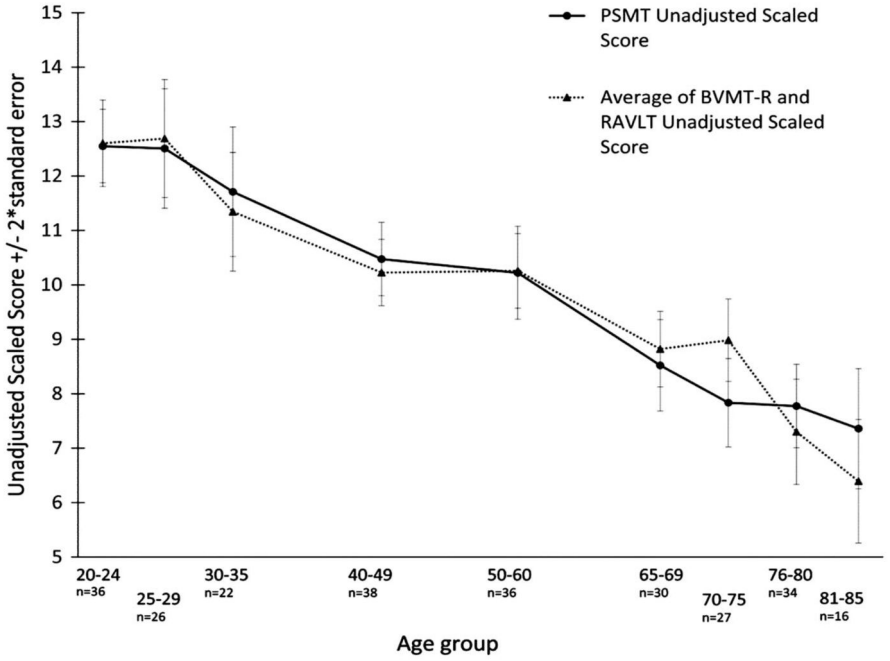

Fig. 2.

Distribution of PSMT and Average Gold Standard Unadjusted Scaled Scores by Age. Symbol marks the mean and lines extend one standard error.

Table 1

Pearson correlation coefficients between PSMT and gold standard measures

	PSMT
RAVLT	0.64
BVMT-R	0.65
Both (Average of RAVLT and BVMT-R)	0.72
PPVT-4	0.04

Table 2

Associations of PSMT versus average of BVMTR and RAVLT with age.

	PSMT* (Overall *SD* = 3.0)		Average of BVMT & RAVLT** (Overall *SD* = 3.0)	
Age	**Adjusted mean (*SE*)**	**ES*** vs. 20 to 24**	**Adjusted mean (*SE*)**	**ES*** vs. 20 to 24**
20 to 24	12.7 (0.4)	Ref	12.7 (0.4)	Ref
25 to 29	12.6 (0.5)	− 0.03	12.6 (0.5)	− 0.03
30 to 35	11.6 (0.5)	− 0.36	11.2 (0.5)	− 0.50
40 to 49	10.4 (0.4)	− 0.78	10.1 (0.4)	− 0.87
50 to 60	10.0 (0.4)	− 0.88	10.1 (0.4)	− 0.88
65 to 69	8.1 (0.4)	− 1.55	8.3 (0.4)	− 1.47
70 to 75	7.2 (0.5)	− 1.82	8.2 (0.5)	− 1.50
76 to 80	7.5 (0.4)	− 1.75	6.8 (0.4)	− 1.96
81 to 85	7.1 (0.6)	− 1.89	5.9 (0.6)	− 2.29
p (overall)	<.001		<.001	
p (ages 20–60)	.001		<.001	

Note. Means were adjusted for gender, race/ethnicity, and education using linear regression. The adjusted means reflect the average of the predicted scaled score for the various categories of the controlled variables.

* PSMT = highest possible scaled score for ages 20–60 = 16.84, highest possible scaled score for ages 65–85 = 12.14.

** Average of BVMT & RAVLT = highest possible scaled score for ages 20–85 = 18.46.

*** ES (effect size) = difference in adjusted means/overall SD. Ref indicates reference category for effect size.

Table 3

Associations of PSMT versus average of BVMTR and RAVLT with various demographics, overall health, and school problems

	PSMT		Average of BVMT & RAVLT	
	Adjusted mean (*SE*)	**ES* vs. Ref**	**Adjusted mean (*SE*)**	**ES* vs. Ref**
Gender				
Male	9.3 (0.2)	Ref	9.3 (0.2)	Ref
Female	10.3 (0.2)	0.35	10.1 (0.2)	0.26
p (overall)	<.001		.009	
Race/ethnicity				
White	10.0 (0.2)	Ref	10.3 (0.2)	Ref
Black	9.6 (0.3)	− 0.14	9.4 (0.3)	− 0.29
Hispanic	9.7 (0.4)	− 0.12	9.3 (0.4)	− 0.34
p (overall)	.388		.006	
Education				
<HS	9.3 (0.3)	− 0.39	9.3 (0.3)	− 0.36
HS Grad	9.5 (0.2)	− 0.32	9.3 (0.2)	− 0.35
College	10.4 (0.2)	Ref	10.3 (0.2)	Ref
p (overall)	.002		.002	
Overall health				
Excellent/very good	10.1 (0.2)	Ref	10.0 (0.2)	Ref
Good	9.6 (0.3)	− 0.17	9.3 (0.3)	− 0.26
Fair/poor	8.6 (0.5)	− 0.49	8.6 (0.5)	− 0.48
p (overall)	.011		.003	
School problems				
None	10.2 (0.2)	Ref	10.0 (0.2)	Ref
Any school problems	9.4 (0.2)	− 0.24	9.4 (0.2)	−0.19
P (overall)	.018		.056	

Note. Means were adjusted for age, gender, race/ethnicity, and education, excluding the factor being examined, using linear regression. The adjusted means reflect the average of the predicted scaled score at the mean age for the various categories of the controlled variables.

*ES (effect size) = difference in adjusted means/overall *SD*. Ref indicates reference category for effect size.