British Journal of Clinical Psychology (2003), 42, 407–425

www.bps.org.uk

Test–retest reliability, practice effects and reliable change indices for the recognition memory test

Chris M. Bird[1], Kyriaki Papadopoulou[1], Paola Ricciardelli[1], Martin N. Rossor[2] and Lisa Cipolotti[1]*

[1] Department of Neuropsychology, The National Hospital for Neurology and Neurosurgery Group, London, UK

[2] The Dementia Research Group, The National Hospital for Neurology and Neurosurgery, London, UK

Objectives. The Recognition Memory Test (RMT) is widely used; however, important characteristics such as reliability and stability over time were largely unknown. In this study, we document for the first time test–retest reliabilities, practice effects, and Reliable Change (RC) indices for this test.

Design. A sample of 206 normal adults (aged 40–70) were administered twice either the same version or two different versions of the RMT, with a 1-month interval between assessments. The normal sample was split into two groups; a young (aged 40–54) and an old (55–70) group.

Results. Test–retest reliabilities were modest when using either the same versions or different versions. Practice effects were abolished when different versions of the RMT were used. In contrast, practice effects were clearly present on the same version of the non-verbal subtest for both control groups. However, practice effects were present on the same versions of the verbal subtest only in the old group. RC indices were rather large when using the same or different versions.

Conclusion. Although modest, the test–retest reliability of the RMT is no worse than those reliabilities reported for other commonly used recall memory tests. Thus, the inherent clinical advantages of using a recognition paradigm make its use desirable. Usage of different versions of the RMT enables us to avoid practice effects. However, the RC indices indicate that large changes in scores are needed to detect a significant improvement or decline in an individual's performance.

* *Requests for reprints should be addressed to Lisa Cipolotti, Department of Neuropsychology, Box 37, The National Hospital for Neurology & Neurosurgery, Queen Square, London WC1N 3BG, UK (e-mail: l.cipolotti@ion.ucl.ac.uk)*

Repeated neuropsychological assessments play a central role in the monitoring of a variety of neurological conditions (e.g. Cipolotti & Warrington, 1995). Indeed, this use of the cognitive baselines has increased in importance. Repeated neuropsychological examinations are generally used to provide indications as to whether a pattern of cognitive deficit associated with brain damage is changing and, if so, in what way. For example, repeated assessments are used to evaluate the effectiveness of pharmacological and/or behavioural treatments for head injuries (e.g. Benedict, 1989) and surgical treatments for intractable epilepsy and Parkinson's Disease (e.g. Chelune, Naugle, Lüders, Sedlak, & Awad, 1993; Scott *et al.*, 1998). Repeated neuropsychological assessment also provides invaluable information about rates and patterns of cognitive decline in patients with neurodegenerative disorders (e.g. Perry & Hodges, 2000; Wilson, Gilley, Bennett, Beckett, & Evans, 2000). Reliable repeated assessment of memory functions is essential in all these pathologies.

Given the extensive use of repeated assessments, there are certain properties that neuropsychometric tests must have to be acceptable (see, for example, the Professional Affairs Board, 1980). These properties include reliability over time, and either resistance to practice effects or well-known practice effects, such that scores at reassessments can be adjusted according to the expected gains.

Reliability

Test–retest reliability refers to the correlation between scores obtained by the same individuals on the same test, separated by some period of time. It provides a measure of the variability that can be expected due to day-to-day fluctuations in a number of different factors, such as concentration, fatigue, etc. Test–retest reliabilities have been documented for a range of neuropsychometric tests. Generally, good or at least adequate reliabilities have been reported for tests of general intelligence (e.g. 0.82 for the WAIS-R; 0.7–0.9 for Raven's Progressive Matrices, e.g. Lezak, 1995, and Wechsler, 1981), of reading (e.g. 0.98 for the National Adult Reading Test; Nelson 1981; Crawford *et al.*, 1989), of naming (e.g. 0.62 to 0.89 for the Boston Naming Test; Kaplan, Goodglass, & Weintraub, 1983; Mitrushima & Satz, 1991; 0.92 for the Graded Naming Test; Bird *et al.*, in press), of phonemic fluency (e.g. 0.82; Harrison, Buxton, Husain, & Wise, 2000), and of speed and attention (e.g. 0.80 for the Symbol Digit Modalities Test; Smith, 1982).

In contrast, poorer reliabilities have been reported for memory tests (see Dikmen, Heaton, Grant, & Temkin, 1999). Amongst the different memory tests, the Wechsler Memory Scale (WMS; Wechsler, 1945) has been the most intensively studied. In a group of healthy older adults, Mitrushina and Satz (1991) found reliabilities of the WMS Logical Memory—immediate recall, and WMS Logical Memory—delayed recall to range from 0.62 to 0.81 over three annual probes. Similarly, in a large group of normal and neurologically impaired adults, Dikmen *et al.* (1999) found roughly comparable reliabilities of 0.58–0.70. Again, comparable values were found in a group of essential hypertensives (0.55–0.74), and in a group of chronic smokers (0.47–0.69; McCaffrey, Ortega, Orsillo, Nelles, & Haase, 1992).

Reliabilities of other memory tests are similarly poor. Mitrushina and Satz (1991) reported reliabilities for the Rey Auditory Verbal Learning Task (RAVLT; Rey, 1964; Taylor, 1959) and delayed recall of the Rey Osterrieth Complex Figure (Rey, 1941) to range from 0.41 to 0.79 and 0.57 to 0.77, respectively. The delayed match to sample reliabilities of the Cambridge Neuropsychological Test Automated Battery (CANTAB;

Sahakian & Owen, 1992) were all under 0.4 (Lowe & Rabbitt, 1998). Reliabilities of the component measures of the Selective Reminding Task (SRT; Buschke, 1973) ranged from 0.46 to 0.64 (Dikmen *et al.*, 1999). This common finding of relatively poor reliability of memory tests has been attributed to the increased variability that human memory performance has when compared with other cognitive skills (Dikmen *et al.*, 1999).

Practice effects

Practice effects are distinct from day-to-day fluctuations in performance and refer to a bias that is introduced at the second test session, due to familiarity with the test procedure and also specific test items. Thus, it is theoretically possible for a test to be very reliable and yet show large effects of practice. It is widely recognized that tests with a speeded component, with an infrequently practised response, or with an easily conceptualized solution are likely to show significant practice effects (Dodrill & Troupin, 1975; Lezak, 1995). Practice effects have been documented for some of the most commonly used neuropsychometric tests. For example, significant gains at follow-up assessments have been reported on the WAIS-R (Rawlings & Crewe, 1992), Stroop Test (Connor, Franzen, & Sharp, 1988), and Digit Symbol Modalities Test, (Smith, 1982) and in attenuated form in the NART (Crawford *et al.*, 1989). The majority of these studies quote the mean practice effect for the population.

