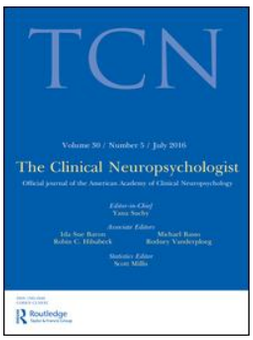

The Clinical Neuropsychologist

ISSN: 1385-4046 (Print) 1744-4144 (Online) Journal homepage: https://www.tandfonline.com/loi/ntcn20

Temporal Stability of the Dutch Version of the Wechsler Memory Scale—Fourth Edition (WMS-IV-NL)

Zita Bouman, Marc P. H. Hendriks, Albert P. Aldenkamp & Roy P. C. Kessels

To cite this article: Zita Bouman, Marc P. H. Hendriks, Albert P. Aldenkamp & Roy P. C. Kessels (2015) Temporal Stability of the Dutch Version of the Wechsler Memory Scale—Fourth Edition (WMS-IV-NL), The Clinical Neuropsychologist, 29:sup1, 30-46, DOI: 10.1080/13854046.2015.1137354

To link to this article: https://doi.org/10.1080/13854046.2015.1137354

Published online: 25 Feb 2016.

The Clinical Neuropsychologist, 2015
Vol. 29, No. S1, S30–S46, http://dx.doi.org/10.1080/13854046.2015.1137354

Routledge
Taylor & Francis Group

Temporal Stability of the Dutch Version of the Wechsler Memory Scale—Fourth Edition (WMS-IV-NL)

Zita Bouman[1,2], Marc P. H. Hendriks[1,2], Albert P. Aldenkamp[1,3,4,5], and Roy P. C. Kessels[2,6]

[1]Academic Centre for Epileptology, Kempenhaeghe, Heeze, the Netherlands
[2]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, the Netherlands
[3]Department of Neurology and School for Mental Health and Neuroscience, Maastricht University Medical Centre, Maastricht, the Netherlands
[4]Department of Neurology, University Hospital Gent, Gent, Belgium
[5]Faculty of Electrical Engineering, Signal Processing System Group, Technical University Eindhoven, Eindhoven, the Netherlands
[6]Department of Medical Psychology, Radboud University Medical Center, Nijmegen, the Netherlands

Objective: The Wechsler Memory Scale—Fourth Edition (WMS-IV) is one of the most widely used memory batteries. We examined the test–retest reliability, practice effects, and standardized regression-based (SRB) change norms for the Dutch version of the WMS-IV (WMS-IV-NL) after both short and long retest intervals. *Method*: The WMS-IV-NL was administered twice after either a short (M = 8.48 weeks, SD = 3.40 weeks, range = 3–16) or a long (M = 17.87 months, SD = 3.48, range = 12–24) retest interval in a sample of 234 healthy participants (M = 59.55 years, range = 16–90; 118 completed the Adult Battery; and 116 completed the Older Adult Battery). *Results*: The test–retest reliability estimates varied across indexes. They were adequate to good after a short retest interval (ranging from .74 to .86), with the exception of the Visual Working Memory Index (r = .59), yet generally lower after a long retest interval (ranging from .56 to .77). Practice effects were only observed after a short retest interval (overall group mean gains up to 11 points), whereas no significant change in performance was found after a long retest interval. Furthermore, practice effect-adjusted SRB change norms were calculated for all WMS-IV-NL index scores. *Conclusions*: Overall, this study shows that the test–retest reliability of the WMS-IV-NL varied across indexes. Practice effects were observed after a short retest interval, but no evidence was found for practice effects after a long retest interval from one to two years. Finally, the SRB change norms were provided for the WMS-IV-NL.

Keywords: Neuropsychological assessment; Episodic memory; Test battery; Test–retest reliability; Practice effects; Reliable change.

INTRODUCTION

In clinical practice, repeated neuropsychological assessments across time are often necessary to monitor patients with a variety of neurological and psychiatric disorders (Heilbronner et al., 2010; Lezak, Howieson, Bigler, & Tranel, 2012). For instance,

Address Correspondence to: Zita Bouman, MSc, Donders Institute for Brain Cognition and Behaviour, Radboud University Nijmegen, Montessorilaan 3, 6525 HR Nijmegen, the Netherlands. E-mail: z.bouman@donders.ru.nl

(*Received 29 June 2015; accepted 24 December 2015*)

repeated assessments are used to monitor the efficacy of cognitive rehabilitation, pharmalogical, or neurosurgical treatments (Chelune, 2003; Chelune, Naugle, Lüders, Sedlak, & Awad, 1993; Schoenberg et al., 2012). Also, serial assessments provide insight in the course of cognitive decline in patients with neurodegenerative disorders, such as dementia (Duff, Chelune, & Dennett, 2012). As memory problems are the most prevalent cognitive deficits in a variety of clinical pathologies, reliable repeated assessment of memory functioning plays a crucial role in neuropsychological evaluations. Therefore, there is a demand for evaluation of test–retest reliability and practice effects of memory tests.

A variety of tests and batteries exist that measure different aspects of memory functioning (see Lezak et al., 2012 for a comprehensive overview). The Wechsler Memory Scale (WMS) is one of the most widely used memory batteries worldwide (Rabin, Barr, & Burton, 2005). Several studies have examined test–retest reliability and practice effects of different versions of the WMS, such as the WMS (Dikmen, Heaton, Grant, & Temkin, 1999; McCaffrey, Ortega, & Haase, 1993; Mitrushina & Satz, 1991; Wechsler & Stone, 1945), the Wechsler Memory Scale—Revised (WMS-R: Ivnik, Smith, Malec, Petersen, & Tangalos, 1995; Theisen, Rapport, Axelrod, & Brines, 1998; Wechsler, 1987), and Wechsler Memory Scale—Third Edition (WMS-III: Iverson, 2001; Lo, Humphreys, Byrne, & Pachana, 2012; Wechsler, 1997). However, the test–retest reliability of the WMS-IV has been scarcely addressed so far (Wechsler, Holdnack & Drozdick, 2009; Holdnack, Drozdick, Weiss, & Iverson, 2013).

Previous research has shown that there are differences in test–retest reliability among different cognitive domains. The test–retest reliability of memory tests is generally found to be poorer than that of tests assessing other cognitive functions (Calamia, Markon, & Tranel, 2013; Ivnik et al., 1995; McCaffrey et al., 1993; McCaffrey & Westervelt, 1995), especially when the retest interval is longer (Domino & Domino, 2006). It has been suggested that normal human memory performance is variable and that caution needs to be exercised when interpreting repeated memory assessments (Dikmen et al., 1999), as poor reliability estimates may cause problems such as failing to detect actual changes in research or in clinical practice.

