The Clinical Neuropsychologist, 2012, 26 (1), 102–115
http://www.psypress.com/tcn
ISSN: 1385-4046 print/1744-4144 online
http://dx.doi.org/10.1080/13854046.2011.640642

Psychology Press
Taylor & Francis Group

Overcoming the Challenge of Re-assessing Logical Memory

Ralf Schnabel
Department of Psychological Medicine, University of Auckland, Auckland, New Zealand

Practice effects present a challenge for neuropsychological re-assessments. Insufficiently controlled test-learning effects could result in "improved" test scores on re-assessment, which could wrongly be interpreted as recovery when in fact the underlying cognitive function has remained unchanged or deteriorated. Logical memory is highly sensitive to practice effects. Clients often remember the commonly used stimulus stories of the Wechsler Memory Scales (WMS) in subsequent re-assessments. Therefore alternative test stimuli are needed for research and clinical practice. This study undertook the development and statistical evaluation of a new set of logical memory stories, which can be utilised interchangeably with the traditional Wechsler stories. Empirical testing with different client groups ($n = 240$) confirmed that the newly created test stimuli have highly compatible structural and statistical properties to the WMS stories.

Keywords: Practice effect; Neuropsychological assessment; Logical memory; Wechsler Memory Scale.

INTRODUCTION

"Practice effects" or "test-learning effects" can impact on test scores and compromise the validity of a neuropsychological re-assessment (McCaffrey, Duff, & Westervelt, 2000; McCaffrey, Ortega, & Haase, 1993). Improved performance on cognitive testing may be mistaken as "significant recovery" or "positive response to treatment" when in fact the client remembers their previous assessment and the test stimuli involved, and performs quite well based on the training effect of a previous assessment (Goldberg, Keefe, Goldman, Robinson, & Harvey, 2010).

Various factors have been shown to mitigate the likelihood of practice effects, including the characteristics of the specific test (McCaffrey et al., 2000), the time span between initial and subsequent testing (Salthouse, Schroeder, & Ferrer, 2004), and the level of the clients' impairment (Lezak, Howieson, & Loring, 2004). The risk for incurring practice effects decreases when a longer time interval between test and re-test is observed, and when the degree of memory impairment is profound; nevertheless, statistically significant practice effects may still occur after 18 months even in cognitively challenged client groups (Heaton, Gladsjo, et al., 2001).

A growing body of literature has emerged analyzing the test/re-test problem from a statistical perspective and offering guidance in interpreting score changes on repeated testing. The Reliable Change Index (Jacobson & Truax, 1991) and the Reliability-Stability Index (Chelune, Naugle, Luders, Sedlak, & Awad, 1993) provide formulae by which a change in an individual's score on repeated testing can

Address correspondence to: Ralf Schnabel, Department of Psychological Medicine, University of Auckland, Private Bag 92019, Auckland, New Zealand. E-mail: rsch087@aucklanduni.ac.nz

Accepted for publication: November 8, 2011. First published online: December 16, 2011.

be judged as statistically significant. Increasingly complex procedures have since been developed to control for confounds such as practice effects and regression-to-mean by introducing constants for the expected practice effect and by using regression models. A synopsis of methods was provided by Collie, Darby, Falleti, Silbert, and Maruff (2002). Efforts to statistically determine and control practice effects were made by Martin et al. (2002) and Sawrie, Chelune, Naugle, and Luders (1996). The authors analyzed test/re-test data of intractable-epilepsy sufferers on the waiting list for receiving surgery ($n = 42$ and 51, respectively) and determined the raw score improvement on re-testing for different neuropsychological tests. ‘‘Corrected Reliable Change’’ indices were then calculated by factoring in the predicted test-learning effect for each sub-test. Restrictions in the generalizability of the constant apply due to high variability of practice effects for individual clients and sub-tests, fixed re-test intervals, and the use of a small and very specific clinical population. Heaton, Temkin, et al. (2001) investigated test/re-test reliabilities and practice effects of the Halstead-Reitan Battery (Dikmen, Heaton, Grant, & Temkin, 1999) and of older versions of the Wechsler Adult Intelligence Scale (Wechsler, 1955, 1981), based on different populations, including a large non-clinical participant group ($n = 384$), a group with schizophrenia ($n = 69$), and a group with recent brain trauma ($n = 33$). Despite similarities across samples in reliability coefficients and practice effects, norms for change did not generalize adequately from non-clinical to clinical groups. In comparing different methods of establishing change, the more complex regression-based prediction models did not prove to be superior compared to the practice-effect-corrected Reliable Change Index. Using a substantial sample of older clients ($n = 445$), Duff et al. (2005) presented data on test/re-test stability and practice effects of the Repeatable Battery for the Assessment of Neuropsychological Status (Randolph, 1998). No significant practice effects were found based on a re-test interval of over 1 year.

