British Journal of Clinical Psychology (2004), 43, 197–210

www.bps.org.uk

Monitoring cognitive changes: Psychometric properties of six cognitive tests

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

[1] Department of Neuropsychology;
[2] Dementia Research Group, National Hospital for Neurology and Neurosurgery, London, UK

Objectives. Repeated neuropsychological assessments are often used to monitor change in cognitive functioning over time. Thus, knowledge about the reliability and stability of neuropsychological tests and the effects of age and IQ is of paramount importance. In this study we document, for six cognitive tests: test–retest reliabilities, practice effects, reliable change (RC) indices corrected for practice, and the impact of premorbid IQ and age.

Design. A sample of 188 normal adults (aged 40–70 years) were administered, on two occasions, one or more of the following tests: the Graded Naming Test (GNT), the Silhouettes Test, two tests of verbal fluency, the Modified Wisconsin Card Sorting Test, and a new test of speed and attention (the Symbol Digit Test). There was a 1-month interval between assessments. At first assessment, all participants were administered the revised National Adult Reading Test (NART).

Results. The test–retest reliability of the tests ranged from very good (the GNT and Silhouettes Test) to moderate (verbal fluency tests and Symbol Digit Test) and to poor (Modified Card Sorting Test). Significant, although modest, practice effects were found on all tests. RC indices were generally large except for the Graded Naming Test and the Silhouettes Test. Premorbid IQ scores significantly correlated with performance on all the tests, the exception being semantic fluency. Age only correlated with the Silhouettes Test and the new Symbol Digit Test. Neither NART IQ nor age correlated with practice effects.

Conclusion. The psychometric properties of the GNT and Silhouettes Test indicated that they are useful tools for monitoring even small cognitive changes. In contrast, the verbal fluency tests and the new Symbol Digit Test are only suitable for monitoring large changes in performance. The Modified Card Sorting Test is an unreliable tool for monitoring 'executive' functions.

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

A comprehensive neuropsychological assessment requires the evaluation of a variety of different cognitive functions. Thus, neuropsychological assessments often include tests tapping nominal, perceptual, 'executive' functions and speed/attention as well as memory functions (e.g. Lezak, 1995). Repeated neuropsychological assessments allow the monitoring of cognitive functions over time (e.g. Cipolotti & Warrington, 1995). The cognitive baselines provided by assessments 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 (e.g. Bruggemans, van de Vijver, & