JOURNAL OF CLINICAL AND EXPERIMENTAL NEUROPSYCHOLOGY
2010, i*First*, 1–11

Psychology Press
Taylor & Francis Group

Improving digit span assessment of short-term verbal memory

David L. Woods,[1,2,3,4] Mark M. Kishiyama,[1] E. William Yund,[1] Timothy J. Herron,[1] Ben Edwards,[1] Oren Poliva,[1] Robert F. Hink,[1] and Bruce Reed[2,5]

[1]Human Cognitive Neurophysiology Laboratory, VANCHCS, Martinez, CA, USA
[2]UC Davis Department of Neurology, Sacramento, CA, USA
[3]Center for Neurosciences, UC Davis, Davis, CA, USA
[4]UC Davis Center for Mind and Brain, Davis, CA, USA
[5]Alzheimer's Disease Center, VANCHCS, Martinez, CA, USA

We measured digit span (DS) in two experiments that used computerized presentation of randomized auditory digits with performance-adapted list length adjustment. A new mean span (MS) metric of DS was developed that showed reduced variance, improved test–retest reliability, and higher correlations with the results of other neuropsychological test results when compared to traditional DS measures. The MS metric also enhanced the sensitivity of forward versus backward span comparisons, enabled the development of normative performance criteria with subdigit precision, and elucidated changes in DS performance with age and education level. Computerized stimulus delivery and improved scoring metrics significantly enhance the precision of DS assessments of short-term verbal memory.

Keywords: Digit span; Metrics; Computer; Aging; Education; Executive function; Attention; Verbal memory; Auditory.

INTRODUCTION

Measures of forward and backward digit span (DS) are among the oldest and most widely used neuropsychological tests of short-term verbal memory (Richardson, 2007). For decades they have been a component of the widely used Wechsler memory scales (WMS) and Wechsler intelligence scales for adults and children (Wechsler, 1997a, 1997b). In each case, digit span is measured for forward- and reverse-order (backward) recall of digit sequences. Digit sequences are presented beginning with a length of 2 digits, and two trials are presented at each increasing list length. Testing ceases when the participant fails to accurately report either trial at one sequence length or when the maximal list length is reached (9 digits forward, 8 backward). The total number of lists reported correctly is combined across forward span (FS) and backward span (BS) to produce a Wechsler total correct score.

In traditional DS testing digit salience is influenced by two factors. First, because each list is read aloud by an examiner there are variations in the rate, intensity, emphasis, and clarity of digit enunciations on each presentation as well as variations in clarity between different examiners (Reeves, Schmauder, & Morris, 2000; Silverman, 2007). In addition, the digits in each list are not selected randomly. Certain digit sequences (e.g., the local telephone area code) may inflate digit span in geographic regions where particular digit sequences are overlearned and underestimate span where digit sequences conflict with previously overlearned strings (e.g., the digit sequence "415" in a region where the local telephone area code is "451").

There are also two problems in the methods used by the Wechsler Memory Scale–Third Edition (WMS–III; Wechsler, 1997b) to sample digit list lengths. First, WMS–III testing involves the presentation of two digit

Supported by grants from the Office of Research and Development, Rehabilitation R&D Service, Department of Veterans Affairs and by National Institutes of Health (NIH) Grant ES014038. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Address correspondence to David L. Woods, Neurology Service (127E), VANCHCS, 150 Muir Rd., Martinez, CA 95553; USA (E-mail: dlwoods@ucdavis.edu).

http://www.psypress.com/jcen
DOI: 10.1080/13803395.2010.493149

lists at each span length, beginning with 3 digits in FS testing and 2 digits in BS testing. This procedure is relatively inefficient for participants with normal memory spans (e.g., 7 digits in FS) because eight trials are delivered before the participant reaches list lengths that challenge memory capacity. A second, more serious problem relates to the suboptimal sampling of list lengths that bound the participants' maximal DS: Testing ceases as soon as the participant misses two lists of the same length. This procedure assumes that the participant's "true" maximum length (ML) span can be assessed by only four list presentations: two at the ML and two above. However, it may seriously underestimate the ML of participants who are distracted or who encounter idiosyncratically difficult digit strings (e.g., permutations of their telephone area code) at a particular length.

In addition, the standard Wechsler total correct metric of DS performance is problematic for two reasons. First, it conflates inconsistent performance with limits in maximal DS. For example, a participant with variable performance who misses one trial in FS testing at lengths of 3, 4, 5, 6, 7, and 8 will have the same total correct score as another participant who accurately reports all trials at lengths 3, 4, and 5, but who fails twice at list length 6. Second, because different participants receive different numbers of trials, the variance of the total correct metric is high relative to its mean and is highly skewed. This inflates standard deviations, as reflected in a high coefficient of variation (COV, the standard deviation divided by the mean), which is reportedly 23.4% for FS and 36.4% for BS (Wilde, Strauss, & Tulsky, 2004). The high variance reduces sensitivity to clinical abnormalities. For example, the average Wechsler total correct score for patients with mild Alzheimer's disease shows mean *z* score differences of only –0.22 for FS and –0.44 for BS (Wilde et al., 2004), indicating that the typical patient with mild Alzheimer's disease is poorly discriminated from age-matched control participants. In addition, the total correct score typically combines performance scores for FS and BS. Historically, these scores were combined to reduce the relative impact of digit span testing on overall IQ measures (Ramsay & Reynolds, 1995). However, subsequent studies have found different FS and BS abnormalities in various clinical populations (Carlesimo, Fadda, Lorusso, & Caltagirone, 1994; Kramer et al., 2003).

The Wechsler maximal span, the longest digit sequence accurately reported in WMS–III testing, is also often included in clinical reports. However, the Wechsler maximal span is also problematic for two reasons. First, it measures DS performance to a precision of only 1 digit: nearly as large as the standard deviation of the measure (Ardila, 2007). Second, it ignores performance variability. Thus, a participant who misses 1 digit on lists of 5, 6, and 7 digits and then misses both lists at length 8 will have the same Wechsler maximal span as a participant who performs flawlessly on lists 5, 6, and 7 digits, but misses both lists at length 8.

Insofar as Wechsler maximal span reflects an underlying continuously varying probabilistic function, some participants may have maximal spans that are best characterized by noninteger values. For example, if repeated testing were possible, a participant with a true maximal span of 7.5 digits would be expected to produce observed spans of 7 digits and 8 digits on equal numbers of WMS–III test sessions. In order to detect such intermediate spans in a single test, list lengths bounding the maximal span must be sampled on multiple trials.

The current study examined digit span performance using the computerized delivery of randomized digit lists that adaptively increased and decreased to repeatedly sample the lower and upper bounds of DS. During the test, it was also possible to characterize the longest list correctly reported before two errors occurred at the same list length (as in WMS–III maximal span) and the consistency of performance prior to making two successive errors at the same list length (as in WMS–III total correct). However, testing continued after two errors had been made at the same list length until the full set of trials had been delivered. This made it possible to evaluate new metrics that more fully captured the statistical properties of DS performance. In Experiment 1, 30 young volunteers participated in three separate test sessions in order to compare the variability and test–retest reliability of different metrics. In Experiment 2, DS and other cognitive tests were performed by 763 adults (age range 18–65 years) in a broad community sample obtained as part of an epidemiological study of environmental influences on cognition. This made it possible to evaluate the influences of demographic variables (e.g., age, educational level, etc.) on DS performance and to evaluate the correlations between different metrics of DS performance and the results of other tests of memory function.

EXPERIMENT 1

Method

Participants

A total of 31 participants took part in Experiment 1 after giving written informed consent, following institutional review board (IRB) regulations of the Veterans Affairs Northern California Health Care System (VANCHCS). The participants included 16 men and 15 women between the ages of 18 and 46 years (mean age = 26 years) with an average of 14.8 years of education. One female participant, who used a mnemonic strategy and had a BS of 13 digits, was excluded from the analysis.

Apparatus and stimuli

Forward and backward digit span testing was performed midway through the 90-minute California Cognitive Assessment Battery (CCAB), a set of 15 computerized tests and three adaptive questionnaires.[1]

In order to evaluate test–retest reliability, each participant underwent three complete CCAB test sessions at intervals ranging from 5 to 11 days.

Testing was performed in a quiet testing room using a standard PC controlled by Presentation software (Versions 13 and 14, Neurobehavioral Systems, Albany, CA). The PC was equipped with two monitors, one visible to the participant and one visible to the experimenter. Responses were recorded by the experimenter using a PC-gaming keyboard and mouse. First, the FS testing procedure was explained orally to the participant. Then, spoken digits (1–9) that had been digitally recorded (44.1 kHz, 16 bits) and normalized in mean sound intensity (70 dB SPL) were delivered binaurally through headphones at the rate of 1/s. Digits were randomly sampled without replacement up to list lengths of 9 digits (with single digit duplications when participants' spans surpassed 9) with the additional constraints that successive digits could not occur in regular ascending or descending sequences with equal consecutive step sizes (e.g., 123, 876, 357, 864, or 369).

A warning cue followed the final digit at an interval of 1.0 s, cueing participants to repeat the digit string. The digit sequence was displayed on the examiner's monitor during list presentation, and responses were transcribed by the examiner using the computer keyboard. The experiment logfile included the identity and timing of each digit presented, as well as the identity and timing of each response as transcribed by the examiner.

Participants received 14 trials with list lengths adaptively adjusted to reflect participant performance. Forward testing began at lists of three digits with list length increasing following a 1:2 staircase—that is, a single correct response increased the length of the subsequent list by one digit, while two incorrect responses were needed to reduce the list length by one digit. Following FS testing, the participant received 14 trials of BS testing with the digit sequence reported in backward order. BS testing began at lists of two digits.

Digit span scoring

The data from individual trials were analyzed using four different automated scoring metrics. Two measures estimated DS following procedures similar to those of the WMS–III. The two-error maximum length (TE-ML) measure recorded the maximal list length successfully recalled prior to missing two successive lists of the same length. Since digit lists were delivered using a 1:2 staircase, the TE-ML reflected the total number of trials correct prior to two successive misses. To evaluate response consistency prior to achieving the TE-ML, we quantified two-error total trials (TE-TT), the total number of trials (both correct and incorrect) presented prior to two successive errors at the same list length. The TE-TT measure, like the total trial correct measure obtained in the WMS–III, reflects the consistency of performance prior to achieving the TE-ML.

Two metrics were also evaluated that utilized the responses from all 14 trials: the maximum length (ML), the longest list correctly reported on any of the 14 trials, and mean span (MS), the list length where 50% of lists would be correctly reported based on an estimation using psychophysical procedures (Killion, Niquette, Gudmundsen, Revit, & Banerjee, 2004). The MS baseline was set at 0.5 digits less than the initial list length (i.e., 2.5 digits in FS) and was incremented by the fraction of digit strings accurately reported at each succeeding list length.

Table 1 illustrates the testing procedure using data from a single participant. Trial length (column 2) increased with each correct report before the participant's first miss on length 7 (Trial 5). Thereafter, trial lengths varied between 6 and 9 digits. The participant's TE-ML was 6 (reached on Trial 4, prior to two successive misses at lists of length 7), and the TE-TT was 4 (reflecting flawless responding prior to the TE-ML). The participant went on to achieve a ML of 8 (Trial 11). The MS was 7.08 as calculated by adding the hit rate for each list length (e.g., 3=1.0, 4=1.0, 5=1.0, 6=1.0, 7=0.25, 8=0.33, and 9=0.0, sum=4.58) to the baseline value of 2.5.

TABLE 1
Digit span trial scoring

Trial	*Length*	*Presented*	*Response*	*Outcome*
1	3	8 5 3	8 5 3	1
2	4	1 8 4 5	1 8 4 5	1
3	5	7 9 6 8 2	7 9 6 8 2	1
4	6	7 9 4 2 5 3	7 9 4 2 5 3	1
5	7	1 7 3 9 2 4 5	1 7 ***9 3*** 2 4 5	0
6	7	4 7 1 3 8 2 9	4 7 1 3 ***2 8*** 9	0
7	6	9 2 1 7 3 6	9 2 1 7 3 6	1
8	7	3 9 4 8 1 5 6	3 9 ***1*** 4 8 *X X*	0
9	7	8 9 7 2 5 6 4	8 9 7 2 5 6 4	1
10	8	2 7 8 1 3 6 5 9	2 7 ***1 3 8*** 6 5 9	0
11	8	4 9 6 7 5 2 8 3	4 9 6 7 5 2 8 3	1
12	9	2 3 1 7 9 4 6 5 8	2 ***7 3 1*** 9 ***6 4 1*** 8	0
13	9	4 1 5 7 8 6 9 2 3	4 1 5 7 ***6 9*** 2 3	0
14	8	8 5 4 9 6 2 3 1	8 5 4 9 6 ***3 2 1***	0

Note. Test results from a single forward span test for one participant. A total of 14 trials were presented, with list length (column 2) increasing after each correct trial and decreasing after two successive incorrect trials at the same list length. The lists presented are shown in column 3 and the response in column 4. The correctness of the response is shown in column 5 (1=correct, 0=incorrect). Errors are shown in bold italics.

[1] The CCAB includes the following computerized tests and questionnaires: finger tapping, simple reaction time, symbol–digit, Stroop, digit span forward and backward, phonemic and semantic verbal fluency, card sorting, verbal list learning, spatial span, trail making, symmetry detection, design fluency, the Wechsler Test of Adult Reading (WTAR), visual feature conjunction, the Paced Auditory Serial Addition Task (PASAT), the Cognitive Failures Questionnaire (CFQ), the posttraumatic stress disorder (PTSD) symptoms checklist, and a traumatic brain injury (TBI) questionnaire.

Results

Comparing metrics of DS

Table 2 shows the mean and standard deviations of the four different performance metrics evaluated in Experiment 1, for FS and BS. For FS, mean TE-ML over the three days of testing averaged 7.36 digits, with a range of 5.7 to 9.0 digits in different individuals. TE-ML scores underestimated the maximal span obtained over all 14 digit list presentations (ML=7.87 digits) by 0.51 digits. TE-TT averaged 5.91 trials, indicating that participants missed an average of 0.55 trials prior to achieving their TE-ML. For FS testing, mean MS scores (7.41 digits) were similar to mean TE-ML scores.

In BS testing, TE-ML scores averaged 5.80 digits with a range of 4.0 to 10.33 digits for different participants. Participants achieved an average ML of 6.48 digits, 0.68 digits longer than the TE-ML. TE-TT averaged 5.49 trials, indicating that participants missed an average of 0.69 trials prior to achieving their TE-ML. Mean MS scores were significantly higher than mean TE-ML scores (6.00 vs. 5.80), $t(89)$, $p < .04$. This likely reflected the fact that some participants were confused by the BS test procedure early in the test sequence and hence produced short TE-ML scores because of two successive misses at a relatively short list length. Standard deviations and COV were reduced for FS in comparison with BS for all measures.

We evaluated the reliability of TE-ML scores across the three days of testing. For FS, TE-ML scores within participants varied by 0–3 digits across testing days with an average absolute span difference between days of 0.96 digits. For backward span, spans varied by 0–3 digits across testing days with an average absolute span difference of 0.93 digits.

Table 3 shows the test–retest correlation matrix for different metrics of FS and BS across the three days of testing. Identity correlations reflect the accuracy with which a particular measure obtained on one day of testing predicts its score on a different day, while correlations across metrics reflect how well the scores on one metric predict the scores of other metrics obtained on different testing days. Across all measures, all correlations were significantly higher for BS than FS testing (12 of 12 comparisons, $p < .001$, sign test). This suggests that the ranking of participants across trials was more consistent in BS than in FS testing.

For FS, test–retest correlations were high for repeated ML (.68) and MS (.67) metrics. Correlations were also high between ML and MS metrics (.67), suggesting that both metrics were capturing similar aspects of DS performance. Test–retest reliability was considerably lower for the two-error measures TE-ML (.39) and particularly TE-TT (.12). The low test–retest reliability of the TE-TT metric suggests that the small number of trials missed prior to achieving the TE-ML was highly variable across test sessions. Indeed, the ML and MS measures obtained on one day of testing were better predictors of TE-ML and TE-TT scores on other testing days than were the TE-ML and TE-TT metrics themselves.

A similar pattern of test–retest reliability was seen for BS testing. Test–retest correlations were high for repeated ML (.81) and MS (.84) metrics, and correlations were also high between ML and MS metrics (.83). Test–retest reliability was considerably lower for TE-ML (.67) and TE-TT (.53) metrics. Again, ML and MS

TABLE 2
Means, standard deviations, and coefficients of variation for different digit span metrics obtained from forward and backward span of Experiment 1

		DS metrics		
		M	*SD*	*COV (%)*
FS	TE-ML	7.36	0.88	11.96
	TE-TT	5.91	0.97	16.41
	ML	7.87	0.97	12.33
	MS	7.41	0.94	12.69
BS	TE-ML	5.80	1.42	24.48
	TE-TT	5.49	1.81	32.97
	ML	6.48	1.36	20.99
	MS	6.00	1.32	22.00

Note. *M*=mean; *SD*=standard deviation; COV=coefficient of variation; DS=digit span; FS=forward span; BS=backward span. TE-ML=two-error maximum length; TE-TT=two-error total trials; ML=maximum length over all 14 trials; MS=mean span over 14 trials. The results have been averaged over three test sessions.

TABLE 3
Mean pairwise correlations among different measures of FS and BS testing across three test sessions

	Test–retest correlations of DS metrics							
	FS				*BS*			
	TE-ML	*TE-TT*	*ML*	*MS*	*TE-ML*	*TE-TT*	*ML*	*MS*
TE-ML	.39	.26	.51	.52	.67	.59	.76	.75
TE-TT		.12	.40	.38		.53	.68	.64
ML			.68	.67			.81	.83
MS				.67				.84

Note. DS=digit span; FS=forward span; BS=backward span. TE-ML=two-error maximum length; TE-TT=two-error total trials; ML=maximum length over all trials; MS=mean span over all trials.

measures obtained on one day of testing were better predictors of TE-ML and TE-TT scores on other testing days than were TE-ML and TE-TT metrics themselves.

Table 4 shows mean FS–BS difference scores for each metric along with associated standard deviations and COVs. FS exceeded BS by more than one digit for ML, TE-ML, and MS metrics. Standard deviations and COVs were larger for TE-ML and particularly TE-TT metrics than for ML or MS. We also compared the incidence of extreme FS–BS difference scores for the two metrics (ML and TE-ML) with single digit precision and for the subdigit precision MS metric. A total of 53% of test sessions showed FS–BS differences ≥ 2 digits or ≤ –2 digits with the TE-ML metric, and 37% of the test sessions produced similarly extreme differences with the ML metric. In contrast, only 3% of test sessions showed such extreme FS–BS differences with the MS metric.

The mean correlations between FS and BS are shown in Table 5. Despite the fact that these measures were obtained on the same day of testing, all correlations were substantially less than the correlations obtained for repeated FS or BS testing across days (Table 3). The correlation matrix reveals that FS scores are somewhat better predictors of BS scores than vice versa. FS ML (r = .46) and MS (r = .48) metrics were better predictors of BS performance than TE-ML and TE-TT were on BS (respectively, .21 and .10). As in the previous test–retest comparisons, ML and MS metrics in FS were better predictors of TE-ML and TE-TT performance in BS than were the TE-ML and TE-TT metrics themselves.

TABLE 4
Mean differences between FS and BS, standard deviations, and coefficients of variation for different metrics

	M	*SD*	*COV (%)*
TE-ML	1.56	1.38	88.40
TE-TT	0.42	1.75	416.70
ML	1.39	1.12	80.50
MS	1.41	1.07	75.90

Note. M = mean; *SD* = standard deviation; COV = coefficient of variation; FS = forward span; BS = backward span. TE-ML = two-error maximum length; TE-TT = two-error total trials; ML = maximum length over all trials; MS = mean span over all trials.

TABLE 5
Mean correlations between FS and BS for the different metrics, averaged over the three testing sessions

	BS			
FS	*TE-ML*	*TE-TT*	*ML*	*MS*
TE-ML	.21	.14	.32	.34
TE-TT	.15	.10	.27	.28
ML	.42	.39	.46	.50
MS	.41	.37	.45	.48

Note. FS=forward span; BS=backward span. TE-ML=two-error maximum length; TE-TT=two-error total trials; ML=maximum length over all trials; MS=mean span over all trials.

Learning effects

Participants showed small learning effects over three testing days on all metrics, averaging 0.30 digits from Day 1 to Day 3 for FS and 0.45 digits for BS on the MS metric. An analysis of variance (ANOVA) of MS scores with participants, day, and test type (FS or BS) as factors showed a significant effect of day, $F(2, 58) = 7.93$, $p < .001$. There was also a highly significant effect of test type, reflecting the fact that forward spans exceeded backward spans, $F(1, 29) = 52.40$, $p < .0001$. However, there was no significant Day × Test Type interaction, $F(2, 58) = 0.37$, *ns*, indicating that learning effects of similar magnitude were obtained during FS and BS testing.

Discussion

DS scores, measured using the two-error DS as in the WMS–III, appeared to vary continuously both in repeated tests of individual participants and across the participant population. Thus, each participant's TE-ML score sampled the continuously varying, probabilistic distribution of digit span capacity with single-digit measurement precision. Insofar as span capacity can be represented by a probability distribution reflecting the likelihood of correctly remembering digit lists of different lengths, continued sampling of DS performance would be expected to improve the accuracy of its estimation. It is therefore unsurprising that the ML metric, obtained over all 14 trials, was a more reliable predictor of future performance than the TE-ML metric that was obtained, on average, following 7.91 trials in FS and 7.49 trials in BS.

Comparison with previous digit span testing results

The TE-ML values obtained in the first session of Experiment 1 were somewhat longer than the digit spans of WMS–III normative data for young adults ages 20–29 (Wechsler, 1997b) for both FS (7.10 vs. 6.74 digits) and BS (5.53 vs. 5.07 digits) spans. Moreover, the standard deviations of the TE-ML measures were lower than those reported in WMS–III control data for both FS (0.88 vs. 1.31 digits) and BS (1.42 vs. 1.57 digits). Thus, the measurement precision of the TE-ML metric in the current experiment appeared to be slightly superior to that of the Wechsler DS metric in WMS–III normative data. This may have in part reflected an increase in the clarity and regularity of digit presentations as well as a reduction in variance due to randomized digit sampling. Alternatively, reduced variance may have reflected the use of a more homogenous population of participants, whose high average level of education (mean=14.8 years) may also help to explain the increases in mean span.

Learning effects

Participants' spans improved slightly but significantly across test sessions due to procedural learning. The short intervals between test sessions and the participants' knowledge that they would be repeating DS testing may have enhanced procedural learning effects in the current study. However, repeated testing with the Wechsler Adult Intelligence Scale (WAIS) or WMS–III might be expected to produce somewhat greater improvement, because these tests involve the repeated presentation of identical digit lists and hence might produce additional improvements due to content learning. In any case, the results suggest that learning effects should be taken into consideration when interpreting the results of repeated DS testing of individual participants.

Correlations between FS and BS

Correlational analysis showed relatively high correlations across FS and BS tests performed on separate days in comparison with the correlations between FS and BS tests performed on the same day. This confirms previous suggestions that FS and BS measure partially distinct cognitive operations (Ramsay & Reynolds, 1995) and is consistent with reports of their different clinical sensitivity profiles (Carlesimo et al., 1994; Kramer et al., 2003). As suggested by Lezak (1995), this implies that the standard WMS procedure of combining the total correct scores of FS and BS may reduce clinical sensitivity to neurocognitive deficits, particularly those that selectively impair BS.

Metrics for quantifying digit span performance

Our results indicate that the precision of digit span assessment was significantly improved when performance was quantified with the ML and MS metrics. This superiority was reflected in lower coefficients of variation and increased test–retest reliability. There were two essential differences between MS and ML metrics: (a) The MS but not the ML metric reflected average digit span performance and hence would be expected to be relatively insensitive to the total number of lists presented. (b) The MS metric quantified DS with subdigit precision while the ML metric quantified DS with single digit precision. The finer grained distribution of MS scores permits the selection of performance criteria with controlled false-positive rates (e.g., 5%). In contrast, single-digit precision does not permit the selection of criterion performance levels with desired false-positive rates because quantification is restricted to cardinal digit values. In addition, the subdigit precision of the MS metric reduced measurement rounding errors in FS–BS difference scores.

EXPERIMENT 2

Experiment 2 was performed to examine digit span performance in a larger population of participants varying in age and educational background. It also permitted the examination of correlations between different DS metrics and performance on other widely used neuropsychological tests.

Method

In Experiment 2, a 10-list test was administered to 763 community volunteers in Rotorua, New Zealand, who were participating in a study investigating the effects of hydrogen sulfide exposure on health. FS and BS testing was performed midway through a brief 30-min computerized assessment battery that included six tests from the CCAB. The methods were similar to those used in Experiment 1 with the following exceptions: (a) Only 10 digit lists were used, to reduce the time required for digit span testing; (b) FS testing began at 5 digits, and BS testing began at 4 digits; and (c) scoring was modified so that the examiner could check an "all correct" box to indicate the correct report of the entire digit string. Data were also gathered on four paper-and-pencil memory tests for a subset of the participants (N = 749). The tests included the Hopkins Verbal Learning Test (Revised (HVLT-R; Shapiro, Benedict, Schretlen, & Brandt, 1999), the Benton Visual Retention Test (BVRT; Sivan, 1992), the Digit Symbol test (Joy, Kaplan, & Fein, 2004), and the National Adult Reading Test (NART; O'Caroll & Gilleard, 1986). Because a large number of demographic and scoring correlations were examined, a strict criterion (p < .01) was used for evaluating statistical significance.

Participants

Participants ranged in age from 18 to 65 yrs (mean age = 46.5 years) with an average of 12.5 years of education. Participant recruitment was designed to obtain a random selection of Rotorua residents of these ages, stratified according to high, medium, or low levels of presumptive hydrogen sulfide exposure. Participants had to have resided in the city for 3 or more years. The only exclusions were inability to speak and write English, disability that would prevent visiting the study facility, and blindness. All participants signed written consent forms approved by the Northern Y Regional Ethics Committee in Rotorua and by the IRB for the University of California (UC) Davis/VANCHCS Clinical and Translational Science Center (CTSC).

Results

Metrics of DS performance

Table 6 shows means and variance measures for the different DS metrics for FS and BS in Experiment 2. The pattern of results was similar to that seen in Experiment 1. FS TE-ML scores averaged 6.35 digits, slightly less than MS scores (6.52) and nearly one half a digit less than ML scores (6.77). Standard deviations were somewhat increased relative to Experiment 1, reflecting the less homogenous population and the reduction in measurement precision due to shorter test duration. BS TE-ML scores were decreased by 1.74 digits with respect to FS scores. On average, the ML over the 10 trials was 0.68 digits longer than the TE-ML during BS testing.

Figure 1 (upper) shows the population distributions of scores for the different measures of FS. TE-ML, MS,

TABLE 6
Means, standard deviations, and coefficients of variation for FS and BS measures in Experiment 2

		M	*SD*	*COV (%)*
FS	TE-ML	6.35	1.15	18.05
	TE-TT	2.94	1.48	50.27
	ML	6.77	1.03	15.27
	MS	6.52	1.00	15.40
BS	TE-ML	4.61	1.22	26.56
	TE-TT	2.69	1.38	51.24
	ML	5.19	1.09	20.96
	MS	4.91	1.06	21.49

Note. M = mean; *SD* = standard deviation; COV = coefficient of variation; FS = forward span; BS = backward span. TE-ML = two-error maximum length; TE-TT = two-error total trials; ML = maximum length over all trials; MS = mean span over all trials.

and ML were positively skewed. TE-ML showed a median span of 6 with 80% of participants showing spans between 5 and 7 digits. The lower tail of the TE-ML distribution was broad with 20.2% of participants showing TE-MLs of 5 and 2.9% of participants showing TE-MLs of 4. The median MS span was 6.45 with 92% of participants producing MS scores between 5 and 7 digits. Also, there was less spread into the lower tail of the MS than the TE-ML metric, with only 5.0% showing MS scores of 5.08 or below. ML measures had a median of 7 digits, with more than 86% of participants showing maximal spans between 6 and 8 digits.

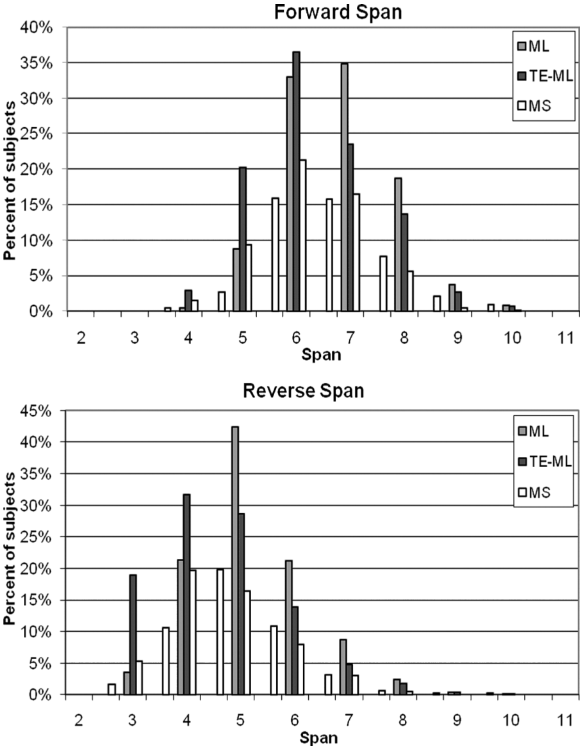

Figure 1. Population distributions of FS (upper) and BS (lower) scores for TE-ML, ML, and MS for participants in Experiment 2. For MS the percentages of scores within each 0.5-digit interval are shown. FS=forward span; BS=backward span. TE-ML=two-error maximum length; ML=maximum length over all trials; MS=mean span over all trials.

Although average MS and TE-ML scores were similar in the population as a whole, there were significant differences in individual participants. Overall, 13.7% of participants missed two trials relatively early in the test sequence but went on to achieve MS scores that were at least 0.9 digits beyond their TE-ML, and 4.3% of participants had TE-ML scores that were at least 0.9 digits longer than their MS scores. For participants with abnormal TE-ML scores (i.e., 4 or less), 64% also had abnormally low MS spans (< 5.08). However, of the 5% of participants who showed significant abnormalities for the MS measure, only 37% had abnormal TE-ML scores.

Figure 1 (lower) shows the distribution of BS scores, where TE-ML spans fell between 4 and 6 for 74% of participants. The lower tail of the TE-ML distribution was discontinuous, in part because testing began with strings of length 3. Overall, 18.9% of participants had TE-ML scores of 3, but no participants produced scores below 3. MS scores fell between 3.5 and 6.5 in 85.3% of participants with a median of 4.83. In the lower tail, 3.8% of participants produced BS MS scores below 3.25 with a minimum of 2.5. In BS testing, 19.1% of participants missed two trials relatively early in the test sequence, but went on to achieve MS that exceeded their TE-ML by at least 0.9 digits, and 4.3% of participants had TE-ML scores that exceed MS scores by at least 0.9 digits. Further investigation showed that of the participants with low BS TE-MLs only 20% had abnormal MS scores. In contrast, of the participants with low MS scores, 72% also had low TE-MLs. ML measures had a median of 5 digits, with 75% of participants showing BS ML scores between 5 and 7 digits. An analysis of the distributions of the different metrics showed that skewness was reduced for ML in comparison with the other metrics.

Correlations between different measures of FS and BS are shown in Table 7. Again, the highest correlations were found for ML (*r* = .50) and MS (*r* = .56), whereas

TABLE 7
Correlations between different metrics of FS and BS in Experiment 2

	BS			
FS	*TE-ML*	*TE-TT*	*ML*	*MS*
TE-ML	.43	.25	.48	.50
TE-TT	.32	.21	.35	.37
ML	.46	.27	.50	.54
MS	.49	.28	.53	.56

Note. FS=forward span; BS=backward span. TE-ML=two-error maximum length; TE-TT=two-error total trials; ML=maximum length over all trials; MS=mean span over all trials.

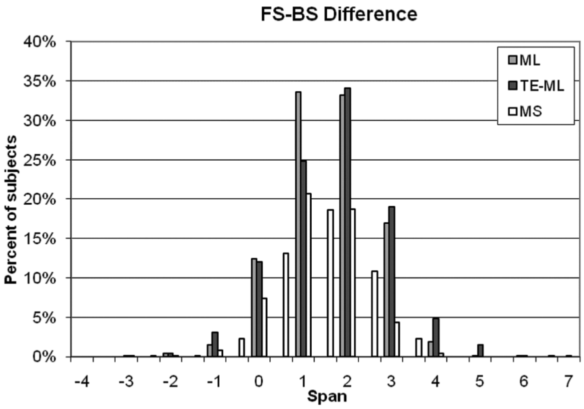

Figure 2. The population distribution of FS–BS difference scores for TE-ML, ML, and MS metrics in Experiment 2. FS=forward span; BS=backward span. TE-ML=two-error maximum length; ML=maximum length over all trials; MS=mean span over all trials.

TABLE 8
Correlations between DS metrics, age, and education level

		Age	*Ed*
FS	TE-ML	–.07	.11
	TE-TT	–.02	.05
	ML	–.03	.13
	MS	–.06	.15
BS	TE-ML	–.14	.15
	TE-TT	–.04	.07
	ML	–.14	.20
	MS	–.13	.19

Note. Ed=education. DS=digit span; FS=forward span; BS=backward span. TE-ML=two-error maximum length; TE-TT=two-error total trials; ML=maximum length over all trials; MS=mean span over all trials.

lower correlations were found for TE-ML (*r* = .43) and particularly for TE-TT (*r* = .21). Statistical analysis of z-transformed correlations revealed significant differences (*p* < .05) between ML and TE-TT and between MS and both TE-ML and TE-TT.

Figure 2 shows the distribution of FS–BS difference scores. All distributions are roughly normal, but the MS distribution showed less variance, with 82% of participants showing FS–BS differences in the range of 0.5–3.0 digits and difference scores exceeding 3.20 digits observed in 5.0% of the control population. In contrast, broader upper tails were observed for the ML and TE-ML distributions. For TE-ML difference scores, 6.6% of FS–BS differences equaled or exceeded 4 digits, and 25.6% equaled or exceeded 3 digits. For ML difference scores, 2.1% of difference scores equaled or exceeded 4 digits, and 19% equaled or exceeded 3 digits.

Correlations with age and education

The correlations between DS metrics, age, and education levels are shown in Table 8. None of the FS metrics correlated significantly with age. However, BS measures of ML, TE-ML, and MS showed small but significant negative correlations with age (*r* = –.13 to –.14), *t*(672) = 3.65 to 3.95, *p* < .0001. All FS and BS metrics, except TE-TT, correlated positively with years of education (*r* = .11 to .20), *t*(672) ranged from 3.07 to 5.75, *p* < .0001, with larger correlations uniformly observed for BS than for FS measures.

Correlations with the results of other neuropsychological tests

Correlations between digit span metrics for FS and BS and selected measures from the HVLT-R, the Digit Symbol Test, the NART, and the BVRT are shown in Table 9. With the exception of TE-TT, all DS metrics showed significant positive correlations with measures of verbal recall (HVLT-R total recall, HVLT-R delayed recall, BVRT correct, and Digit Symbol performance) and significant negative correlations with errors (BVRT errors and NART). BS was a better predictor of performance on other neuropsychological measures than was FS: Excluding TE-TT, all 18 correlations were greater for BS than FS (Sign test, *p* < .00001).

Correlations with neuropsychological test results varied slightly with the different DS metrics with the MS metric correlating more highly than either the TE-ML or TE-TT metrics (Sign test, *p* < .0005) as well as the ML metric (9 of 12 comparisons, Sign test, *p* < .05). The ML and TE-ML metrics did not differ significantly from each other. However, all 12 correlations with other neuropsychological tests results were higher for both metrics than with TE-TT (Sign test, *p* < .0005).

Discussion

Comparisons with previous normative digit span test results

TE-ML scores in Experiment 2 were similar to those reported for the age-matched WMS–III control data sample (Wechsler, 1997b) for both FS (6.4 vs. 6.6 digits) and BS (4.6 vs. 4.9 digits). However, the variance of the TE-ML metric in the current study was reduced in comparison with the variance reported in the WMS–III control data for both FS (1.15 vs. 1.31 digits) and BS (1.22 vs. 1.57 digits). Normative data from the WMS–III (Wechsler, 1997b) reported an average FS–BS difference score of 1.6 in participants aged 18–20 years, which increased to 1.8 in participants aged 55–65 years. In the current study, the TE-ML difference scores averaged 1.7 digits, similar to that predicted from the WMS–III normative data on the basis of the mean age of our participants. However, the standard deviation of the TE-ML FS–BS difference (1.3) was slightly lower than the variance reported for age-matched WMS–III normative data (1.4). Because of the 1:2 staircase procedure used in the current study, the TE-ML measures were obtained

TABLE 9
Correlations between DS metrics and other neuropsychological tests for FS and BS

			TE-ML	*TE-TT*	*ML*	*MS*
FS	Hopkins Verbal Learning Test-Revised	Total recall	.20	.12	.20	.22
		Delayed	.13	.07	.12	.14
	Digit Symbol		.22	.12	.22	.23
	National Adult Reading Test	Errors	–.25	–.18	–.25	–.28
	Boston Visual Retention Test	Correct	.16	.10	.16	.18
		Errors	–.24	–.16	–.25	–.28
BS	Hopkins Verbal Learning Test-Revised	Total recall	.30	.09	.32	.33
		Delayed	.27	.09	.3	.29
	Digit Symbol		.28	.11	.32	.32
	National Adult Reading Test	Errors	–.27	–.12	–.32	–.33
	Boston Visual Retention Test	Correct	.22	.10	.26	.26
		Errors	–.33	–.13	–.38	–.40

Note. DS = digit span; FS = forward span; BS = backward span. TE-ML = two-error maximum length; TE-TT = two-error total trials; ML = maximum length over all trials; MS = mean span over all trials.

from fewer trials than the corresponding measures obtained from the WMS–III. Thus, the reduced variance of the TE-ML metrics of FS, BS, and FS–BS would appear likely to reflect the increased clarity and regularity of digit sequence presentations and possibly the use of randomized digit lists.

Metrics of digit span performance

The MS and ML metric again appeared to offer a number of advantages in comparison with the TE-ML or TE-TT metrics: reduced variance, higher correlations between FS and BS, and higher correlations with the results of other neuropsychological tests. The tightened distribution of MS measures would increase its clinical sensitivity in comparison with TE-ML span measures. For example, the abnormality threshold of the MS metric for FS testing (5.08) was more than 1 digit greater than the abnormality threshold (4) of the TE-ML metric. Moreover, the finer grain of the MS distribution enabled the 5% false-positive rates for excessive FS–BS differences to be established with precision (3.08 digits), whereas the large TE-ML difference threshold of 4 produced a false-positive rate of 6.6%. As a result, the MS measure would be more sensitive to abnormal FS–BS differences in comparison with TE-ML measures.

Comparison of the results in Experiment 1 and Experiment 2

The results of Experiment 2 revealed that BS decreased with age and increased with education, whereas FS increased with educational level alone. The larger increase in BS than FS scores between Experiment 1 and Experiment 2 is consistent with the results of prior studies showing significant effects of education on both FS and BS (Gregoire & Van der Linden, 1997) and larger age-related differences in BS scores (Babcock & Salthouse, 1990; Hayslip & Kennelly, 1982; Hester, Kinsella, & Ong, 2004).

GENERAL DISCUSSION

Digit presentation effects on the precision of digit span assessment

The TE-ML metric showed lower variance in both experiments than did WMS–III normative data scored with a similar algorithm. One explanation is that the computerized digit delivery reduced variability in the rate and clarity of digit presentation associated with variable digit articulation within and across examiners. The use of constrained random digit sequences also enabled multiple tests at the same list length without concern for the repetition of particular digit sequences and may have improved the generalizability of results in regions where particular digit sequences in the WMS–III digit lists occur systematically (e.g., telephone area codes).

Improving the measurement of DS performance

The use of list lengths that sampled the upper and lower bound of digit span revealed that the TE-ML procedure of terminating testing after two errors underestimated the true ML span by more than 0.5 digits. The poorest metric was the TE-TT. The TE-TT, like the widely used WMS–III total correct score, reflects the consistency of performance at subthreshold list lengths. In comparison with the TE-ML, the TE-TT showed higher variance, a greater coefficient of variation, poorer test–retest reliability, lower correlations both with itself and with other measures of digit span performance across separate days of testing, and poorer correlations with scores on other neuropsychological tests of memory.

Sampling critical list lengths adaptively on additional trials improved the assessment of DS. In comparison with TE-ML and TE-TT metrics, MS and ML showed improved test–retest reliability, lower variance, and higher correlations between FS and BS, and were better predictors of performance on other memory tests. MS was preferable to the ML metric because of its insensitivity to the number of lists presented and its finer measurement grain that permitted the establishment of normal performance criteria with precise false-positive rates.

Comparisons with other computerized subtests of digit span

A number of computerized cognitive tests have been developed that assess memory for digits (Wild, Howieson, Webbe, Seelye, & Kaye, 2008). For example, the Cognitive Drug Research (CDR) test battery (Parrott, Garnham, Wesnes, & Pincock, 1996) and the Cognitive Drug Research computerized Assessment System (COG-DRAS; Simpson, Wesnes, & Christmas, 1989) assess memory for digits using a modified Sternberg task while the Automated Neuropsychology Assessment Metrics (ANAM; Kane & Kay, 1992) uses a digit set comparison task. However, these tasks measure recognition memory for digit strings rather than verbal working memory as assessed in traditional DS testing.

There are also three computerized test batteries that include verbal working memory assessment of digit span: the NeuroCog FX (Fliessbach, Hoppe, Schlegel, Elger, & Helmstaedter, 2006), IntegNeuro (Gordon, Cooper, Rennie, Hermens, & Williams, 2005), and MicroCog (Powell et al., 1993). These tests differ from the CCAB DS test in several important ways. First, the CCAB DS test is designed to enhance the efficiency of test administration by a trained examiner who administers the test. Other computerized tests are designed for unsupervised self-administration by the test participant. While self-administration enhances test efficiency, it can increase spurious variation in DS performance due to motivation, emotional lability, poor comprehension of test instructions, and lack of computer literacy (Wild et al., 2008). Second, the CCAB DS test uses calibrated auditory digit presentation while the other tests present digits visually. There are significant intermodality differences in digit span performance, particularly for BS (Powell & Hiatt, 1996; Ramsay & Reynolds, 1995). Moreover, auditory digit presentation is thought to more directly assess the core systems of verbal working memory (Baddeley, 2003). In addition, in the current experiments participants used verbal report, whereas existing computerized tests require participants to use manual responses with a keypad or touch screen. Manual responses are influenced by the participant's familiarity with computer response devices and by motor or executive control processes. Additionally, the visual search for digits on a keyboard or computer screen requires visuospatial attention and may interfere with the iconic representations of visually presented lists. Finally, although list lengths are adaptively adjusted in one of the computerized batteries, test scoring is based on metrics that are similar to the TE-ML or TE-TC. The current experiments demonstrate that these metrics are less accurate and reliable than MS and ML scoring procedures.