Suggestions have been made that practice effects could be minimized by presenting alternative test stimuli. Regrettably, many established tests do not offer either parallel versions of the tests or alternative stimuli for re-testing (Lezak et al., 2004). Examples of tests with parallel test stimuli include the Rey Auditory-Verbal Learning Test (Schmidt, 1996; Spreen & Strauss, 1998) and the California Learning Test (Delis, Kaplan, Kramer, & Ober, 2000; Spreen & Strauss, 1998); both are list-learning tests of random words or categorized words. Here alternate word lists are available (Lezak et al., 2004, pp. 422–434; Uchiyama et al., 1995). Less satisfactory are tests for which an alternative stimulus exists, in principle, but comprehensive norms only apply to the initial version, not the alternative stimulus. An example is the Rey Complex Figure Test (Meyers & Meyers, 1995), a test for visual comprehension and memory, for which an alternative but insufficiently co-validated stimulus is available (Taylor, 1969). In the absence of empirical data confirming equivalence of initial and alternative stimuli the use of alternative stimuli is restricted to qualitative screening (Hubley, 2010).

A particular challenge that has remained insufficiently resolved applies to logical memory, most commonly appraised by presenting a set of short stories that have to be repeated by the client immediately following narration by the health practitioner and also after a 30-minute delay. Commonly used test stimuli involve short, ‘‘catchy’’ storylines with clear themes and a number of emotive details, which bear considerable risks of being remembered in subsequent re-assessments. The popular Logical Memory Test of the Wechsler Memory Scale (WMS) comprises two stories, one of which has remained unchanged in the last three editions of the test battery. This spans a period of 23 years of regular updates to norms, although the easily learned test stimulus has been retained (Wechsler, 1987, 1997, 2009). Furthermore, the second stimulus story has remained essentially unchanged since the previous revision in 1997. The authors on the Technical and Interpretive Manual of the WMS-IV point out the substantial improvements in scores on repeated presentation of the WMS stimulus stories. After a time interval of 14 to 84 days (mean of 23 days) a representative adult sample of 173 examinees improved on second testing by 1.9 standard score points (0.63 *SD*) for immediate recall and 2.3 (0.77 *SD*) for delayed recall (Wechsler, 2009, pp. 50–51). Based on the continued use of the widely known WMS stories, most clinicians in neuropsychological practice can remember a number of clients asking: *''Are you going to read me that story again with the woman who got robbed?''*

Faced with the challenge of re-assessing logical memory in the light of likely practice effects, attempts have been made to develop alternative test stimuli. Morris, Kunka, and Rossini (1997) devised two stories that matched the WMS stories in the number of semantic units, emotive tone, and readability. In a sample of 50 undergraduate students, moderate levels of correlation were found between individual stories, ranging from .44 to .63. The usefulness of these alternative stories is limited due to colloquial idioms, such as ''quarterback'', and the injury-related content of both stories, such as ''it took seventeen stitches to close the wound''. Another problematic issue is the restriction of the study to students (mean age of 21.6 years) and the relatively small sample size, which do not allow conclusions about the value of these alternative stories in clinical populations.

A further attempt was made by Sullivan (2005), where six stories with matching lexical and linguistic characteristics were developed and colloquialisms were eliminated through external reviewers. All of the six new stories and the two WMS standard stories were presented to a sample of 32 undergraduate psychology students (mean age of 21 years). Subsequently, three pairs of new stories were created using the six stories by grouping together stories which, when combined as a pair, were most comparable to the pair of WMS stories. Sullivan reported similar means of remembered items for the three pairs of new stories (28.3, 28.5, and 28.9), and the published WMS-R mean (25.7) for the examined age group. As in the study by Morris et al. (1997), uncertainty remains regarding the compatibility of new and standard stories in clinical populations, given the use of a small, young, and academic test sample. Furthermore, Sullivan did not publish the actual alternate stories, reporting only on the procedure of stimulus creation and the statistical findings; hence these cannot be replicated in clinical settings.

The most recent attempt to provide alternate forms of logical memory was presented by Cunje, Molloy, Standish, and Lewis (2007). The authors developed three very short stories (''paragraphs''), unrelated to the WMS. Each of their stories comprised a short description of an animal's activity (''The red fox ran across the ploughed field'') followed by two sentences describing the natural environment surrounding the animal. Encouraging levels of consistency were documented for the three stories in different client groups, including, mild cognitive impairment

($n = 45$), dementia ($n = 55$), and controls ($n = 46$). This consistency was demonstrated both for immediate and delayed recall. The stories were not designed as alternate stimuli to the WMS stories, but served the specific research interest of the authors who were investigating cognition in different clinical populations. Accordingly no age norms are available for these new stories, and the results cannot be used to calculate compounded memory indices provided by the WMS.

The current study sought to address the need for the availability of empirically validated test stimuli that can be used as alternatives to the established WMS stories in clinical populations. This will be beneficial in addressing the concerns regarding uncontrolled practice effects both in clinical practice and research settings.

METHOD

Development of alternative test stimuli

Logical memory in the WMS-IV is assessed by presenting two test-stories, each story consisting of 25 separate semantic entities in the format of congruent, sequential, and emotionally engaging, but non-threatening storylines. The total items remembered from both stories are summed together, resulting in a raw score between 0 and 50. The items recalled are noted at two time-points: immediately after presentation (immediate recall) and with a 20- to 30-minute delay (delayed recall). We developed two alternative stories with similar formal characteristics, each containing 25 semantic items and involving a coherent, plausible, and moderately emotive narrative. A peer review was undertaken to confirm cultural appropriateness of the story lines and the vocabulary used, including absence of colloquialisms.

A small pilot study ($n = 60$) was undertaken to ensure ease of readability and comprehension of the stories. In addition, objective scoring criteria for each semantic unit were established. Inter-scorer reliability was assessed by audio-taping the responses of 20 clients for the standard and new stories and then presenting the recordings to two health professionals for independent scoring. An acceptable inter-scorer reliability ($r = .97$) was observed. A structural comparison between the WMS standard stories and newly developed stories demonstrated similar but non-identical linguistic characteristics (Table 1).

Table 1. Formal and linguistic parameters of WSM IV standard stories and new stories

	Standard stories			New stories		
Parameters	A-story	B-story	Combined stories	A-story	B-story	Combined stories
Words (*n*)	65	86	151	69	75	144
Characters (*n*)	278	388	666	317	361	678
Sentences (*n*)	3	5	8	5	6	11
Words per sentence (mean)	21.6	17.2	18.8	13.8	12.5	13
Characters per word (mean)	4.1	4.4	4.3	4.4	4.6	4.5
Passive sentences (%)	33	0	12	0	0	0
Flesch Reading Ease	74.2	63.4	68.3	73.8	6.1	67.2
Flesch-Kincaid Grade Level	8.2	8.6	8.4	6.3	7.8	7.1

Compared to the pair of standard stories, the pair of new stories was found to have slightly fewer words, shorter sentences, and was generally easier to understand according to the Flesch-Kincaid Grade Level and the percentage of passive sentences. The Flesch Reading Ease scores were largely identical for standard and new stories. Given the similarities noted in the number of items recalled between standard and new stories in the pilot study, no further alterations were made. The final versions of the two new stories and their scoring criteria are presented in Table 2.

Participants

A total of 160 clinical clients were recruited from a community-based Psychological Assessment and Rehabilitation Centre in Auckland, New Zealand, over a period of 18 months. This consisted of 80 clients with mild Traumatic Brain Injury (mTBI) and 80 clients with Major Depressive Episode (MDE). A control group of 80 healthy volunteers was recruited informally.

Clients were included in the mTBI sample when they were referred with such a diagnosis made by a multi-disciplinary team, based on standard diagnostic parameters (Carroll et al., 2004; Ruff et al., 2009). The criteria comprise a traumatic disruption of brain function, as manifested by at least one of the following: any loss of consciousness, any loss of memory for events immediately before or after the accident, any alteration in mental state at the time of the accident, and focal neurologic deficit(s) that may or may not be transient. The diagnostic parameters further include that the injury-related loss of consciousness does not exceed 30 minutes; an initial Glasgow Coma Scale score of 13–15 needs to be obtained after 30 minutes, and post-traumatic amnesia of less than 24 hours needs to be documented. All clients included in the sample were in the post-acute stage, ranging between 2 to 10 months post-injury at the time of assessment. None of the mTBI clients had a current or historic diagnosis of mental health disorder; none of the mTBI clients had a history of acquired brain injury (other than the current mTBI), including previous significant TBI, cerebral vascular accident, tumour, neuro-toxic exposure, HIV, dementia, or other cerebral conditions. Clients with mTBI who were suspected of extending incomplete test-effort according to the criteria of the Test of Memory Malingering (TOMM; Tombaugh, 2003) were excluded from the sample.

Clients with MDE were included in the study upon referral under such diagnosis according to DSM-IV TR criteria (American Psychiatric Association, 2000), assessed by the client's general practitioner (general medical officer) and confirmed separately by a psychiatrist. MDE was the only mental health diagnosis at the time of the assessment. For 77.5% of clients selected for the study this was their first formally diagnosed depressive episode. For 22.5% of the depression sample the current MDE represents a relapse; this sub-group can, by DSM IV TR definition, be classified as suffering a MDE in the context of a Major Depressive Disorder. This sample did not include clients in the depressive cycle of a diagnosed bipolar disorder. About 9% of clients had a previous anxiety disorder diagnosis. None of the clients had a history of psychotic disorder, including schizophrenia, or a

Table 2. Alternative test stories

Item	Criteria	Score	
Logical Memory Story A			
1. Maria	*Maria* or variation of name	0	1
2. Anderson	*Anderson* required	0	1
3. was a law student	Indication being a student/scholar of law	0	1
4. at Otago University.	*Otago University* or *University in Dunedin*	0	1
5. She and her two friends	Indication *two other people* joined her	0	1
6. Anna and	*Anna* or variation of name	0	1
7. Michael	*Michael* or variation of name	0	1
8. went skiing	Indication of any type of snow sport	0	1
9. in Queenstown	*Queenstown* required	0	1
10. over the winter holidays.	Indication of any type of vacation in the cold season	0	1
11. They arrived in the afternoon,	Indication of arrival in the afternoon	0	1
12. checked into their hotel	Indication of going to hotel	0	1
13. and went out for dinner.	Indication of going out for an evening meal	0	1
14. This night	Indication that it happened on the day of arrival	0	1
15. Maria fell ill	Indication of sickness	0	1
16. with fever,	Indication of elevated body temperature	0	1
17. nausea,	Indication of nausea of any type	0	1
18. headache,	Indication of pain/ache affecting the head	0	1
19. and stomach cramps.	Indication of stomach cramps or pain	0	1
20. The doctor advised her	Indication of medical attention	0	1
21. to stay in bed	Indication of bed-rest	0	1
22. for two days	*Two days* required	0	1
23. and to drink tea.	Indication of drinking tea	0	1
24. Maria recovered quickly	Indication of timely recovery	0	1
25. enjoying the rest of her holiday.	Indication of successful continuation of holiday	0	1
Logical Memory Story B			
1. Amanda	*Amanda* or variation of name	0	1
2. Wright	*Wright* required	0	1
3. was driving	Indication of driving	0	1
4. to the supermarket in her	Indication of shopping-related destination	0	1
5. blue	*Blue* required	0	1
6. Toyota	*Toyota* required	0	1
7. along Church Road,	*Church* Road or *Church* Street required	0	1
8. when she saw	Indication of noticing visually	0	1
9. a white	*White* required	0	1
10. limousine.	Indication of stretch-vehicle	0	1
11. She was excited	Indication of emotional arousal	0	1
12. thinking this may be a celebrity's car	Indication of famous person	0	1
13. visiting her town	Indication of temporary visit	0	1
14. for a concert.	Indication of public performance	0	1
15. She slowed down	Indication of slowing her car	0	1
16. and tried to get a closer look.	Indication of trying to see celebrity	0	1
17. Just then	Indication of simultaneous event	0	1
18. her two	*Two* required	0	1
19. young	Indication of youth	0	1
20. children,	Indication of children	0	1
21. sitting in their back seats,	Indication of back seat or back of the car	0	1
22. started quarrelling.	Indication of quarrel, or becoming noisy	0	1
23. She told them	Indication of Amanda verbally attending to her children	0	1
24. to be quiet,	Indication of request to behave	0	1
25. and continued her trip.	Indication of continuation of trip	0	1

personality disorder. Participating MDE clients had no history of acquired brain injury, including brain trauma, or other diagnosed cerebral conditions.

Inclusion criteria for the control group were absence of current or past mental health disorders (of any type), and absence of current or past acquired brain injury. Participants from all three groups had not been previously exposed to cognitive testing and were presented with the standard and new stories for the first time. The specific clinical client groups were chosen for this study as they are frequently subjected to neuro-cognitive research, are highly prevalent clinical populations, and were accessible to the author.

Design and procedure

The sets of WMS-IV standard stories and new stories were presented, in random order, to clients with MDE ($n = 80$), clients with mTBI ($n = 80$), and members of a control sample ($n = 80$), resulting in a total of 240 participants. Each client was given both standard stories and both new stories, whereby the order of presentation was alternated between consecutive clients. For instance, the first client was presented with the two new stories followed by the two standard stories, and the next client first heard the two standard stories followed by the two new stories. Within each set of stories the order was not varied; the first standard story was always followed by the second standard story, in accordance with the instructions of the WMS-IV test manual; correspondingly, the first new story was always followed by the second new story. Clients were asked to re-tell each sub-story immediately after hearing it. After recording the number of items recalled by the client, the next sub-story was presented. Less than 2 minutes elapsed between noting the recall of each sub-story and presenting the next sub-story. The stories of both sets were presented without a substantial break between sets. In accordance with the WMS-IV test instructions, raw scores were calculated by adding the number of items remembered from both sub-stories. After a 20–30 minute delay clients were asked to again re-tell standard and new stories, and the number of correct items was recorded for each set. For each client four raw scores were obtained: immediate recall of standard and new stories, and delayed recall of standard and new stories. The conversion of raw scores into scaled scores was undertaken based on the client's age at the time of the assessment and the WMS-IV conversion tables. The WMS-IV subtask Logical Memory Recognition was not presented for either standard or new stories.

Statistical methods

Each set of stories (standard and new) resulted in two types of recall scores: raw scores (number of items remembered) and scaled scores (age-weighted). Stratified by client group (mTBI, MDE, controls), correlation analyzes between recall scores for the standard stories and new stories were done separately for immediate and delayed recall. A mixed model ANOVA was used to explore the within-participant (new vs standard stories) and between-participant relationships (MDE, mTBI, controls), separately for each of the raw and scaled logical memory scores (immediate and delayed). Mixed model and independent sample *t*-tests were used to analyze the impact of education, age, and gender on immediate and delayed recall, both for standard and new stories. Wilcoxon rank sums tests were employed to investigate how the order of presentation of story sets impacted on scores. These analyses were done using PASW/SPSS version 18 (IBM Corporation, New York, USA).

RESULTS

Broadly consistent with the epidemiological distribution of the diagnostic categories is the apparent gender bias in this sample (American Psychiatric Association, 2000; Rickels, von Wild, & Wenzlaff, 2010; Tagliaferri, Compagnone, Korsic, Servadei, & Kraus, 2006). Men were significantly more frequently represented in the mTBI group, while the MDE groups had a significantly higher proportion of females than males (Table 3). The MDE group was significantly older than the mTBI and control groups. In addition, participants in the MDE group were significantly more highly educated than those in the MTBI and control groups; they had the highest proportion of participants (32.5%) with 16 or more years

Table 3. Participants' sociodemographic characteristics by study group

	Study group		
Characteristics	MDE ($n = 80$)	mTBI ($n = 80$)	Controls ($n = 80$)
Age (mean ± *SD*)	48.9 ± 9.1*	44.6 ± 11.6	40.4 ± 14.0
Age (min/max)	22–65	28–64	16–69
Gender (*n*, %)			
Female	49 (61.3)	26 (32.5)**	44 (55.0)
Male	31 (38.8)***	54 (67.5)	36 (45.0)
Years of education (*n*, %)			
≤ 8	0 (0)	0 (0)	5 (6.3)
9 to 11	2 (2.5)	9 (11.3)	15 (18.8)
12	24 (30.0)	31 (38.8)	30 (37.5)
13 to 15	28 (35.0)	30 (37.5)	21 (26.3)
≥ 16	26 (32.5)****	10 (12.5)	9 (11.3)
min/max	11–28	8–23	7–20
Ethnicity (*n*, %)			
Caucasian	75 (93.8)*****	56 (70.0)	63 (78.8)
Maori/Pacific	2 (2.5)	17 (22.3)*****	9 (11.3)
Asian	2 (2.5)	5 (6.3)	3 (3.8)
Other	1 (1.3)	1 (1.3)	5 (6.3)

MDE, Major Depressive Episode.

mTBI, Mild Traumatic Brain Injury.

*p** < .005; MDE group is older than mTBI and control group.

*p*** < .005; fewer women in mTBI group than in other groups.

*p**** < .005 fewer men in MDE group than women.

*p***** < .005; education ≥ 16 years occurs more frequently in MDE group than in other groups.

*p****** < .005; Caucasian more frequent in MDE group; Pacific/Maori more frequent in the mTBI group.

of education. The mTBI and control groups had more comparable levels of education. This educational bias in favor of the MDE group appears to be due to different referring agencies from which the clinical samples were sourced. In New Zealand blanket cover is provided by the National Accident Compensation Corporation for TBI and other personal injuries, resulting in clients with varying socio-economic levels being referred for assessment. In contrast, comprehensive neuro-cognitive assessment for depression is accessible mostly to clients with private insurance cover or private funds. The suspected socio-economic bias of the MDE group is further confirmed by the ethnic distribution, comprising the significantly highest number of Caucasian clients. In contrast, Maori/Pacific clients are significantly more frequently represented in the mTBI group than in other groups. The control group is well consistent with the ethnic distribution of the New Zealand population (Statistics New Zealand, 2010).

Within each study group the distributions of logical memory scores obtained from immediate and delayed assessments were comparable for WMS standard stories and the new stories developed in the present study (Table 4). This was demonstrated for the raw and scaled scores. Some differences were noted between different client groups, whereby the MDE and control group obtained similar results; the mTBI group obtained significantly lower scores in immediate and delayed recall. Both sets of stories (standard and new) demonstrated these group differences equally.

A mixed model ANOVA was used to explore the within-participants (new vs standard stories) and between-participants (MDE, mTBI, controls) relationships. There were no differences in immediate or delayed logical memory scores between the standard and new stories (within-participants differences) ($p > .88$), and effect sizes were small—immediate raw score ($\text{eta}^2 = .001$), immediate scaled score (eta squared = .003), delayed raw score (eta squared = .001), delayed scaled score (eta squared < .001). Differences between the diagnostic groups (MDE, MTBI, control) were significant ($p < .001$), but the effect sizes were small to medium—immediate recall raw score (eta squared = .034), immediate recall scaled score (eta squared = .038), delayed recall raw score (eta squared = .091), delayed recall scaled score (eta squared = .074).

In addition, Pearson correlations computed for the logical memory scores from the standard and new stories (raw scores and scaled scores) demonstrated very strong correlations in all three client groups and for both recall modalities (Table 5). The correlation coefficients ranged from .79 to .94, documenting high concurrent validity.

With regard to the order of presentation of story-sets (standard followed by new stories, or new followed by standard stories) Wilcoxon rank sums tests documented no overall differences in raw scores and scaled scores for both immediate and delayed recall. Standard and new stories equally confirmed that younger participants recalled more items (raw scores) than older participants immediately after story presentation ($p < .05$) and with 30 minutes delay ($p \leq .05$); no difference between stories was documented ($p > .35$). No age-related impact on scaled scores was found both according to standard and new stories (difference between age groups: $p > .63$; difference between stories: $p > .39$). No gender differences were found either for immediate ($p > .66$) or for delayed recall

Table 4. Differences between standard and new stories and between study groups

		Standard stories			New stories				*p***			*p****		
Logical memory	*n*	Mean	SD	Min/max	Mean	SD	Min/max	*p**	MDE group	mTBI group	Control group	MDE group	mTBI group	Control group
MDE group														
Immediate recall raw scores	80	27.69	6.50	14/41	27.64	6.49	15/42	.83	–	<.001	.09	–	<.001	.11
Delayed recall raw scores	80	24.15	5.80	9/37	24.06	5.54	8/38	.73	–	<.001	.21	–	<.001	.32
Immediate recall scaled scores	80	10.93	2.47	5/16	10.95	2.49	5/16	.78	–	<.001	.22	–	<.001	.16
Delayed recall scaled scores	80	11.03	2.27	5/16	10.99	2.17	4/16	.68	–	<.001	.05	–	<.001	.07
mTBI group														
Immediate recall raw scores	80	23.99	5.85	14/40	23.84	5.43	14/41	.56	<.001	–	.01	<.001	–	.001
Delayed recall raw scores	80	19.09	6.33	7/37	19.06	6.24	7/35	.92	<.001	–	<.001	<.001	–	<.001
Immediate recall scaled scores	80	9.39	2.42	5/15	9.40	2.25	5/16	.91	<.001	–	<.001	<.001	–	<.001
Delayed recall scaled scores	80	8.89	2.63	4/16	8.89	2.50	4/15	1.00	<.001	–	<.001	<.001	–	<.001
Control group														
Immediate recall raw scores	80	26.21	5.10	14/42	26.34	4.95	16/40	.72	.09	.01	–	.11	.001	–
Delayed recall raw scores	80	23.11	5.18	10.38	23.31	4.94	12/35	.57	.21	<.001	–	.32	<.001	–
Immediate recall scaled scores	80	10.50	2.01	5/16	10.51	1.89	6/15	.93	.22	<.001	–	.16	<.001	–
Delayed recall scaled scores	80	10.38	2.11	5/16	10.43	1.95	6/15	.74	.05	<.001	–	.07	<.001	–

MDE, Major Depressive Episode.
mTBI, Mild Traumatic Brain Injury.
*p** Difference between standard and new stories.
*p*** Difference between study groups (MDE, mTBI, control) according to standard stories.
*p**** Difference between study groups (MDE, mTBI, control) according to new stories.

Table 5. Pearson correlations between logical memory scores derived from standard and new stories by study group

		Pearson correlation coefficients	
Logical memory scores by study group	*n*	Immediate recall*	Delayed recall*
MDE			
Standard stories/new stories: raw scores	80	.94	.92
Standard stories/new stories: scaled scores	80	.94	.93
mTBI			
Standard stories/new stories: raw scores	80	.91	.93
Standard stories/new stories: scaled scores	80	.92	.94
Controls			
Standard stories/new stories: raw scores	80	.80	.80
Standard stories/new stories: scaled scores	80	.79	.77

*$p < .001$.
MDE, Major Depressive Episode.
mTBI, Mild Traumatic Brain Injury.

($p > .70$), according to standard and new stories (difference between stories $p > .74$). A significant impact of clients' levels of education on story-recall was noted, comparing less-educated clients (12 years or less) with more-educated clients (more than 12 years). Higher-educated clients performed significantly better both in immediate and delayed recall then less-educated clients ($p < .01$). This effect was demonstrated equally by standard and new stories ($p > .63$).

DISCUSSION

The results demonstrate that the newly created test stimuli have compatible structural and statistical properties as the established stories of the WMS-IV logical memory test for three clinical groups evaluated. Consistently high levels of compatibility were demonstrated for MDE, mTBI, and healthy control participants, representing focal areas of neuropsychological interest. Compatibility was also shown for different points of recall; that is, immediate recall and delayed recall of the newly acquired verbal (logical) information. Very high levels of correlation in raw scores and scaled scores were demonstrated between standard stories and new stories for the three groups and two points of assessment. Analyzing raw and scaled scores there were no significant differences between the standard and new stories, irrespective of whether recall was assessed immediately or after a 30-minute delay, within any of the study groups. Clinical application extending beyond the demographic characteristics of the current samples should be considered experimental at this time, and further clinical validation is needed.