In addition to test–retest reliability from a pure psychometric perspective, practice effects may further complicate the interpretation of repeated memory assessment. Specifically, practice effects are improvements in the test performance on re-evaluations that do not reflect genuine improvement in the underlying construct, but may be related to other processes such as recollection of specific items, learned test-taking strategies, or familiarity with the test-occasion (Calamia, Markon, & Tranel, 2012). Notably, if all participants scores increase or decrease in the same amount, the reliability is still high. Therefore, it is possible that a test has a high reliability, but at the same time reveals large practice effects.

By definition, memory tests are especially susceptible to practice effects, because repeated assessment will enhance retrieval of specific items (Lezak et al., 2012; McCaffrey et al., 1993). However, findings on practice effects in the literature are difficult to compare, since practice effects are influenced by many different factors which vary across studies. For example, population-specific effects such as younger age and higher intellectual ability are related to larger practice effects (Dikmen et al., 1999; McCaffrey & Westervelt, 1995; Mitrushina & Satz, 1991; Rabbitt, Diggle, Smith, Holland, & Mc Innes, 2001; Rapport et al., 1997). Moreover, an increasing number of

re-administrations (Collie, Maruff, McStephen, & Darby, 2003; Ferrer, Salthouse, Stewart, & Schwartz, 2004; Theisen et al., 1998) and shorter lengths of the test–retest interval are associated with larger practice effects (Calamia et al., 2012).

Until now, only two studies have addressed retest effects of the latest edition of the WMS, the Wechsler Memory Scale—Fourth Edition (WMS-IV), which are reported in the test's Technical Manual (Holdnack & Drozdick, 2009) and in the Advanced Clinical Interpretation publication (Holdnack et al., 2013). These studies used the test–retest sample of the US WMS-IV, which consists of 244 participants (173 completed the Adult Battery and 71 the Older Adult Battery). The WMS-IV was administered twice after test-intervals from 2 to 12 weeks ($M = 23$ days). For the Adult Battery, the test–retest reliability coefficients for the index scores ranged between .81 and .83, and average increases ranged from 4.3 points (Visual Working Memory Index) to 13.7 points (Delayed Memory Index). For the Older Adult Battery, the test–retest reliability coefficients for the index scores ranged from .80 to .87, and average increases ranged from 10.6 points (Auditory Memory Index) to 12.4 points (Immediate Memory Index). Based on these results, considerable increments in the WMS-IV index scores seem to occur after short time intervals of several weeks in a healthy sample. Accordingly, Holdnack and colleagues (2013) provided regression equations for all WMS-IV indexes and subtests that can be used to predict reliable change in repeated assessments.

However, it is still unclear whether these performance increments continue to persist after longer retest intervals. It has been suggested that practice effects diminish when the time passes, but several studies have shown that practice effects may persist over longer time intervals up to 7–13 years (Basso, Carona, Lowery, & Axelrod, 2002; Salthouse, 2010; Salthouse, Schroeder, & Ferrer, 2004; Salthouse & Tucker-Drob, 2008). With respect to previous versions of the WMS, few studies have reported practice effects after long test–retest intervals in healthy participants (Chelune et al., 1993; Dikmen et al., 1999; Lo et al., 2012; Mitrushina & Satz, 1991). Overall, these studies found that the magnitude of practice effects after longer retest intervals varies per subtest (e.g., long lasting practice effects were commonly found on the subtests Logical Memory, and Verbal Paired Associates), and were also influenced by demographic variables such as age and intelligence (e.g., long-lasting practice effects were generally seen in younger adults and higher educated participants).

The present study examined the test–retest reliability and practice effects of the Dutch version of the WMS-IV (WMS-IV-NL: Hendriks, Bouman, Kessels, & Aldenkamp, 2014; Wechsler, 2009). Two different test–retest intervals were studies; one group of healthy participants was re-examined after a short retest interval (3–16 weeks) and another group after a long retest interval (12–24 months). It is expected that the test–retest reliability and practice effects after a short retest interval will be comparable to those found in the US WMS-IV test–retest sample (Holdnack & Drozdick, 2009; Holdnack et al., 2013). In addition, it is expected that the test–retest reliability may be somewhat lower after a long retest interval compared to a short retest interval. Moreover, we expect long-lasting learning effects on the verbal and visual episodic memory tasks (i.e., Auditory Memory Index, Visual Memory Index, Immediate Memory Index and Delayed Memory Index), but not on the visual working memory tasks (i.e., Visual Working Memory Index). Furthermore, we generated standardized regression-based (SRB) change norms to provide statistical directions for detecting significant changes at an individual level for use in clinical practice.

METHODS

Participants

The sample consisted of 234 healthy persons (age range 16–90 years, mean age = 59.55, *SD* = 21.36; 100 males) from the WMS-IV-NL standardization sample (Hendriks et al., 2014). Of these participants, 118 completed the WMS-IV-NL Adult Battery, and 116 completed the WMS-IV-NL Older Adult Battery. Participants from different age groups and with different educational levels were recruited by trained assessors through their network, via advertisement and via a database of Pearson Assessment, the Netherlands. The sample-selection was based on the Dutch population according to census results from the Central Office for Statistics of the Netherlands (CBS, 2011). The sample was stratified according to age, sex, education level, and ethnicity; and the participants were only included if they met the inclusion criteria: primary language is Dutch; no significant hearing of visual impairment; no psychiatric or neurologic disorder, no substance abuse affecting cognitive functioning; and no use of medicines affecting cognitive functioning.

Of these participants, 134 (49.3% Adult Battery and 50.7% Older Adult Battery) were reassessed after a short interval of approximately 8.48 weeks (*SD* = 3.40, range 3–16 weeks), and 100 (52% Adult Battery and 48% Older Adult Battery) were reassessed after a longer interval of approximately 17.87 months (*SD* = 3.48, range 12–24 months). With respect to the frequency distribution, neither the short nor the long-interval data were skewed (skewness and kurtosis coefficients >−1 and <1). Participant characteristics are summarized in Table 1. The WMS-IV-NL standardization study was approved by the Institutional Review Board of Radboud University, Nijmegen and written informed consent was obtained.