However, recent studies have suggested that this is inadequate. Rabbitt and colleagues (Rabbitt, Diggle, Smith, Holland, & McInnes, 2001) reported that the practice effects shown by individuals vary according to their age, their ability, and some complex function of the task. Along these lines, Rapport, Brines, Axelrod, and Theisen (1997) found that practice effects in full-scale IQ scores on the WAIS-R depended on the initial IQ score. Thus, individuals with higher IQ scores at first assessment showed a greater improvement at subsequent assessments than those with lower initial IQ scores. Practice effects in memory tests also show population-specific effects. For example, Mitrushina and Satz (1991) reported that age affected the magnitude of practice effects on the delayed recall of the Rey–Osterrieth complex figure and the verbal reproduction subtest of the WMS. All age groups showed practice effects, but the younger participants showed a greater improvement at retest.

Another important variable influencing practice effects in memory tests is whether or not alternative forms are available. Benedict and Zgaljardic (1998) reviewed the evidence for practice effects in various memory tests, including for example the WMS, the SRT, and the RAVLT. When the same versions of the tests were used at retest, substantial practice effects were documented. However, when using alternative versions, practice effects were either abolished or greatly reduced. The authors also carried out a study investigating practice effects when using either the same versions or different versions of the Hopkins Verbal Learning Test–Revised (HVLT-R; Benedict, Schretlen, Groninger, & Brandt, 1998) and the Brief Visuospatial Memory Test–Revised (BVMT-R; Benedict, 1997). Practice effects were pervasive when the same version of the two tests was administered twice. However, practice effects were either absent or greatly reduced when alternative versions of the HVLT-R and the BVMT-R were used.

Reliable change indices

All tests include a degree of measurement error, which gives rise to fluctuations in individuals' scores between assessments (Chelune *et al.*, 1993). Indeed, even in the

healthy population, fluctuations in scores are generally present. The magnitudes of these changes in scores are generally assumed to be normally distributed (Jacobson & Traux, 1991). RC indices provide a measure of how large a change in score between two assessments must be to be clinically significant (Jacobson & Traux, 1991). These indices have been defined as a change in score that only occurs 5% (Jacobson & Traux, 1991) or 10% (Chelune *et al.*, 1993) of the time in a population. Thus, if one takes the 10% cut-off, RC indices are confidence intervals constructed about the mean change in score that are only exceeded 5% of the time in the positive direction and 5% of the time in the negative direction, assuming there has been no real change.

Chelune *et al.* (1993) made an important modification to the procedure used to calculate RC indices. The authors acknowledged that if a test shows practice effects, the upper and lower RC indices must be adjusted by the mean practice effect gain across the population. For example, if a test had RC indices of ± 7 and a mean practice effect of 2, the RC indices corrected for practice will be -5 and +9.

To the best of our knowledge, despite the obvious importance of RC indices, they have only been documented for a few neuropsychological tests. These include the WAIS, WAIS-R, WMS-R, and the Trail Making test (Chelune *et al.*, 1993; Dikmen *et al.*, 1999).

If, as suggested by Dikmen *et al.* (1999), human memory performance has greater variability when compared with other cognitive skills, this would result in large RC indices. In line with this, Chelune *et al.* (1993) reported very large RC indices for the subtests of the WMS-R (from ± 16.5 for visual memory to ± 20.7 for delayed recall). We are not aware of RC indices being reported for any other memory tests.

The recognition memory test

The Recognition Memory Test (RMT; Warrington, 1984) comprises a verbal (words) and a non-verbal (unfamiliar faces) subtest. It is commonly included in neuropsychological batteries, which are used both in routine clinical assessments and in clinically oriented research. Examples of this are represented by studies investigating the effects of neurosurgical treatments for epilepsy (e.g. Baxendale, 1997; Naugle, Chelune, Schuster, & Lüders, 1994), Parkinson's Disease (e.g. Scott *et al.*, 1998), colloid cyst removal (e.g. Aggleton *et al.*, 2000) as well as numerous studies of amnesic syndromes (e.g. Mayes, Meudell, & MacDonald, 1991; Parkin & Hunkin, 1993; for a review, see Aggleton & Shaw, 1996). The RMT has also been extensively used in research into neurodegenerative disease. Thus, this test has been used in studying the early memory changes associated with familial Alzheimer's Disease (e.g. Fox, Warrington, Seiffer, Agnew, & Rossor, 1998); the distinct cognitive profiles associated with Pick's Disease and Alzheimer's Disease (e.g. Chan *et al.*, 2001; Mummery *et al.*, 2000) and with early-onset autosomal dominant familial Alzheimer's Disease caused by mutation of the presenilin 1 gene (e.g. Janssen *et al.*, 2000, 2001). Many of these studies involved repeated administrations of the RMT.

Despite its extensive use, there have been only two very preliminary reports investigating test–retest reliability and practice effects (Coughlan & Hollows, 1985; Soukup, Bimbela, & Schiess, 1999). Coughlan and Hollows (1985) tested 30 normal subjects (age range 24–61) on two different versions of the RMT. The interval between first and second assessments was 1–6 days. The test–retest reliability was relatively low (0.55 for the verbal subtest and 0.63 for the non-verbal subtest). No practice effects were reported. However, it should be noted that the 30 subjects obtained rather high

scores at first assessment. Clearly, this could have masked any practice effects. Soukup *et al.* (1999) investigated reliability and practice effects on the same non-verbal subtest of the RMT. They tested 40 neurological patients and reported a test–retest reliability of 0.81 over an interval ranging from 2 to 20 months. No practice effects were found. The difference between test–retest reliabilities reported by these studies (approximately 0.6 in a healthy population vs. 0.81 in a neurological population) may be a direct result of the psychometric properties of the RMT itself. The RMT was designed for diagnostic purposes, and the range of scores possible by healthy subjects is rather narrow. However, the test has a long 'tail', enabling the severity of impairment in patient groups to be assessed. Thus, the scores of a group of patients may have a larger range than a group of healthy participants, and test–retest reliabilities are likely to be higher when the range of scores is greater. It should be noted that both studies used a rather small sample and did not consider variables such as the age of the participants. Neither of these two studies documented RC indices.

The lack of documented test characteristics such as test–retest reliability and practice effects has led some investigators to question the acceptability of the RMT (e.g. Kapur, 1987; Lezak, 1995). Indeed, Kapur concluded that this test '... suffers from significant defects ...' (p. 146), which not only limits its '... usefulness as a procedure for use in routine neuropsychological assessment ...' (p. 146) but also fails to meet all the criteria for acceptable test procedures.