CONCLUSIONS

Examiner-administered computerized tests of DS offer a number of advantages over traditional paper-and-pencil tests. First, they improve the consistency and clarity of digit list delivery and permit the use of randomized digit lists. Second, the use of adaptive adjustment of digit list length using a 1:2 staircase shortens the time required to measure spans near the limits of average participants' capacities and therefore optimizes the sampling of digit span performance. The traditional approach of ending the test after two errors fails to adequately sample performance, and metrics based on this approach neglect useful information that is present in performance variability around maximal span. Acquiring more samples improves the reliability of span measures particularly when combined with a mean span (MS) metric that provides subdigit estimates of DS performance and generates a more continuous distribution of DS scores increasing sensitivity to FS–BS difference scores. Examiner-administered computer-controlled measures of digit span can significantly enhance the reliability and precision of digit span assessments of short-term verbal memory.

Original manuscript received 11 February 2010
Revised manuscript accepted 25 April 2010
First published online

REFERENCES

Ardila, A. (2007). Normal aging increases cognitive heterogeneity: Analysis of dispersion in WAIS–III scores across age. *Archives of Clinical Neuropsychology*, *22*(8), 1003–1011.

Babcock, R. L., & Salthouse, T. A. (1990). Effects of increased processing demands on age differences in working memory. *Psychology and Aging*, *5*, 421–428.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189–208.

Carlesimo, G. A., Fadda, L., Lorusso, S., & Caltagirone, C. (1994). Verbal and spatial memory spans in Alzheimer's and multi-infarct dementia. *Acta Neurologica Scandinavica*, *89*(2), 132–138.

Fliessbach, K., Hoppe, C., Schlegel, U., Elger, C. E., & Helmstaedter, C. (2006). NeuroCogFX—a computer-based neuropsychological assessment battery for the follow-up examination of neurological patients. *Fortschritte der Neurologie-Psychiatrie*, *74*(11), 643–650.

Gordon, E., Cooper, N., Rennie, C., Hermens, D., & Williams, L. M. (2005). Integrative neuroscience: The role of a standardized database. *Clinical EEG and Neuroscience*, *36*(2), 64–75.

Gregoire, J., & Van der Linden, M. (1997). Effect of age on forward and backward digit spans. *Aging, Neuropsychology, and Cognition*, *4*(2), 140–149.

Hayslip, B., & Kennelly, K. J. (1982). Short-term memory and crystallized-fluid intelligence in adulthood. *Research on Aging*, *4*, 314–332.

Hester, R. L., Kinsella, G. J., & Ong, B. (2004). Effect of age on forward and backward span tasks. *Journal of the International Neuropsychological Society*, *10*(4), 475–481.

Joy, S., Kaplan, E., & Fein, D. (2004). Speed and memory in the WAIS–III Digit Symbol–Coding subtest across the adult lifespan. *Archives of Clinical Neuropsychology*, *19*(6), 759–767.

Kane, R. L., & Kay, G. G. (1992). Computerized assessment in neuropsychology: A review of tests and test batteries. *Neuropsychology Review*, *3*, 1–117.

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, *116*(4, Pt. 1), 2395–2405.

Kramer, J. H., Jurik, J., Sha, S. J., Rankin, K. P., Rosen, H. J., Johnson, J. K., et al. (2003). Distinctive neuropsychological patterns in frontotemporal dementia, semantic dementia, and Alzheimer disease. *Cognitive and Behavioral Neurology*, *16*(4), 211–218.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

O'Caroll, R. E., & Gilleard, C. J. (1986). Estimation of premorbid intelligence in dementia. *British Journal of Clinical Psychology*, *25*(2), 157–158.

Parrott, A. C., Garnham, N. J., Wesnes, K., & Pincock, C. (1996). Cigarette smoking and abstinence: Comparative effects upon cognitive task performance and mood state over 24 hours. *Human Psychopharmacology*, *11*, 391–400.

Powell, D. H., & Hiatt, M. D. (1996). Auditory and visual recall of forward and backward digit spans. *Perceptual and Motor Skills*, *82*(3, Pt. 2), 1099–1103.

Powell, D. H., Kaplan, E. F., Whitla, D., Weintraub, S., Catlin, R., & Funkenstein, H. H. (1993). Microcog: Assessment of cognitive functioning (Version 2.1) [Computer software]. San Antonio, TX: The Psychological Corporation

Ramsay, M. C., & Reynolds, C. R. (1995). Separate digits tests: A brief history, a literature review, and a reexamination of the factor structure of the Test of Memory and Learning (TOMAL). *Neuropsychology Review*, *5*(3), 151–171.

Reeves, C., Schmauder, A. R., & Morris, R. K. (2000). Stress grouping improves performance on an immediate serial list recall task. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(6), 1638–1654.

Richardson, J. T. (2007). Measures of short-term memory: A historical review. *Cortex*, *43*(5), 635–650.

Shapiro, A. M., Benedict, R. H., Schretlen, D., & Brandt, J. (1999). Construct and concurrent validity of the Hopkins Verbal Learning Test–Revised. *Journal of Clinical Neuropsychology*, *13*(3), 348–358.

Silverman, M. J. (2007). The effect of paired pitch, rhythm, and speech on working memory as measured by sequential digit recall. *Journal of Music Therapy*, *44*(4), 415–427.

Simpson, P. M., Wesnes, K. A., & Christmas, L. (1989). A computerised system for the assessment of drug-induced performance changes in young, elderly and demented populations. *British Journal of Clinical Pharmacology*, *27*, 711–712.

Sivan, A. B. (1992). *Benton Visual Retention Test* (5th ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale—administration and scoring manual* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997b). *WMS–III administration and scoring manual*. San Antonio, TX: The Psychological Corporation.

Wild, K., Howieson, D., Webbe, F., Seelye, A., & Kaye, J. (2008). Status of computerized cognitive testing in aging: A systematic review. *Alzheimer's & Dementia*, *4*(6), 428–437.

Wilde, N. J., Strauss, E., & Tulsky, D. S. (2004). Memory span on the Wechsler scales. *Journal of Clinical and Experimental Neuropsychology*, *26*(4), 539–549.