Neuropsychological tests

All participants were administered the WMS-IV-NL. This memory battery is divided into an Adult Battery for use in participants aged 65–90, and an Older Adult

Table 1. Participant characteristics

	Adult Battery			Older Adult Battery		
	Short retest interval	Long retest interval	*p*	Short retest interval	Long retest interval	*p*
N	66	52		68	48	
Age (Mean, *SD*)	41.73 (17.15)	43.94 (16.21)	ns	77.40 (7.86)	75.67 (6.32)	ns
Sex (M/F)	31/35	20/32	ns	27/41	22/26	ns
Education level (Low/average/high)	24/24/18	11/25/16	ns	36/19/13	21/16/10	ns
NART IQ (Mean, *SD*)	101.44 (12.90)	104.45 (13.68)	ns	102.28 (12.94)	104.37 (9.20)	ns

Notes: Education level was classified according to the Central Office for Statistics of the Netherlands (CBS, 2011), which is based on the International Standard Classification of Education (ISCED: United Nations Educational, Scientific and Cultural Organization Institute for Statistics (UNESCO-UIS, 2011). Intelligence was measures using the Dutch version of the National Adult Reading Test (NART: Nelson & Willison, 1991; Schmand et al., 1992).

Battery for use in participants aged 65–90. The WMS-IV-NL Adult Battery comprises one optional subtest, the Brief Cognitive Status Exam (BCSE), and a total of six primary subtests of which two visual working memory tests: Spatial Addition (SA) and Symbol Span (SSP), and four subtests with immediate and delayed recall conditions: Logical Memory I and II (LM), Verbal Paired Associates I and II (VPA), Visual Reproduction I and II (VR), and Designs I and II (DE). These six primary subtests contribute to five index scores: Auditory Memory (AMI), Visual Memory (VMI), Visual Working Memory (VWMI), Immediate Memory (IMI), and Delayed Memory (DMI). The WMS-IV-NL Older Adult Battery comprises the BCSE and a selection of four primary subtests (SSP, LM, VPA and VR) and four index scores (AMI, VMI, IMI, and DMI).

The Dutch version of the WMS-IV was developed to be equivalent to original published US version of this test battery. The nonverbal visual stimuli are identical, and the instructions, auditory stimuli and scoring criteria were translated and adapted to the Dutch language. Pilot studies (first pilot study $n = 60$; second pilot study $n = 120$) were performed to check and improve the Dutch language adaptation of the WMS-IV. Moreover, an expert group consisting of clinical neuropsychologists from the Netherlands and Belgium checked the Dutch adaptation after both pilot studies (see also Hendriks et al., 2014; for a detailed description of the development of the WMS-IV-NL). Moreover, a previous study revealed that the factor structures of the Dutch and US WMS-IV standardization samples were invariant, which strengthens the case of equivalence (Bouman, Hendriks, Kerkmeer, Kessels, & Aldenkamp, 2015).

Also, the Dutch version of the National Adult Reading Test (NART) was used as an estimate of intelligence (IQ) (Nelson & Willison, 1991; Schmand, Lindeboom, & Van Harskamp, 1992)

Procedures

Administration of the WMS-IV-NL was performed in accordance with the test manual (Hendriks et al., 2014; Wechsler, 2009). The trained assessors were neuropsychologists or research assistants who completed an interactive training about the WMS-IV-NL administration and scoring. Moreover, their performance was monitored and evaluated before and at multiple times during the study. The standardization study of the WMS-IV-NL was accomplished by 93 independent trained assessors, the short-term retest assessment was performed by 31 assessors, and the long-term retest assessment was performed by 9 assessors. All participants in the short-term group were tested by the same assessor twice. However, due to logistic constraints, only 21 participants in the long-term retest group were tested by the same assessor twice.[1]

[1]To examine whether the performance increments differed when the WMS-IV-NL was administered twice by the same examiner or by different examiners, we conducted a 2 × 2 × 5 mixed factor multivariate analysis of covariance was performed with Examiner (2 levels: same vs. different) as the between-subject factor, Time (2 levels: baseline vs. re-evaluation) and Index Scores (5 levels: AMI, VMI, VWMI, IMI, and DMI) as the repeated factors, and Interval (short vs. long) as the covariate. In both the Adult and Older Adult Batteries, no significant main effect was found for Examiner, nor were there significant interaction effects for Time × Examiner, Index Scores × Examiner or Time × Index Scores × Examiner ($p > .05$). This indicates that there is no difference in performance increments when the WMS-IV-NL is administered twice by the same examiner or by two different examiners.

Statistical analyses

Demographic variables were compared for the groups who were re-evaluated after a short or a long retest interval using analyses of variance (age and NART IQ) and chi-squared tests (sex and education level). For all analyses we used the scaled scores to enhance the comparability to the test–retest studies reported for the US WMS-IV (Holdnack & Drozdick, 2009; Holdnack et al., 2013). Also, these age-corrected scores are more insightful for clinicians. All analyses were performed using the Statistical Package for Social Sciences (SPSS) version 19.0, and all effects are reported as significant at $p < .05$.

Test–retest reliability of the WMS-IV-NL was assessed by Pearson correlation coefficients between the scores from the baseline and re-evaluations (i.e., short- and long-term intervals). Moreover, the reliability coefficients of the short- and long-term retest groups were compared. Also, the reliability coefficients of the short-term retest group were compared to the test–retest reliabilities reported in the US WMS-IV-NL Technical Manual (Holdnack & Drozdick, 2009) using Fisher *r*-to-*z* transformation.

To examine the practice effects, we conducted a number of steps. First, we compared the baseline measures of the short- and long-term retest groups using a one-way between-groups multivariate analysis of variance (MANOVA) with Interval (2 levels: short vs. long) as between-subject factor and baseline measures of the index scores (5 levels: AMI, VMI, VWMI, IMI, and DMI) as dependent variables. Next, in order to study whether the short- and long-term retest groups showed different performance increments across time, we conducted a mixed between-within subjects MANOVA with Interval (2 levels: short vs. long) as the between-subject factor, Time (2 levels: baseline vs. re-evaluation) as the within-subject factor, and Index Scores (5 levels: AMI, VMI, VWMI, IMI, and DMI) as dependent variables. Subsequently, we performed paired samples *t*-tests for the separate groups (i.e., short- and long-term retest) with Time (2 levels: baseline vs. re-evaluation) as within-subject factor and each of the five WMS-IV-NL index scores as dependent variable (Corrected alpha from .01 to reduce the risk of Type I errors).

In addition, we used a multivariate SRB approach to determine reliable change on the WMS-IV-NL index scores after short- and long-term retest intervals. According to the procedure described by McSweeny, Naugle, Chelune, and Lüders (1993), multiple regression analyses were employed to derive equations for predicting WMS-IV-NL index scores at re-evaluation from baseline test performance and other predictors. Specifically, hierarchical regression models were performed with the retest WMS-IV-NL index score as dependent variable, and baseline scores, test–retest interval (days), and demographic variables (age, education level, and sex) as predictors. Education level (low, average, and high) was classified according to the Central Office for Statistics of the Netherlands (CBS, 2011), which is based on the International Standard Classification of Education (ISCED: United Nations Educational, Scientific and Cultural Organization Institute for Statistics (UNESCO-UIS, 2011). Only predictor variables that were significant at the .05 level were retained in the model. Next, the intercept and regression coefficients form these models were used to estimate the predicted retest index scores for all participants, and the 90 and 95% confidence intervals were applied to all individual's to determine base rates of significant improvements, declines, and stability on the WMS-IV-NL scores (see also a procedure utilized by Temkin, Heaton, Grant, & Dikmen, 1999).

RESULTS

The groups did not significantly differ with respect to age, sex distribution, education level, and NART IQ score ($p < .05$) (see Table 1).

Test–retest reliability

The mean WMS-IV-NL index scores across time for both the short- and long-term retest groups are presented in Table 2. Additionally, the mean WMS-IV-NL subtest scores across time for the short- and long-term retest groups can be found in Supplementary Table 1. The correlations between the scores from the baseline and re-evaluations were all significant ($p < .001$). Specifically, for the short retest interval, we found adequate to good correlations for the WMS-IV-NL index scores (ranging from $r = .74$ for VMI Adult Battery to $r = .86$ for AMI Older Adult Battery), with the exception of the VWMI in the Adult Battery ($r = .59$). With the exception of the VWMI in the Adult Battery, all reliability coefficients are comparable to the test–retest reliabilities reported in the US WMS-IV-NL Technical Manual ($p < .05$) (Holdnack & Drozdick, 2009). For the long retest interval, we found somewhat lower (but not statistically different) correlations on most of the index scores than after a short retest interval (ranging from $r = .56$ for VWMI Adult Battery to .77 for VMI Adult Battery) (see Table 2).

Practice effect

Adult Battery. Table 2 shows the mean WMS-IV-NL Adult Battery index scores across time for both groups (i.e., short- and long-term intervals). The groups (Interval: short vs. long) did not differ significantly at the baseline measures.

The mixed factor MANOVA revealed a significant main effect for Interval ($F(1, 112) = 9.13$, $p < .003$, $\eta_p^2 = .08$), with participants who were re-evaluated after the short retest interval performing significantly higher than participants who were re-evaluated after the long retest interval ($p < .001$). In addition, a main effect of Time ($F(1, 112) = 48.48$, $p < .001$, $\eta_p^2 = .30$) was observed. Contrast analyses revealed that overall, participants performed better at the re-evaluation than at the baseline measure ($p < .001$). No significant main effect was found for Index Scores ($F(2.092, 234.345) = 2.42$, $\eta_p^2 = .02$).

Additionally, the interaction effect of Interval × Index Scores was not significant, $F(2.092, 268.207) = 3.15$, $\eta_p^2 = .03$. Interval and Time showed significant interactions, $F(1, 112) = 35.74$, $p < .001$, $\eta_p^2 = .24$. That is, participants who were re-evaluated after the short retest interval performed better the second time ($p < .001$), whereas the performance of participants who were re-evaluated after the long retest interval did not differ between the two assessments. Also, the interaction between Time and Index Scores was significant, $F(2.395, 268.207) = 3.15$, $p < .036$, $\eta_p^2 = .03$. The three-way interaction Interval × Time × Index Scores was not significant, $F(2.395, 268.207) = 2.22$, $p = .101$, $\eta_p^2 = .02$. Subsequently, Table 2 shows that with the exception of the VWMI, performance on all index scores increased significantly from baseline to re-evaluation after a short interval. Specifically, the overall group showed a mean increase of approximately 10, 9, 4, 11, and 10 points on AMI, VMI, VWMI, IMI, and DMI, respectively. In contrast, the performance on all WMS-IV-NL index scores did not differ between the

Table 2. Mean WMS-IV-NL index scores across time for both groups (i.e., short- and long-term intervals)

	Short retest interval						Long retest interval						Diff baseline measures			Diff *r**	
	Baseline	Re-evaluation	*r*	*t*	*p*	η_p^2	Baseline	Re-evaluation	*r*	*t*	*p*	η_p^2	*F*	*p*	η_p^2	*z*	*p*
Adult Battery	(*n* = 66)						(*n* = 52)										
AMI	101.38 (15.05)	111.73 (14.66)	.81	−8.84	<.001	.55	102.44 (15.49)	103.68 (13.59)	.66	−.85	.399	.01	.14	.707	.00	1.75	.080
VMI	103.11 (13.60)	111.67 (13.63)	.74	−6.94	<.001	.43	98.62 (15.56)	98.72 (15.67)	.77	−.24	.812	.00	2.79	.097	.02	−.37	.711
VWMI	103.03 (13.65)	107.52 (13.68)	.59	−2.84	.006	.11	98.90 (14.59)	99.90 (13.85)	.56	−.37	.713	.00	2.50	.116	.02	.24	.810
IMI	101.65 (13.70)	112.94 (13.76)	.76	−9.27	<.001	.58	100.02 (15.08)	101.00 (13.89)	.73	−.75	.456	.01	.38	.540	.00	.35	.726
DMI	103.56 (13.88)	113.94 (14.13)	.79	−8.94	<.001	.56	101.23 (16.23)	101.92 (15.06)	.69	−.43	.672	.00	.71	.403	.01	1.17	.242
Older Adult Battery	(*n* = 68)						(*n* = 48)										
AMI	98.24 (16.22)	104.10 (15.89)	.86	−5.68	<.001	.33	104.31 (14.61)	101.96 (16.77)	.61	1.36	.180	.04	4.28	.041	.04	3.01	.003
VMI	99.56 (13.11)	108.19 (13.16)	.72	−7.22	<.001	.44	100.81 (16.77)	103.75 (15.15)	.58	−1.04	.304	.02	.20	.653	.00	1.26	.208
IMI	98.56 (15.78)	106.09 (16.05)	.80	−6.21	<.001	.37	105.82 (14.78)	103.21 (14.44)	.70	1.15	.256	.03	4.43	.038	.04	1.19	.234
DMI	98.32 (14.21)	106.74 (14.67)	.82	−7.94	<.001	.49	102.79 (15.43)	102.46 (16.61)	.68	.32	.7484	.00	2.59	.110	.02	1.69	.091

*Differences between test–retest reliability coefficients of the short retest and the long retest interval groups calculated using Fisherman's *z* transformation.

baseline and re-evaluation assessment after a long interval (mean overall group difference between −1 and 1 points). Table 3 shows the percentages of participants gaining 5, 10, 15, 20, 25, and 30 points per Index Score.

Older Adult Battery. Table 2 shows the mean WMS-IV-NL Older Adult Battery index scores across time for both groups (i.e., short- and long-term intervals). A significant main effect of Interval (short vs. long) was found for the baseline measures, $F(4, 111) = 3.24$, $p < .015$, $\eta_p^2 = .10$, with the long-term retest group performing significantly higher at the baseline measures of AMI, IMI, and DMI.