Our study aimed to investigate the RMT's test–retest reliability, practice effects using both the same and an alternative version, and RC indices in both a young and an old control population.

Method

Participants

A total of 206 healthy volunteers participated in the study. Participants were recruited through posters placed in the National Hospital as well as in local churches, community centres, and an engineering company. The participants were aged between 40 and 70 years (mean age = 56.1; *SD* = 8.6) and had 13.1 years of education (*SD* = 3.7). There were 73 males and 133 females. This group was split into two age groups: a young and an old group. The mean age of the young group was 48.9 (*SD* = 4.0; range = 40–54; *N* = 102). The mean age of the old group was 63.1 (*SD* = 4.9; range = 55–71; *N* = 104).

Materials

Recognition memory for words (RMW)

There were two versions of this test (A and B). Version A was the published RMW (Warrington, 1984). Version B was an alternative version developed in the Neuropsychology Department of the National Hospital for Neurology and Neurosurgery. The words used for Version B were selected using identical criteria to those used for the published RMW. As with the RMW, these 50 words were 'short' (four to six letters), high in frequency (A or AA on the Thorndike-Lorge, 1944), and of relatively low imageability. The stimulus words were each typed in upper-case letters on a white card (6" × 4"), and then bound in a test booklet. In the recognition condition, both the stimulus words and the distractor items were typed in upper-case letters in two

columns of 25 pairs. The presentation order of the target items was not maintained in the recognition task, and the left/right position of the targets was randomized.

Recognition memory for faces (RMF)

Similarly, there were two versions of this test (A and B). Version A was the published RMF (Warrington, 1984), while Version B was a matched alternative version developed in the Neuropsychology Department. Version B comprised 50 black and white photographs of unfamiliar Caucasian male faces and 50 distractor faces. A wide variety of non-distinctive facial types were selected, as in the original version. Distractor items were chosen to be visually similar to the target. Each stimulus photograph was mounted on a white card (6" × 4") and then bound in a test booklet. For the recognition condition, a stimulus face and a distractor face were mounted one next to the other on white cards (8" × 5"). As with the procedure adopted for the words, the presentation order of the target items was not maintained in the recognition task, and the left/right position of the targets was randomized.

The national adult reading test (NART)

The NART was administered to all participants at the second assessment to obtain an estimation of the IQ of our sample. We followed the procedure as described in the second edition (Nelson & Willison, 1991).

Design

The Recognition Memory Test (RMT) was administered to both young and old age groups on two different occasions. There was a 1-month interval between the two assessments. The test was presented either in its original version (Version A, Warrington, 1984) or in the alternative Version B. On each occasion and within each age group, the two versions of the recognition memory test were administered with the alternation design shown in Table 1.

Table 1. Design for the two assessments

	Assessment 1	Order	Assessment 2	Order
SAME	RMT (Version A)	W F	RMT (Version A)	W F
DIFFERENT	RMT (Version A)	F W	RMT (Version B)	W F
SAME	RMT (Version B)	W F	RMT (Version B)	F W
DIFFERENT	RMT (Version B)	F W	RMT (Version A)	F W

Note. W = verbal subtest; F = non-verbal subtest.

Within each age group, half of the participants were tested with the SAME versions and the other half with DIFFERENT versions. At the second assessment, the NART was also administered.

Procedure

The same standardized administration procedure as that described in the RMT and NART manuals was used.

Results

NART and demographics

The 206 subjects who participated in both assessments had a mean NART IQ of 113.3 (*SD* = 10.9). Scores ranged from 79 to 129. There were no significant differences between the participants that were administered the SAME version of the RMT and the participants that were administered a DIFFERENT version, in terms of NART IQ, sex prevalence, and years of education. In addition, the two age groups were matched for NART IQ, sex prevalence, and years of education.

RMT

We first investigated whether our alternative version (B) of the RMT was matched for difficulty with the published version (A). The mean scores for both versions of the test at the first assessment are shown in Table 2.

Table 2. Mean scores correct obtained on both versions of the Recognition Memory Test

	RMW *M* (*SD*)	RMF *M* (*SD*)
Version A (*N* = 110)	46.6 (3.2)	43.4 (3.8)
Version B (*N* = 94)	46.3 (3.5)	42.4 (4.0)
p	0.58	0.07

Note. RMW = Recognition Memory Test (Words subtest); RMF = Recognition Memory Test (Faces subtest). *M* = Mean, *SD* = Standard Deviation.

The scores on both versions of the RMT were compared using an independent sample *t* test. There were no significant differences between the two versions of the RMW, although there was a trend difference for the two versions of the RMF. Given the large groups of participants involved in the study, this was considered unproblematic. Therefore, the two versions of the RMT were approximately matched for difficulty.

In a second analysis, we tested for a correlation between age, NART IQ, and performance on the RMT at first assessment (scores from both versions were combined). Age and performance on the NART were not significantly correlated with each other. Unsurprisingly, there were significant negative correlations between age and performance on the RMW and the RMF (-0.16, $p < 0.05$ and -0.22, $p < 0.01$, respectively). There was a significant correlation between performance on the NART and the RMW (0.18, $p < 0.02$). However, the correlation between performance on the NART and the RMF was not significant. This is slightly surprising, as Warrington (1984) found significant correlations between performance on two measures of general intelligence (Raven's Advanced Matrices and the Mill Hill Vocabulary Test; Raven, 1965, respectively) and performance on both the RMW and RMF. This may be because the NART provides a better estimate of verbal rather than non-verbal intellectual functioning. It is possible that verbal intelligence does not contribute substantially to performance on the RMF.

Thirdly, we compared performance of our sample on the RMT with the published normative data (Warrington, 1984). To make this an accurate comparison, we split our sample into two groups aged 40–54 and 55–70, as in the original standardization study.

Both the young and old groups in our sample performed better on the RMW than the published data, as tested by an independent samples t test (our sample, young group mean = 46.6, SD = 2.9 vs. 45.3, SD = 3.4, t = 3.06, df = 203, $p < 0.01$; our sample, old group mean = 46.0, SD = 3.7 vs. 43.0, SD = 4.5, t = 5.05, df = 194, $p < 0.001$). Our sample's performance on the RMF was not significantly different from the published data for either young or old groups (our sample, young group mean = 43.8, SD = 3.7 vs. 44.3, SD = 3.5; our sample, old group mean = 42.1, SD = 4.0 vs. 42.4, SD = 3.8). It is unclear why the performance of our participants should be better than the published control data on the RMW and not the RMF. Since the NART scores of our sample were rather high, this may indicate a generally high level of verbal ability in the sample. Another possibility is that the performance of the RMW in the population has generally increased, as was the case for a commonly used test of nominal skills (the Graded Naming Test; McKenna & Warrington, 1983; Warrington, 1997).