Limitations of this study include a possible socio-economic bias of the MDE group with a disproportionately high representation of Caucasian and well-educated participants. While the mTBI and Control group are comparatively well matched in ethnicity and education to their respective reference populations, the use of a solely New Zealand test population may reduce the generalizability of results.

No comparison of standard and new stories had been undertaken for the sub-test Logical Memory Recognition. It should also be considered that scaled scores for the new stories were calculated based on the WMS-IV normative sample (Wechsler, 2009). There are insufficient data to assert that raw scores for the new stories will universally correspond to the score distribution provided by the WMS-IV normative sample for the standard stories. Furthermore the clinical samples did not include clients aged 16–21 years or older than 65 years, suggesting that additional validation efforts be undertaken for these age groups. Clearly clinicians should not derive demographically adjusted norms (e.g., correcting for education, gender and ethnicity) for the new logical memory stimuli using the Advanced Clinical Solutions (NCS Pearson, 2009) normative information. Further validation with different clinical samples and a greater diversity of healthy normative peers, particularly ethnic minorities, is needed.

It is worth noting, however, that even excellent alternative test stimuli provide no blanket protection against practice effects. Although learning of test material is the most obvious pitfall, Goldberg et al. (2010) pointed out that practice effects can occur as a result of decreasing test anxiety, having greater familiarity with the test settings, procedural learning, and improvement in test-taking strategy. In a double-blind, placebo-controlled pharmaceutical study, moderate improvements on repeated testing were documented even on a test (verbal list learning) in which alternative stimuli were used (Keefe et al., 2008). Goldberg et al. (2010) suggest that additional efforts are required to address the risk of practice effects unrelated to stimulus learning. Such efforts could include a period of surrogate testing or lead-in testing at the beginning of the assessment to increase clients' familiarity with the testing procedure, both from the perspective of anxiety management and from operational competency of clients on how to strategically approach test tasks.

REFERENCES

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision [DSM-IV-TR]*. Washington, DC: American Psychiatric Association.

Carroll, L. J., Cassidy, J. D., Holm, L., Kraus, J., Coronado, V. G., & The WHO Collaborating Centre Task Force on Mild Traumatic Brain Injury. (2004). Methodological issues and research recommendations for mild traumatic brain injury: The WHO Collaborating Centre Task Force on Mild Traumatic Brain Injury. *Journal of Rehabilitation Medicine, 43*, 113–125.

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

Collie, A., Darby, D. G., Falleti, M., Silbert, B. S., & Maruff, P. (2002). Determining the extent of cognitive change after coronary surgery: A review of statistical procedures. *The Annals of Thoracic Surgery*, *73*(6), 2005–2011. doi: 10.1016/s0003-4975(01)03375-6.

Cunje, A., Molloy, D., Standish, T. I., & Lewis, D. L. (2007). Alternate forms of logical memory and verbal fluency tasks for repeated testing in early cognitive changes. *International Psychogeriatrics*, *19*(1), 65–75.

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

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

Duff, K., Beglinger, L. J., Schoenberg, M. R., Patton, D. E., Mold, J., Scott, J. G., et al. (2005). Test–retest stability and practice effects of the RBANS in a community dwelling elderly sample. *Journal of Clinical and Experimental Neuropsychology*, *27*(5), 565–575. doi: 10.1080/13803390490918363.

Goldberg, T. E., Keefe, R. S., Goldman, R. S., Robinson, D. G., & Harvey, P. D. (2010). Circumstances under which practice does not make perfect: A review of practice effect literature in schizophrenia and its relevance to clinical treatment studies. *Neuropsychopharmacology*, *35*, 1053–1062.

Heaton, R. K., Gladsjo, J. A., Palmer, B. W., Kuck, J., Marcotte, T. D., & Jeste, D. V. (2001). Stability and course of neuropsychological deficits in schizophrenia. *Archives of General Psychiatry*, *58*(1), 24–32.

Heaton, R. K., Temkin, N., Dikmen, S., Avitable, N., Taylor, M. J., Marcotte, T. D., et al. (2001). Detecting change: A comparison of three neuropsychological methods, using normal and clinical samples. *Archives of Clinical Neuropsychology*, *16*(1), 75–91. doi: 10.1016/s0887-6177(99)00062-1.

Hubley, A. M. (2010). Using the Rey-Osterreith and Modified Taylor Complex Figures with older adults: A preliminary examination of accuracy score comparability. *Archives of Clinical Neuropsychology*, *25*(3), 197–203.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.

Keefe, R., Malhotra, A., Meltzer, H., Kane, J., Buchanan, R., Murthy, A., et al. (2008). Efficay and safety of denepezil in patients with schizophrenia or schizoaffective disorder: significant placebo/practice effects in a 12-week, randomised, double-blind, placebo-controlled trial. *Neuropsychopharmacology*, *33*, 1217–1228.

Lezak, M., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological Assessment* (4th ed.). New York: Oxford University Press.

Martin, R., Sawrie, S., Gilliam, F., Mackey, M., Faught, E., Knowlton, R., et al. (2002). Determining reliable cognitive change after epilepsy surgery: Development of reliable change indices and standardized regression-based change norms for the WMS-III and WAIS-III. *Epilepsia*, *43*(12), 1551–1558.

McCaffrey, R., Duff, K., & Westervelt, H. (2000). *Practitioner's Guide to Evaluating Change with Neuropsychological Assessment Instruments*. New York: Kluwer Academic/Plenum Press.

McCaffrey, R., Ortega, A., & Haase, R. F. (1993). Effects of repeated neuropsychological assessments. *Archives of Clinical Neuropsychology*, *8*(6), 519–524.

Meyers, J., & Meyers, K. (1995). *Rey Complex Figure Test and Recognition Trial*. Odessa, FL: Psychological Assessment Resource,Inc.

Morris, J., Kunka, J. M., & Rossini, E. D. (1997). Development of alternate paragraphs for the Logical Memory subtest of the Wechsler Memory Scale–Revised. *Clinical Neuropsychologist*, *11*(4), 370–374.

NCS Pearson (2009). *Advanced clinical solutions for WAIS–IV and WMS–IV (ACS): Administration and scoring manual*. San Antonio, TX: Psychological Corporation.

Randolph, C. (1998). *Repeatable Battery for the Assessment of Neuropsychological Status manual*. San Antonio, TX: The Psychological Corporation.

Rickels, E., von Wild, K., & Wenzlaff, P. (2010). Head injury in Germany: A population-based prospective study on epidemiology, causes, treatment and outcome of all degrees of head-injury severity in two distinct areas. *Brain Injury*, *24*(12), 1491–1504.

Ruff, R. M., Iverson, G. L., Barth, J. T., Bush, S. S., Broshek, D. K., Policy, N. A. N., et al. (2009). Recommendations for diagnosing a mild traumatic brain injury: A National Academy of Neuropsychology education paper. *Archives of Clinical Neuropsychology*, *24*(1), 3–10.

Salthouse, T., Schroeder, D., & Ferrer, E. (2004). Estimating retest effects in longitudinal assessments of cognitive functioning in aldults between 18 and 60 years of age. *Developmental Psychology*, *40*, 813–822.

Sawrie, S. M., Chelune, G. J., Naugle, R. I., & Luders, H. O. (1996). Empirical methods for assessing meaningful neuropsychological change following epilepsy surgery. *Journal of the International Neuropsychological Society*, *2*(6), 556–564.

Schmidt, M. (1996). *Rey Auditory Verbal Learning Test Handbook*. Los Angeles: Western Psychological Services.

Spreen, O., & Strauss, E. (1998). *A Compendium of Neuropsychological Tests* (2nd ed.). New York: Oxford University Press.

Statistics New Zealand. (2010). *The Social Report*. Wellington: New Zealand Ministry of Social Development.

Sullivan, K. (2005). Alternate forms of prose passages for the assessment of auditory-verbal memory. *Archives of Clinical Neuropsychology*, *20*(6), 745–753.

Tagliaferri, F., Compagnone, C., Korsic, M., Servadei, F., & Kraus, J. (2006). A systematic review of brain injury epidemiology in Europe. *Acta Neurochirurgica*, *148*(3), 255–268.

Taylor, L. B. (1969). Localisation of cerebral lesions by psychological testing. *Clinical Neurosurgery*, *16*, 269–287.

Tombaugh, T. N. (2003). The Test of Memory Malingering (TOMM) in forensic psychology. *Journal of Forensic Neuropsychology*, *2*(3&4), 69–96.

Uchiyama, C. L., D'Elia, L. F., Dellinger, A. M., Becker, J. T., Selnes, O. A., Wesch, J. E., et al. (1995). Alternate forms of the auditory-verbal learning test: Issues of test comparability, longitudinal reliability, and moderating demographic variables. *Archives of Clinical Neuropsychology*, *10*(2), 133–145. doi: 10.1016/0887-6177(94)e0034-m.

Wechsler, D. (1955). *Wechsler Adult Intelligence Scale*. New York: Psychological Corporation.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale – Revised*. New York: Psychological Corporation.

Wechsler, D. (1987). *Wechsler Memory Scale – Revised*. New York: Psychological Corporation.

Wechsler, D. (1997). *Wechsler Memory Scale III*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2009). *Wechler Memory Scale – Fourth Edition – Technical and Interpretive Manual* (4th ed.). San Antonio, TX: Pearson.