The mixed factor MANOVA revealed a significant main effect for Time ($F(1, 110) = 17.69$, $p < .001$, $\eta_p^2 = .14$). Contrasts revealed that participants performed better at the re-evaluation than at the baseline measure ($p < .001$). No significant main effects were found for either Interval, $F(1, 110) = .08$, $\eta_p^2 < .001$, nor Index Scores, $F(1.368, 150.533) = .43$, $\eta_p^2 < .004$.

Additionally, the interaction effect of Interval × Index Scores was not significant $F(1.368, 150.533) = 1.55$, $\eta_p^2 = .01$. Interval and Time showed significant interactions, $F(1, 110) = 27.73$, $p < .001$, $\eta_p^2 = .18$. That is, participants who were re-evaluated after the short retest interval performed better the second time ($p < .001$), whereas the performance of participants who were re-evaluated after the long retest interval did not differ between the two assessments. Also, the interaction between Time and Index Scores was significant, $F(1.662, 182.782) = 5.48$, $p < .008$, $\eta_p^2 = .05$. The three-way interaction Interval × Time × Index Scores was not significant, $F(1.662, 182.782) = 1.21$, $\eta_p^2 = .01$. Furthermore, Table 2 shows that the performance on all index scores increased significantly from baseline to re-evaluation after a short interval. Specifically, the participants showed a mean overall group increase of approximately 6, 9, 8, and 8 points on AMI, VMI, IMI, and DMI, respectively. In contrast, the performance on all index scores did not differ between the baseline and re-evaluation assessment after a long interval (overall group mean differences between −4 and 2 points). Table 3 shows the percentages of participants gaining 5, 10, 15, 20, 25, and 30 points per Index Score.

Table 3. Percentage of participants with change scores (Time 2–Time 1) of different magnitudes for all WMS-IV-NL index scores after short- and long-term intervals for both groups

	Short retest interval						Long retest interval					
	≥5	≥10	≥15	≥20	≥25	≥30	≥5	≥10	≥15	≥20	≥25	≥30
Adult Battery	(n = 66)						(n = 52)					
AMI (%)	70.8	50.8	33.8	13.8	7.7	1.5	27.5	23.5	21.6	11.8	2.0	0
VMI (%)	62.5	46.9	31.36	14.1	3.1	3.1	43.1	23.5	7.8	0	0	0
VW MI (%)	46.2	32.3	16.9	10.8	4.6	3.1	42.3	17.3	9.6	7.7	3.8	1.9
IMI (%)	76.6	60.9	3.4	12.5	7.8	3.1	35.3	23.5	5.9	2.0	0	0
DMI (%)	72.3	40.0	27.7	18.5	10.8	4.6	33.3	25.5	19.6	2.0	0	0
Older Adult Battery	(n = 68)					(n = 48)						
AMI (%)	57.4	32.4	13.2	5.9	1.5	1.5	22.7	15.9	4.5	4.5	2.3	2.3
VMI (%)	73.5	50.0	22.1	7.4	4.4	1.5	37.5	29.2	22.9	12.5	8.3	4.2
IMI (%)	61.8	41.2	25.0	13.2	5.9	1.5	27.3	11.4	2.3	2.3	0	0
DMI (%)	64.7	42.6	25.0	14.7	2.9	1.5	31.8	20.5	13.6	4.5	2.3	2.3

Regression-based measures for change

Results of the regression analyses for predicting retest scores on the WMS-IV-NL index scores are provided in Table 4. In addition, regression-based results for the WMS-IV-NL subtest scores after a short- and a long-term retest interval can be found in Supplementary Tables 2 and 3. Baseline performance was a significant predictor of the retest score for all WMS-IV-NL index scores in both batteries after both short-term and long-term retest intervals. The other predictors were only entered in a selection of equations (see Table 4): age accounted for <5% of the statistical variance in the two equations; and sex accounted for <5.2% of the statistical variance in the equation for VWMI in the Adult Battery after a long retest interval. Moreover, the variables test–retest interval and education level did not meet the criteria for inclusion as predictor in any of the regression equations. In order to calculate the predicted retest score, clinicians may use the regression equations provided in Table 4.

Next, the following equation was used to calculate standardized z-scores: z-score = $(Y_o - Y_p)/SE_{est}$, where Y_o is the observed retest score, Y_p is the predicted retest score, and SE_{est} is the standard error of the estimate from the regression analysis. These z-scores provide individual level determination of change, and z-scores exceeding ±1.64 are significant with an 90% confidence interval and those exceeding ±1.96 are significant with a 95% confidence interval. The base rates of these confidence intervals for the WMS-IV-NL index scores are presented in Table 5.

DISCUSSION

The present study provides test–retest reliability estimates, examined practice effects and presented SRB change norms for the WMS-IV-NL Adult and Older Adult Batteries using large independent samples of healthy participants who were re-assessed either after a short retest-interval (3–16 weeks) or a long retest-interval (12–24 months). To our knowledge, this study is the first to examine the magnitude of test–retest reliability and practice effects for the WMS-IV over longer time intervals than three months.

Consistent with the results reported in the Technical Manual of the US WMS-IV (Holdnack & Drozdick, 2009), short-term test–retest reliability estimates for most of the WMS-IV-NL index scores were adequate to good, with the exception of a low test–retest reliability for the VWMI in the Adult Battery. Long-term test–retest reliability coefficients were generally lower than the short-term estimates, which is in agreement with the notion that test–retest reliability decreases as the retest interval is longer (Domino & Domino, 2006). With respect to previous versions of the WMSs, poor test–retest reliability estimates were reported after a longer retest interval for the WMS (Haltstead–Reitan Neuropsychological Test Battery/WMS: Dikmen et al., 1999) and WMS-III (Lo et al., 2012). Furthermore, similar findings were demonstrated for the Auditory Verbal Learning Test (Geffen, Butterworth, & Geffen, 1994; Uchiyama et al., 1995) and Hopkins Verbal Learning Test (Rasmusson, Bylsma, & Brandt, 1995).

As expected, practice effects for the WMS-IV-NL index scores were most prominent after a short retest interval. Upon re-examination after such short time periods, healthy participants seem to benefit greatly from the first administration. In agreement with previous work (Basso et al., 2002; Holdnack & Drozdick, 2009; Lo et al., 2012; Wechsler, 1997), no significant practice effects were observed for the working memory

Table 4. Regression coefficients and indices of significance

Short interval	R	SE_{est}	Constant	$\beta_{baseline}$	$\beta_{testinterval}$	β_{age}	$\beta_{education}$	β_{sex}	Equation predicted score
Adult Battery (n = 66)									
AMI	.80	8.81	33.138	.774	–	–	–	–	33.138 + (baseline AMI × .774)
VMI	.73	9.39	36.284	.731	–	–	–	–	36.284 + (baseline VMI × .731)
VWMI	.57	11.35	47.982	.575	–	–	–	–	47.982 + (baseline VWMI × .575)
IMI	.74	9.31	37.172	.745	–	–	–	–	37.172 + (baseline IMI × .745)
DMI	.80	8.52	41.684	.769	–	−.177	–	–	41.684 + (baseline DMI × .769) + (age × −.177)
Older Adult Battery (n = 68)									
AMI	.88	7.78	46.042	.863	–	−.345	–	–	46.042 + (baseline AMI × .863) + (age × −.345)
VMI	.72	9.23	36.482	.720	–	–	–	–	36.482 + (baseline VMI × .720)
IMI	.80	9.65	25.652	.816	–	–	–	–	25.652 + (baseline IMI × .816)
DMI	.83	8.27	23.751	.844	–	–	–	–	23.751 + (baseline DMI × .844)
Long interval									
Adult Battery (n = 52)									
AMI	.66	11.04	40.493	.625	–	–	–	–	40.493 + (baseline AMI × .625)
VMI	.77	10.29	21.201	.781	–	–	–	–	21.201 + (baseline VMI × .781)
VWMI	.64	10.77	31.065	.588	–	–	–	6.412	31.065 + (baseline VWMI × .588) + (sex × 6.12)
IMI	.77	9.41	26.989	.734	–	–	–	–	26.989 + (baseline IMI × .734)
DMI	.71	11.20	32.171	.696	–	–	–	–	32.171 + (baseline DMI × .696)
Older Adult Battery (n = 48)									
AMI	.63	13.16	26.354	.725	–	–	–	–	26.354 + (baseline AMI × .725)
VMI	.60	12.98	45.435	.571	–	–	–	–	45.435 + (baseline VMI × .571)
IMI	.70	10.48	33.984	.661	–	–	–	–	33.984 + (baseline IMI × .661)
DMI	.69	12.14	25.901	.745	–	–	–	–	25.901 + (baseline DMI × .745)

Notes: SE_{est} = standard error of the estimate; β = unstandardized beta (slope); testinterval was measured in days; age was measured in years; education level was coded 1: low, 2: average, 3: high; sex was coded 1: male, 2: female.

All index equations use age adjusted standard scores.

Table 5. Base rates of 90 and 95% SRB change norms for the WMS-IV-NL index scores

	90% SRB confidence interval						95% SRB confidence interval					
	Short retest interval			Long retest interval			Short retest interval			Long retest interval		
	Declined	Stable	Improved	Declined	Stable	Improved	Declined	Stable	Improved	Declined	Stable	Improved
Adult Battery												
AMI (%)	3.1	93.8	3.1	3.9	90.2	5.9	3.1	95.4	1.5	2.0	94.1	3.9
VMI (%)	3.1	92.2	4.7	3.9	96.1	0	1.6	95.3	3.1	2.0	98.0	0
VWMI (%)	1.5	90.8	7.7	7.7	88.5	3.8	1.5	96.9	1.5	3.8	96.2	0
IMI (%)	4.7	92.2	3.1	3.9	90.2	5.9	0	100	0	4.0	96.0	0
DMI (%)	0	92.3	7.7	3.9	94.1	2.0	0	98.5	1.5	3.9	96.1	0
Older Adult Battery												
AMI (%)	2.9	91.2	5.9	2.3	90.9	6.8	0	95.6	4.4	2.3	93.2	4.5
VMI (%)	8.8	86.8	4.4	6.3	89.6	4.2	2.9	95.6	1.5	4.2	93.8	2.1
IMI (%)	4.4	91.2	4.4	4.5	93.2	2.3	0	100	0	4.7	95.3	0
DMI (%)	2.9	92.6	4.4	6.8	90.9	2.3	0	95.5	1.5	4.5	93.2	2.3

Notes: Adult Battery short retest interval $n = 66$; Adult Battery long retest interval $n = 52$; Older Adult Battery short retest interval $n = 68$; Older Adult Battery long retest interval $n = 48$.

measure. Furthermore, we found that practice effects were more pronounced in the Adult Battery than in the Older Adult Battery (McCaffrey et al., 1993; McCaffrey & Westervelt, 1995; Mitrushina & Satz, 1991; Tulsky & Zhu, 1997). Presumably, healthy participants with optimal memory function are able to remember specific stimuli of the episodic memory subtests or effective test-taking strategies which may have resulted in performance increments. Furthermore, they may have benefitted from familiarity with the test procedure (Anastasi & Urbina, 1997; Calamia et al., 2012).

Our findings extend those of previous research (Holdnack & Drozdick, 2009; Holdnack et al., 2013) in that we examined repeated assessments of the WMS-IV-NL Adult and Older Adult Batteries after a long retest interval, whereas previous studies only used re-assessments up to three months. We found no evidence for practice effects for the long-term retest group. That is, performance on all WMS-IV-NL index scores remained stable across the baseline and re-evaluation after a long interval from 12 to 24 months (overall group mean differences between −1 and 1). This suggests that the performance on the WMS-IV-NL after a long retest-interval of 12–24 months is influenced to a lesser extent by factors such as recollection of specific test items, test-taking strategies, or familiarity with the test-occasion. In the future, it would be interesting to evaluate whether such a long time interval may also diminish effects of practice in other neuropsychological measures.

For clinicians who are interested in determining whether a patient's change in performance on the WMS-IV-NL index scores is clinically meaningful and reliable, we have provided SRB equations in Table 4. These SRB equations correct for practice effects using an individual's baseline performance as a predictor of the retest WMS-IV-NL index score (Duff et al., 2012; Temkin et al., 1999). Moreover, it is possible to correct for other demographic variables that could potentially impact memory performance at re-evaluation, such as the test–retest interval, age, education level, and sex. One could argue that simple reliable change approaches that reveal cut-offs may be more easily used in clinical practice, but they do not correct for practice effects, regression to the mean and other potential predictors. Furthermore, another advantage of SRB change norms is that they are converted into *z*-scores, i.e., a common metric that allows us to make direct comparisons of changes in performance across different neuropsychological tests (Duff et al., 2012).

Also, we took into account other variables, such as test–retest interval and the demographic variables age, education level, and sex. These variables may affect repeated assessments using memory tests. Our results show that these variables did not have significant effects on most of the regression equations in our study. These findings are in agreement with several previous studies reporting that baseline measures alone predict a large amount of the variance and that subsequent variables (e.g., retest interval and demographics) predict none or only small amounts of the variance (Temkin et al., 1999).