Test–retest reliability

Pearson correlation coefficients between performances on each assessment were taken as measures of test–retest reliability. The results of the subjects who performed the SAME version and the results of the subjects who had performed the DIFFERENT versions were analysed separately (see Table 3).

Table 3. Test–retest reliabilities of the Recognition Memory Test

	RMW Reliability	RMF Reliability
SAME (N = 102)	0.69	0.76
DIFFERENT (N = 102)	0.53	0.41

Note. RMW = Recognition Memory Test (Words subtest); RMF = Recognition Memory Test (Faces subtest).

Table 4. Test–retest reliabilities in groups split by age

		RMW Reliability	RMF Reliability
SAME	Young group (aged 40–54)	0.66	0.74
	Old group aged 55–70	0.76	0.76
DIFFERENT	Young group (aged 40–54)	0.33	0.55
	Old group (aged 55–70	0.57	0.27

Note. RMW = Recognition Memory Test (Words subtest); RMF = Recognition Memory Test (Faces subtest).

The correlations between scores at the first assessment and at the second assessment were all highly significant ($p < 0.001$). Overall, we found reliabilities comparable with those reported for other tests of memory. Reliabilities appeared higher for the SAME versions of the RMT, than for the DIFFERENT versions. The test–retest reliability

coefficients were compared using the Fisher transformation and a subsequent z test. The difference between the reliabilities of the SAME and DIFFERENT versions of the RMW tended towards significance ($p = 0.07$). The difference between the reliabilities of the SAME and DIFFERENT versions of the RMF was highly significant ($p < 0.001$).

In a second analysis, the individuals assessed on the SAME and DIFFERENT versions were split further according to age into a young and an old group (see Table 4). The reliabilities were similar, and there was no evidence that these were better in either younger or older subjects. In both the young and the old groups, reliability again appeared better for the SAME rather than for DIFFERENT versions. The differences in the reliabilities of the SAME and DIFFERENT versions were again tested using the Fisher transformation and subsequent z tests. The only differences that reached significance were between SAME and DIFFERENT versions of the RMW in the young group ($p < 0.05$) and between SAME and DIFFERENT versions of the RMF in the old group ($p < 0.001$). It should be noted that for the DIFFERENT versions, the reliabilities are worryingly low in the young group on the RMW and in the old group on the RMF.

Practice effects

Participants who had carried out the SAME versions of the RMT were analysed separately from those that had carried out DIFFERENT versions. Two-tailed paired t tests were used to test whether any changes in scores at retest were significant (see Table 5).

Table 5. Analysis of practice effects in the Recognition Memory Test

	RMW			RMF		
Group	Assessment 1 *M* (*SD*)	Assessment 2 *M* (*SD*)	*p*	Assessment 1 *M* (*SD*)	Assessment 2 *M* (*SD*)	*p*
SAME	46.1 (3.7)	46.5 (3.2)	0.137	43.0 (4.2)	45.6 (3.9)	<0.001
DIFFERENT	46.6 (2.9)	46.9 (3.2)	0.423	42.8 (3.7)	42.8 (4.2)	0.835

Note. RMW = Recognition Memory Test (Words subtest); RMF = Recognition Memory Test (Faces subtest). *M* = Mean correct, *SD* = Standard Deviation.

When the SAME versions were used at retest, gains in scores on the RMW were non-significant. However, there were significant practice effects on the RMF. When DIFFERENT versions were used, practice effects were abolished.

A second analysis investigated whether age affected practice effects. The individuals assessed on the SAME and DIFFERENT versions were further split by age (see Table 6).

When using the SAME version, the young group still showed significant practice effects only on the RMF. However, in the old group, practice effects were present on both the RMW and the RMF. The use of DIFFERENT versions abolished practice effects in both age groups.

Two further analyses investigated whether IQ (as predicted by the NART) affected practice effects. In the first analysis, individuals assessed on the SAME and DIFFERENT versions were split into higher, medium, and lower NART IQ groups. A one-way ANOVA revealed no significant differences in these three groups of controls, between

Table 6. Analysis of practice effects in groups split by age

		RMW			RMF		
Group		Assessment 1 *M* (*SD*)	Assessment 2 *M* (*SD*)	*p*	Assessment 1 *M* (*SD*)	Assessment 2 *M* (*SD*)	*p*
SAME	Young group (aged 40–54)	47.0 (2.8)	46.8 (3.5)	0.697	43.8 (3.8)	46.6 (3.3)	<0.001
	Old group (aged 55–70	45.3 (4.2)	46.2 (3.0)	<0.02	42.1 (4.5)	44.5 (4.1)	<0.001
DIFFERENT	Young group (aged 40–54)	47.3 (2.4)	47.8 (2.4)	0.306	43.7 (3.6)	43.4 (3.8)	0.518
	Old group (aged 55–70	46.0 (3.2)	46.1 (3.7)	0.828	42.0 (3.6)	42.2 (4.4)	0.845

Note. RMW = Recognition Memory Test (Words subtest); RMF = Recognition Memory Test (Faces subtest). *M* = Mean correct; *SD* = Standard Deviation.

the presence or absence of practice effects on the RMT. The second analysis investigated whether there was any correlation between NART IQ and the size of the practice effects. There was no correlation between NART IQ and the size of practice effects either in individuals assessed on the SAME or DIFFERENT versions of both the RMW and the RMF.

Reliable change indices

In order to estimate RC indices, the magnitudes of the changes in scores between two assessments must be normally distributed. We analysed separately the distribution of the changes in RMW and RMF scores for both the SAME and the DIFFERENT version. We found that they were normally distributed in all conditions. To illustrate this, Fig. 1 shows, as an example, the distribution of scores for the all participants assessed on DIFFERENT versions of the RMW.

Having established that the magnitude of changes in scores was normally distributed, we used Chelune *et al.'s* (1993) method to calculate RC indices corrected for practice. A 10% cut-off was adopted, and thus, 90% RC indices were calculated by multiplying the standard deviations of change by the appropriate value from the Normal distribution (1.64 in this case). Given our findings (see above), the only RC indices we needed to correct for practice were those relating to the SAME versions of the RMF and the SAME versions of the RMW in the old group only.

We found rather large RC indices for both age groups, when using both the SAME and DIFFERENT versions of the RMT (see Tables 7 and 8). Overall, RC indices are larger for the RMF than for the RMW, irrespective of whether the SAME or DIFFERENT versions were used. In addition, RC indices for the old group were larger than for the young group when using both the SAME and DIFFERENT versions of both subtests. More specifically, when using the SAME versions of the RMF, large changes in scores are needed to detect improvement. Relatively smaller changes in scores are needed to detect decline. For the SAME versions of the RMW, equally large changes in scores are needed to detect a significant improvement or decline in the young group. In the old group, large changes in scores are needed to detect improvement on the SAME RMW version, and relatively smaller changes in scores are needed to detect a decline in

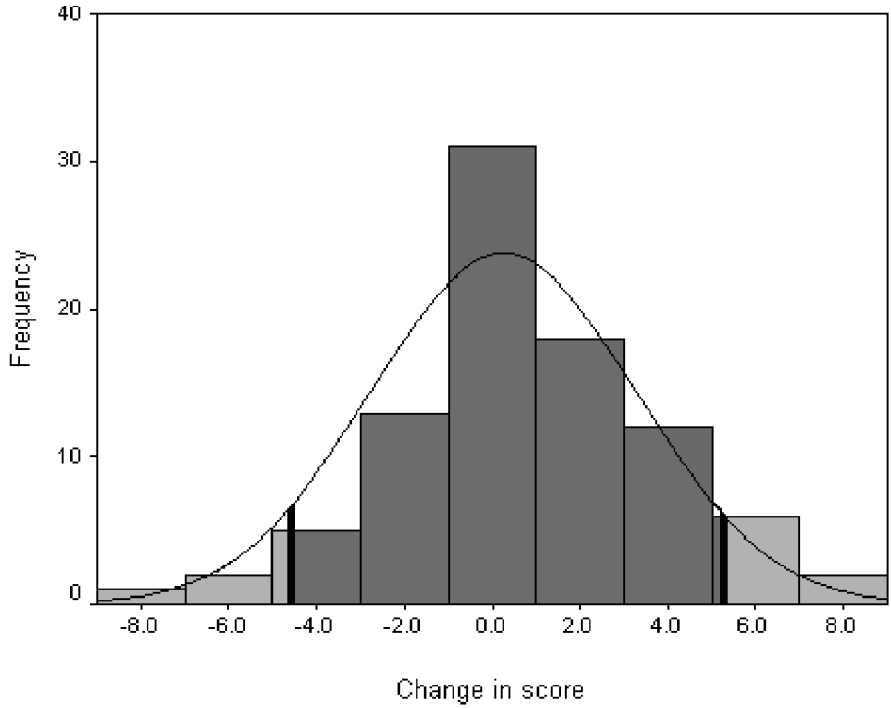

Figure 1. Distribution of the changes in scores for subjects carrying out different versions of the RMW. Bold lines indicate RC indices. Light grey areas fall outside the RC indices.

Table 7. Reliable Change (RC) indices for the Recognition Memory Test when using the SAME version corrected for practice

	RMW RC index	RMF RC index
Young group (aged 40–54)	±4.4	−1.5, +7.1
Old group (aged 55–70)	−3.6, +5.4	−2.6, +7.2

Note. RMW = Recognition Memory Test (Words subtest); RMF = Recognition Memory Test (Faces subtest).

Table 8. Reliable Change (RC) indices for the Recognition Memory Test when using DIFFERENT versions

	RMW RC index	RMF RC index
Young group (aged 40–54)	±4.5	±5.8
Old group (aged 55–70)	±5.3	±8.0

Note. RMW = Recognition Memory Test (Words subtest); RMF = Recognition Memory Test (Faces subtest).

performance. When using DIFFERENT versions of the RMT the magnitude of the changes in scores is the same, for detecting improvement and decline.

It is often useful in clinical practice to convert a patient's raw score on a test into a percentile score, as this permits performance on different tests to be compared directly (e.g. Spreen & Strauss, 1998). Appendix 2 in the manual for the RMT is a conversion table of raw scores to percentile scores (the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles; Warrington, 1984). Thus, an individual's raw score can be classified in terms of which percentile score it falls between (e.g. 10th–25th percentile). We have termed the intervals between these percentile scores as *percentile bands*. For example, a score at the 56th percentile that drops to the 12th percentile would be classed as dropping two percentile bands. In the second analysis, we evaluated the number of percentile band changes corresponding to the RMT's RC indices. The results of this analysis are shown in Tables 9 and 10.

Table 9. Number of percentile bands representing the RC indices when using the SAME version corrected for practice

		RMW Numbers of percentile bands	RMF Number of percentile bands
Young group (aged 40–54)	Improvement	1–2	2–3
	Decline	1–2	1
Old group (aged 55–70)	Improvement	1–2	3–4
	Decline	1–2	1–2

Table 10. Number of percentile bands representing the RC indices when using DIFFERENT versions

		RMW Numbers of percentile bands	RMF Number of percentile bands
Young group (aged 40–54)	Improvement	1–2	2–3
	Decline	1–2	2–3
Old group (aged 55–70)	Improvement	1–2	3–4
	Decline	1–2	3–4

Note. RMW = Recognition Memory Test (Words subtest); RMF = Recognition Memory Test (Faces subtest).

When using either the SAME or DIFFERENT versions of the RMW, changes of one to two percentile bands are needed in order to detect significant improvement or decline in a patient's performance irrespective of their age. For example, to conclude that a significant improvement has occurred on the RMW, an individuals' performance has to change from the 25th to the 75th percentile. Similarly, to conclude that a decline in performance is significant, a change from the 25th to the 5th percentile must be recorded.

In contrast, changes of two to four percentile bands are needed to detect improvement on the RMF across both age groups when using either the SAME or

DIFFERENT versions. When using the SAME version, detecting a significant decline in performance is easier in both age groups. When using DIFFERENT versions of the RMF, performance must decrease by two to three percentile bands in the young group and three to four percentile bands in the old group to detect decline. For example, an elderly patient's performance must drop from the 25th to below the 1st percentile to have significantly deteriorated. Since this study employed a test–retest interval of 1 month, it remains unclear whether the RC indices we documented would remain the same over longer intervals.

Discussion

Test–retest reliability

We found that scores on the RMT at the second assessment were significantly correlated with the scores at the first assessment. This was regardless of whether the same version or a different version was used at retest. Overall, the test–retest reliabilities of the RMT are modest and certainly below the proposed ideal value of $\geq$0.75–0.8 or above (e.g. Anastasi, 1988; Coolican, 1994; Sattler, 1992). Despite this, our result is in keeping with the moderate to low test–retest reliabilities reported for other memory tests (e.g. WMS and RAVLT; Mitrushina & Satz, 1991; SRT; Dikmen *et al.*, 1999). It suggests that perhaps these modest reliabilities should be expected for memory tests.

We investigated whether the use of the same or different versions of the RMT affected its reliability. We found that the reliability of the RMT was slightly better when the same rather than a different version was used at retest. Indeed, the test–retest reliability when using the same version was moderate for the verbal subtest (0.69) and fairly good for the non-verbal subtest (0.76). In line with this, Soukop *et al.* (1999) reported a test–retest reliability index of 0.81 in a small sample of neurological patients on the non-verbal subtest of the RMT. The reliability coefficients for the RMT when participants were administered a different version were poorer than when using the same version. This was true for both the verbal and the non-verbal subtests (0.53 and 0.41, respectively). Only one previous study has documented test–retest reliabilities of the RMT when using different versions at retest, and this study documented reliabilities similar to ours (Coughlan and Hollows, 1985).

There are at least two possible reasons why test retest reliabilities when using the same versions of the RMT may be higher than when using different versions. First, it is possible that when tested with the same version, individuals recall many of their former responses and therefore produce the same pattern of right and wrong responses. This would mean that the two administrations of the test are not independently obtained, and the correlations between them will be spuriously high (e.g. Anastasi, 1988).

Another explanation could be that the test–retest reliability of alternative versions of a test measures not only temporal stability but also consistency of response to different test versions (e.g. Anastasi, 1988). If there are differences between individuals' recognition performance due to the selection of test items, then the reliability when using a different version would be expected to be lower. For example, if an individual found the *specific* faces in version B of the non-verbal version particularly hard to recognize, then it is likely that they would perform comparably at the second assessment. However, if an alternative form was used, they might be expected to

perform better at the second assessment, and consequently the difference in scores between assessment would be greater. If item sampling does underlie this difference in reliabilities, it appears that this effect is greater for non-verbal material, as the differences in reliabilities between the same and different versions are greater for the non-verbal version. This effect should be considered when designing memory tests, especially those for non-verbal material.

Overall, our findings indicate that the RMT does not have any particular advantage over other memory tests in terms of reliability. However, the RMT does have other advantages. In particular, the RMT uses a recognition paradigm with a two-alternative forced-choice (2AFC) structure. This paradigm is less psychologically taxing to the patient than a free-recall procedure, which is often employed in other memory tests (e.g. the WMS-R). Indeed, Coughlan and Hollows (1984) reported that depressed patients performed as well as healthy controls on the RMT. In contrast, the same patients performed much worse than controls on verbal and visual recall tasks. In fact, their scores were so poor that on a visual recall test, for example, they were virtually indistinguishable from those obtained by neurological patients. These findings indicate that the recognition paradigm utilized by the RMT is rather more resistant to psychiatric conditions such as depression. Therefore, it represents a useful tool, for example, when the issue of a differential diagnosis needs to be addressed.

Given this advantage of forced-choice memory tests, it is worth considering whether their reliabilities could be improved. Ideally, a memory test should provide an index of how many items a subject can remember. However, when using a forced-choice procedure, the final score depends on both the number of items a subject remembers and on the number of items the subject correctly guesses. When using a 2AFC procedure, a subject still has at least a 0.5 probability of a correct response, even when guessing randomly. Thus, clearly a substantial degree of variability is introduced into the test. By increasing the number of distractors in the recognition task, correct guesses are less likely; this would obviously produce less variability. Some memory tests with larger numbers of distractors are available (e.g. the verbal and visual recognition subtests of the Doors and People test which use a 4AFC procedure; Baddeley, Emslie, & Nimmo-Smith, 1994; and the Topographical Recognition Memory Test which uses a 3AFC procedure; Warrington, 1996). Unfortunately, no information is available regarding reliability and practice effects for these tests. In addition, no alternative forms are available. Consequently, their suitability for monitoring changes in memory functions is, at this stage, somewhat limited.

Practice effects

In the old group, practice effects when using the same versions of the RMT were pervasive. On the non-verbal subtest, mean gains were approximately 2 to 3 points at retest. On the verbal subtest, mean gains were lower (approximately 1 point). In the young group, practice effects when using the same versions at retest were clearly present on the non-verbal subtest. These mean gains were of similar magnitude to those seen in the old group (nearly 3 points). However, in the young group, we failed to document practice effects for the verbal subtest. This may unfortunately be an artefact, due to the generally high performance of most of the young participants. Ceiling effects would have masked practice effects.

Some caution should be exercised before generalizing our findings to longer test–retest intervals. The only previous study to examine practice effects on the same

versions of the RMT at longer intervals (mean interval = 7 months) found no practice effects on the non-verbal subtest. The verbal subtest was not studied (Soukup, Bimbela, & Schiess, 1999). However, this study enrolled only 40 neurological patients, some of whom had probable Alzheimer-type dementia. This is a disease that is characterized by a progressive decline in memory function (e.g. McKhann *et al.*, 1984). Thus, it is probable that the memory scores of some of these patients may have declined over the time interval considered by the study. This may well have masked practice effects. It therefore remains unclear whether practice effects on the same version of the RMT are present or absent over longer intervals.

In general, very little is known about practice effects on memory tests over longer intervals. To the best of our knowledge, there is only one study, which reported the performance of healthy elderly controls over the course of a year on four verbal and visual memory recall tasks (Mitrushina & Satz, 1991). Practice effects were pervasive in all tasks.

When using different versions at retest, practice effects are completely abolished across age groups on both subtests. This finding replicates and extends the results of Coughlan and Hollows (1985). The authors reported no evidence for practice effects when using different versions of the RMT (test–retest interval = 1–6 days) in their small group of 30 healthy adults. This lack of practice effects when using different versions of the RMT clearly represents a significant advantage over the use of the same version. Thus, to avoid the confounding effects of practice when carrying out multiple assessments, the use of a different version of the RMT is advisable.

The fact that we find practice effects when using the same, but not different, versions allows us to speculate on the basis for this effect. Anastasi (1988) coined the term 'test sophistication' to describe procedural learning. This reflects memory for general task demands and development of effective strategies for performing the test. If the practice effects we document were underpinned by test sophistication, they should have been present when different versions were used. The lack of practice effects when using different versions suggests that the RMT is relatively resistant to these procedural learning effects. The practice effects that we documented when using the same version of the test are likely to represent item-specific learning. In line with this, Benedict and Zgaljardic (1998) documented evidence for item-specific learning on the HVLT-R and the BVMT-R.

Finally, we failed to document IQ-related practice effects on the RMT. Other authors have found such effects, for example, on tests of general intelligence (Rapport *et al.*, 1997) and on the rate of improvement at complex video games (Rabbitt, Banerji, & Szymanski, 1989). Our findings suggest that IQ does not play a significant role in determining practice effects on the RMT.

RC indices

We calculated the RMT's RC indices to establish when a change in a patient's performance is significant. We found that fluctuations in scores in our normal control group tended to be rather large. Consequently, our RC indices when using either the same or different versions are also large. On the verbal subtest, an improvement or decline in performance is associated with a change in score of one to two percentile bands, irrespective of whether the same or different versions are used. On the same non-verbal version, performance must rise by two to four percentile bands to be judged as significantly better. In contrast, a drop in performance of one to two percentile

bands represents a significant decline. On the different non-verbal versions, changes of two to four percentile bands represent both significant improvements and declines. These rather large RC indices suggest that the RMT is capable of detecting significant changes in individuals' scores only when they are reasonably substantial. The RMT is less effective at monitoring subtle changes in scores. In fact, only very large changes in performance represent a significant improvement or decline on the non-verbal version, particularly in the old group. However, the RMT may still be suitable even for detecting small changes in the scores of a population of patients, since individual variability is likely to average out.

It is difficult to ascertain whether large RC indices are typical of memory tests in general. At present, we know of only one study which documented RC indices for a memory test (WMS-R). Chelune *et al.* (1993) reported very large RC indices. In addition, the poor overall test–retest reliability of memory tests (e.g. Dikmen *et al.*, 1999) would imply that memory tests are associated with a rather large degree of variability over time. Consequently, large RC indices should be expected. Therefore, although the RMT's RC indices are larger than ideal, this may simply reflect a common feature of memory tests.

Conclusion

This study indicated that, although modest, the test–retest reliability of the RMT is no worse than those reported for other commonly used memory tests. Given this fact, the inherent advantages of the RMT's recognition paradigm make its usage desirable. If used to monitor changes in memory function, our results suggest that usage of different versions is desirable to avoid practice effects. On an individual basis, to obtain valid judgments regarding changes in a patient's performance, a rather large change in scores needs to be documented, as indicated by the RC indices. In the light of our findings related to practice and RC indices, the importance of investigating reliability and stability for cognitive tests is stressed.

Acknowledgements

We would like to thank Prof. Tim Shallice for helpful comments on an earlier draft of this paper and Ms Hilary Watt for advice on the statistics. We would also like to thank David Jenkins, Bob Broome, Rolandos Louca, and Pat Hawkey, at WS Atkins in Epsom for their help in recruiting volunteers, and Angela Bird for comments on the English. This research was funded in part by Program Grant G9626876 from the Medical Research Council (UK).

References

Aggleton, J. P. & Shaw, C. (1996). Amnesia and recognition memory: A re-analysis of psychometic data. *Neuropsychologia*, *34*, 51–64.

Aggleton, J. P., McMackin, D., Carpenter, K., Hornak, J., Kapur, N., Halpin, S., Wiles, C. M., Kamel, H., Brennan, P., Carton, S., & Gaffan, D. (2000). Differential cognitive effects of colloid cysts in the third ventricle that spare or compromise the fornix. *Brain, 123*(Pt. 4), 800–815.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Baddeley, A., Emslie H., & Nimmo-Smith, I. (1994). *Doors and people*. Bury St Edmunds: Thames Valley Test Company.

Baxendale, S. A. (1997). The role of the hippocampus in recognition memory. *Neuropsychologia, 35*, 591–598.

Benedict, R. H. B. (1989). The effectiveness of cognitive remediation strategies for victims of traumatic head injury: A review of the literature. *Clinical Psychology Review, 9*, 605–626.

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

Benedict, R. H. B., Schretlen, D., Groninger, L., & Brandt, J. (1998). Revision of the Hopkins Verbal Learning Test: Normative data and analysis of inter-form and test–retest reliability. *The Clinical Neuropsychologist, 5*, 125–142.

Benedict, R. H. B., & Zgaljardic, D. J. (1998). Practice effects during repeated administrations of memory tests with and without alternative forms. *Journal of Clinical and Experimental Neuropsychology, 20*, 339–352.

Bird, C. M., Papadopoulou, K., Ricciardelli, P., Rossor, M. N., & Cipolotti, L. (in press). Monitoring cognitive changes: Psychometric properties of six cognitive tests. *British Journal of Clinical Psychology*.

Buschke, H. (1973). Selective reminding for analysis of memory and learning. *Journal of Verbal Learning and Verbal Behavior, 12*, 543–550.

Chan, D., Fox, N. C., Scahill, R. I., Crum, W. R., Whitwell, J. L., Leschziner, G., Rossor, A. M., Stevens, J. M., Cipolotti, L., & Rossor, M. N. (2001). Patterns of temporal lobe atrophy in semantic dementia and Alzheimer's disease. *Annals of Neurology, 49*, 433–442.

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.

Cipolotti, L., & Warrington, E. K. (1995). Neuropsychological assessment. *Journal of Neurology, Neurosurgery & Psychiatry, 58*, 655–664.

Connor, A., Franzen, M., & Sharp, B. (1988). Effects of practice and differential instructions on Stroop performance. *International Journal of Clinical Neuropsychology, 10*, 1–4.

Coolican, H. (1994) *Research methods and statistics in psychology*. London: Hodder & Stoughton.

Coughlan, A. K., & Hollows, S. E. (1984). Use of memory tests in differentiating organic disorder from depression. *British Journal of Psychiatry, 145*, 164–167.

Coughlan, A. K., & Hollows, S. E. (1985). *The adult memory and information processing battery*. Leeds: Coughlan, St James's University Hospital.

Crawford, J. R., Stewart, L. E., Cochrane, R. H., Foulds, J. A., Besson, J. A., & Parker, D. M. (1989). Estimating premorbid IQ from demographic variables: Regression equations derived from a UK sample. *British Journal of Clinical Psychology, 28*(Pt. 3), 275–278.

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.

Dodrill, C. B., & Troupin, A. S. (1975). Effects of repeated administrations of a comprehensive neuropsychological battery among chronic epileptics. *Journal of Nervous and Mental Disease, 161*, 185–190.

Fox, N. C., Warrington, E. K., Seiffer, A. L., Agnew, S. K., & Rossor, M. N. (1998). Presymptomatic cognitive deficits in individuals at risk of familial Alzheimer's disease. A longitudinal prospective study. *Brain, 121*(Pt. 9), 1631–1639.

Harrison, J. E., Buxton, P., Husain, M., & Wise, R. (2000). Short test of semantic and phonological fluency: Normal performance, validity and test–retest reliability. *British Journal of Clinical Psychology, 39*(Pt. 2), 181–191.