One potential confound of the current study is that the short interval group was assessed by the same examiner twice, whereas the assessments in the long-term group were mainly performed by two different examiners. However, because there were no differences in gains when the WMS-IV-NL was administered twice by the same examiner or by two different examiners, it is unlikely that this has influenced the results. Furthermore, assignment of participants to either the short or long retest interval group was done pseudo-randomly; first, the short-interval substudy was performed by

asking participants in the normative sample whether they were willing to take part in this test–retest study (based on stratification criteria and availability). Subsequently, the long-interval substudy was performed in a similar manner (only excluding participants who already took part in the short-delay study).

A potential limitation of our study is that we found significant differences in the baseline measures of the short-term and long-term retest groups for the WMS-IV-NL Older Adult Battery. Because of the pseudo-random group assignment, we cannot think of a potential bias that could have caused these baseline differences. Note, however, that these baseline differences are taken into account statistically in the newly performed regression analyses. Another limitation of this study is the broad time window for the long-interval group. It would be a recommendation for further research to examine a retest interval as a continuous variable, making it possible to determine whether the gain in test performance reaches an asymptote and is no longer clinically meaningful.

A further limitation of the WMS-IV is that no alternate forms are available for this memory battery. Alternate forms of episodic memory tests may reduce the confounding practice effects as they intercept the recollection of specific items (Benedict, 2005; Benedict & Zgaljardic, 1998). Examples of memory tests with alternate forms are the California Verbal Learning Test—II (Delis, Kramer, Kaplan, & Ober, 2000) and the Rivermead Behavioral Memory Test—Third Edition (Wilson et al., 2008). Although it is well known that alternate forms cannot eliminate practice effects because other factors such as recollection of test demands, effective test-taking strategies, or familiarity with the test-procedure influence the effects of practice (Anastasi & Urbina, 1997; Beglinger et al., 2005; Calamia et al., 2012), it is suggested to use memory tests with alternative forms when a short retest interval is required.

In conclusion, the findings of this study show that the WMS-IV-NL has an adequate test–retest reliability. Since the authorized Dutch version of the WMS-IV and the original US version are highly equivalent, our results are likely to apply to the US version and other language versions of the WMS-IV as well. In line with this notion, the findings corroborate previously observed practice effects on the WMS-IV after a short time interval (Holdnack & Drozdick, 2009), but no evidence for practice effects were found after a long time interval of 12–24 months. Furthermore, practice effect-adjusted SRB change norms were provided for the WMS-IV-NL.

ACKNOWLEDGMENTS

We thank Pearson Assessment B. V., Amsterdam, The Netherlands, for authorizing and funding the development of the WMS-IV-NL. We thank Aileen Bremer, Bart Kral, Elaha Naimi, Jessyca Villier, Julius Ströhm, Karina Burger, Kimberly van der Wissel, Rens van Meegen, and Sophie Jellema for their assistance in the data collection.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

FUNDING

This study was funded by Pearson Assessment B.V. and Kempenhaeghe, Academic Centre for Epileptology, Heeze, The Netherlands.

REFERENCES

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York, NY: Macmillan.

Basso, M. R., Carona, F. D., Lowery, N., & Axelrod, B. N. (2002). Practice effects on the WAIS-III across 3- and 6-month intervals. *The Clinical Neuropsychologist (Neuropsychology, Development and Cognition: Section D), 16*, 57–63. doi:10.1076/clin.16.1.57.8329

Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J., … Siemers, E. R. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology, 20*, 517–529. doi:10.1016/j.acn.2004.12.003

Benedict, R. H. (2005). Effects of using same- versus alternate-form memory tests during short-interval repeated assessments in multiple sclerosis. *Journal of the International Neuropsychological Society, 11*, 727–736. doi:10.1017/S1355617705050782

Benedict, R. H., & Zgaljardic, D. J. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology (Neuropsychology, Development and Cognition: Section A), 20*, 339–352. doi:10.1076/jcen.20.3.339.822

Bouman, Z., Hendriks, M. P. H., Kerkmeer, M. C., Kessels, R. P. C., & Aldenkamp, A. P. (2015). Confirmatory Factor Analysis of the Dutch version of the Wechsler Memory Scale—Fourth Edition (WMS-IV-NL). *Archives of Clinical Neuropsychology*, *30*, 228–235. doi:10.1093/arclin/acv013.

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist, 26*, 543–570. doi:10.1080/13854046.2012.680913

Calamia, M., Markon, K., & Tranel, D. (2013). The robust reliability of neuropsychological measures: Meta-analyses of test–retest correlations. *The Clinical Neuropsychologist, 27*, 1077–1105. doi:10.1080/13854046.2013.809795

Centraal Bureau voor Statistiek [Central Office of Statistics for the Netherlands] (2011). *Dataleving Enquête Beroepsbevolking* [Data survey workforce]. Retrieved from http://www.cbs.nl

Chelune, G. J. (2003). Assessing reliable neuropsychological change. In R. Franklin (Ed.), *Prediction in forensic and neuropsychology: Sound statistical practices* (pp. 123–147). Mahwah: Lawrence Erlbaum.

Chelune, G. J., Naugle, R. I., Lüders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology, 7*, 41–52. doi:10.1037/0894-4105.7.1.41

Collie, A., Maruff, P., McStephen, M., & Darby, D. G. (2003). Psychometric issues associated with computerised neuropsychological assessment of concussed athletes. *British Journal of Sports Medicine, 37*, 556–559. doi:10.1136/bjsm.37.6.556

Delis, D. C., Kramer, H. H., Kaplan, E., & Ober, B. A. (2000). *California Verbal Learning Test* (2nd ed.). San Antonio, TX: Psychological Corporation.

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test–retest reliability and practice effects of expanded Halstead–Reitan Neuropsychological Test Battery. *Journal of the International Neuropsychological Society, 5*, 346–356.

Domino, G., & Domino, M. L. (2006). *Psychological testing*. Cambridge, UK: Cambridge University Press.

Duff, K., Chelune, G., & Dennett, K. (2012). Within-session practice effects in patients referred for suspected dementia. *Dementia and Geriatric Cognitive Disorders, 33*, 245–249. doi:10.1159/000339268

Ferrer, E., Salthouse, T. A., Stewart, W. F., & Schwartz, B. S. (2004). Modeling age and retest processes in longitudinal studies of cognitive abilities. *Psychology and Aging, 19*, 243–259. doi:10.1037/0882-7974.19.2.243

Geffen, G. M., Butterworth, P., & Geffen, L. B. (1994). Test–retest reliability of a new form of the auditory verbal learning test (AVLT). *Archives of Clinical Neuropsychology, 9*, 303–316. doi:10.1093/arclin/9.4.303

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., & Hart, R. P. (2010). Official position of the American academy of clinical neuropsychology on serial neuropsychological assessments: The utility and challenges of repeat test administrations in clinical and forensic contexts. *The Clinical Neuropsychologist, 24*, 1267–1278. doi:10.1080/13854046.2010.526785

Hendriks, M. P. H., Bouman, Z., Kessels, R. P. C., & Aldenkamp, A. P. (2014). *Wechsler Memory Scale-Fourth Edition, Dutch Edition (WMS-IV-NL)*. Amsterdam: Pearson Assessment.