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

Janssen, J. C., Hall, M., Fox, N. C., Harvey, R. J., Beck, J., Dickinson, A., Campbell, T., Collinge, J., Lantos, P. L., Cipolotti, L., Stevens, J. M., & Rossor, M. N. (2000). Alzheimer's disease due to an

intronic presenilin-1 (PSEN1 intron 4) mutation: A clinicopathological study. *Brain, 123*(Pt. 5), 894–907.

Janssen, J. C., Lantos, P. L., Fox, N. C., Harvey, R. J., Beck, J., Dickinson, A., Campbell, T. A., Collinge, J., Hanger, D. P., Cipolotti, L., Stevens, J. M., & Rossor, M. N. (2001). Autopsy-confirmed familial early-onset Alzheimer disease caused by the I153V presenilin 1 mutation. *Archives of Neurology, 58*, 953–958.

Kaplan, E. F., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test* (2nd ed.). Philadelphia: Lea & Febiger.

Kapur, N. (1987). Some comments on the technical acceptability of Warrington's Recognition Memory Test. *British Journal of Clinical Psychology, 26*(Pt. 2), 144–146.

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

Lowe, C., & Rabbitt, P. (1998). Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: Theoretical and practical issues. Cambridge Neuropsychological Test Automated Battery. International study of post-operative cognitive dysfunction. *Neuropsychologia, 36*, 915–923.

McCaffrey, R. J., Ortega, A., Orsillo, S. M., Nelles, W. B., & Haase, R. F. (1992). Practice effects in repeated neuropsychological assessments. *The Clinical Neuropsychologist, 6*, 32–42.

McKenna, P., & Warrington, E. K. (1983). *The Graded Naming Test*. Windsor: NFER-Nelson.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS–ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology, 34*, 939–944.

Mayes, A. R., Meudell, P. R., & MacDonald, C. (1991). Disproportionate intentional spatial-memory impairments in amnesia. *Neuropsychologia, 29*, 771–784.

Mitrushina, M., & Satz, P. (1991). Effect of repeated administration of a neuropsychological battery in the elderly. *Journal of Clinical Psychology, 47*, 790–801.

Mummery, C. J., Patterson, K., Price, C. J., Ashburner, J., Frackowiak, R. S., & Hodges, J. R. (2000). A voxel-based morphometry study of semantic dementia: Relationship between temporal lobe atrophy and semantic memory. *Annals of Neurology, 47*, 36–45.

Naugle, R. I., Chelune, G. J., Schuster, J., & Lüders, H. O. (1994) Recognition memory for words and faces before and after temporal lobectomy. *Assessment, 1*, 373–381.

Nelson, H., E. (1981) *The National Adult Reading Test (NART): Test manual*. Windsor: NFER-Nelson.

Nelson, H. E., & Willison, J. R. (1991). *The National Adult Reading Test* (2nd ed.). Windsor: NFER-Nelson.

Parkin, A. J., & Hunkin, N. M. (1993) Impaired temporal context memory on anterograde but not retrograde tests in the absence of frontal pathology. *Cortex, 29*, 267–280.

Perry, R., & Hodges, J., R. (2000). Fate of patients with questionable (very mild) Alzheimer's disease: Longitudinal profiles of individual subjects' decline. *Dementia and Geriatric Cognitive Disorders, 11*, 342–349.

Professional Affairs Board (1980). Technical recommendations for psychological tests. *Bulletin of The British Psychological Society, 33*, 161–164.

Rabbitt, P., Diggle, P., Smith, D., Holland, F., & McInnes, 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.

Rabbitt, P. M. A., Banerji, N., & Szymanski, A. (1989). Space Fortress as an IQ test? Predictions of learning and of practised performance in a complex interactive video-game. *Acta Psychologica, 71*, 243–257.

Rapport, L. A., Brines, D. B., Axelrod, B. N., & Theisen, M. E. (1997). Full scale IQ as mediator of practice effects: The rich get richer. *The Clinical Neuropsychologist, 11*, 375–380.

Raven, J. C. (1965). *The Mill Hill Vocabulary Scale*. London: Lewis.

Raven, J. C. (1965). *Advanced Progressive Matrices Sets I and II*. Oxford: Oxford Psychological Press Ltd.

Rawlings, D. B., & Crewe, N. M. (1992). Test–retest practice effects and test score changes of the WAIS-R in recovering traumatically brain-injured survivors. *The Clinical Neuropsychologist, 6*, 415–430.

Rey, A. (1941) L'examen psychologique dans les cas d'encephalopathie traumatique. *Archives de Psychologie, 28*, 286–340.

Rey, A. (1964) *L'examen clinique en psychologie*. Paris: Presses Universitaires de France.

Sahakian, B. J,. & Owen, A. M. (1992) Computerised assessment in neuropsychiatry using CANTAB. *Journal of the Royal Society of Medicine, 85*, 399–402.

Sattler, J. M., (1992). *Assessment of children* (3rd ed., p. 25). San Diego, CA: Jerome M. Sattler.

Scott, R., Gregory, R., Hines, N., Carroll, C., Hyman, N., Papanasstasiou, V., Leather, C., Rowe, J., Silburn, P., & Aziz, T. (1998). Neuropsychological, neurological and functional outcome following pallidotomy for Parkinson's disease. A consecutive series of eight simultaneous bilateral and twelve unilateral procedures. *Brain, 121*(Pt. 4), 659–675.

Smith, A. (1982). *Symbol Digit Modalities Test (SDMT). Manual (Revised)*. Los Angeles: Western Psychological Services.

Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests*. (2nd edn). New York: Oxford University Press.

Soukup, V. M., Bimbela, A., & Schiess, M. C. (1999). Recognition memory for faces: Reliability and validity of the Warrington Recognition Memory Test (RMT) in a neurological sample. *Journal of Clinical Psychology in Medical Settings, 6*, 287–293.

Taylor, E. M. (1959). *The appraisal of children with cerebral deficits*. Cambridge: Harvard University Press.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Teachers College, Columbia University.

Warrington, E. K. (1984). *Recognition Memory Test*. Windsor: NFER-Nelson.

Warrington, E. K. (1996). *The Camden Memory Tests*. Hove: Psychology Press.

Warrington, E. K. (1997). The Graded Naming Test: A restandardisation. *Neuropsychological Rehabilitation, 7*, 143–146.

Wechsler, D. (1945). A standardised memory scale for clinical use. *Journal of Psychology, 19*, 87–95.

Wechsler, D. (1981). *WAIS-R manual*. New York: The Psychological Corporation.

Wilson, R. S., Gilley, D. W., Bennett, D. D., Beckett, L. A., & Evans, D. A. (2000). Person-specific paths of cognitive decline in Alzheimer's disease and their relation to age. *Psychology and Aging, 15*, 18–28.

Received 22 February 2002; revised version received 2 October 2002