Holdnack, J. A., Drozdick, L., Weiss, L. G., & Iverson, G. L. (Eds.). (2013). *WAIS-IV, WMS-IV, and ACS: Advanced clinical interpretation*. San Antonio, TX: Academic Press.

Holdnack, J. A., & Drozdick, L. W. (2009). *Wechsler memory scale* (Technical and interpretive manual 4th ed.). San Antonio, TX: NCS Pearson.

Iverson, G. L. (2001). Interpreting change on the WAIS-III/WMS-III in clinical samples. *Archives of Clinical Neuropsychology, 16*, 183–191. doi:10.1016/S0887-6177(00)00060-3

Ivnik, R. J., Smith, G. E., Malec, J. F., Petersen, C., & Tangalos, E. G. (1995). Long-term stability and intercorrelations of cognitive abilities in older persons. *Psychological Assessment, 7*, 155–161. doi:10.1037/1040-3590.7.2.155

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment*. New York, NY: Oxford University Press.

Lo, A. H., Humphreys, M., Byrne, G. J., & Pachana, N. A. (2012). Test–retest reliability and practice effects of the Wechsler Memory Scale-III. *Journal of Neuropsychology, 6*, 212–231. doi:10.1111/j.1748-6653.2011.02023.x

McCaffrey, R. J., Ortega, A., & Haase, R. F. (1993). Effects of repeated neuropsychological assessments. *Archives of Clinical Neuropsychology, 8*, 519–524. doi:10.1016/0887-6177(93)90052-3

McCaffrey, R. J., & Westervelt, H. J. (1995). Issues associated with repeated neuropsychological assessments. *Neuropsychology Review, 5*, 203–221. doi:10.1007/BF02214762

McSweeny, A. J., Naugle, R. I., Chelune, G. J., & Lüders, H. (1993). "T Scores for Change": An illustration of a regression approach to depicting change in clinical neuropsychology. *Clinical Neuropsychologist, 7*, 300–312. doi:10.1080/13854049308401901

Mitrushina, M., & Satz, P. (1991). Effect of repeated administration of a neuropsychological battery in the elderly. *Journal of Clinical Psychology, 47*, 790–801. doi:10.1002/1097-4679(199111)47:6<790:AID-JCLP2270470610>3.0.CO;2-C

Nelson, H. E., & Willison, J. R. (1991). *The Revised National Adult Reading Test: Test manual*. Windsor: NFER-Nelson.

Rabbitt, P., Diggle, P., Smith, D., Holland, F., & Mc Innes, L. (2001). Identifying and separating the effects of practice and of cognitive ageing during a large longitudinal study of elderly community residents. *Neuropsychologia, 39*, 532–543. doi:10.1016/S0028-3932(00)00099-3

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20*, 33–65. doi:10.1016/j.acn.2004.02.005

Rapport, L. J., Axelrod, B. N., Theisen, M. E., Brines, D. B., Kalechstein, A. D., & Ricker, J. H. (1997). Relationship of IQ to verbal learning and memory: Test and retest. *Journal of Clinical and Experimental Neuropsychology, 19*, 655–666. doi:10.1080/01688639708403751

Rasmusson, D. X., Bylsma, F. W., & Brandt, J. (1995). Stability of performance on the Hopkins Verbal Learning Test. *Archives of Clinical Neuropsychology, 10*, 21–26. doi:10.1093/arclin/10.1.21

Salthouse, T. A. (2010). Influence of age on practice effects in longitudinal neurocognitive change. *Neuropsychology, 24*, 563–572. doi:10.1037/a0019026

Salthouse, T. A., Schroeder, D. H., & Ferrer, E. (2004). Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. *Developmental Psychology, 40*, 813–822. doi:10.1037/0012-1649.40.5.813

Salthouse, T. A., & Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology, 22*, 800–811. doi:10.1037/a0013091

Schmand, B., Lindeboom, J., & Van Harskamp, F. (1992). *NLV: Nederlandse leestest voor volwassenen* [The Dutch Adult Reading Test: A measure of premorbid intelligence]. Lisse: Swets & Zeitlinger B.V.

Schoenberg, M. R., Rinehardt, E., Duff, K., Mattingly, M., Bharucha, K. J., & Scott, J. G. (2012). Assessing reliable change using the repeatable battery for the assessment of neuropsychological status (RBANS) for patients with Parkinson's disease undergoing deep brain stimulation (DBS) surgery. *The Clinical Neuropsychologist, 26*, 255–270. doi:10.1080/13854046.2011.653587

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society, 5*, 357–369.

Theisen, M. E., Rapport, L. J., Axelrod, B. N., & Brines, D. B. (1998). Effects of practice in repeated administrations of the Wechsler Memory Scale-Revised in normal adults. *Assessment, 5*, 85–92. doi:10.1177/107319119800500110

Tulsky, D., & Zhu, J. (1997). *WAIS-III/WMS-III technical manual*. San Antonio, TX: Psychological Corporation.

Uchiyama, C. L., D'Elia, L. F., Dellinger, A. M., Becker, J. T., Selnes, O. A., Wesch, J. E., … Miller, E. N. (1995). Alternate forms of the auditory-verbal learning test: Issues of test comparability, longitudinal reliability, and moderating demographic variables. *Archives of Clinical Neuropsychology, 10*, 133–145. doi:10.1093/arclin/10.2.133

United Nations Educational, Scientific and Cultural Organisation Institute for Statistics. (2011). *International Standard Classification of Education (ISCED)*. Montreal: Author.

Wechsler, D. (1987). *Wechsler Memory Scale—Revised (WMS-R)*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *Wechsler Memory Scale—Third Edition (WMS-III)*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2009). *Wechsler Memory Scale—Fourth Edition (WMS-IV)*. San Antonio, TX: Pearson Assessment.

Wechsler, D., & Stone, C. P. (1945). A Standardized memory Scale for clinical Use. *The Journal of Psychology, 19*, 87–95.

Wilson, B. A., Greenfield, E., Clare, L., Baddeley, A., Cockburn, J., Watson, P., … Crawford, J. R. (2008). *The Rivermead Behavioural Memory Test—Third Edition*. London: Pearson Assessment.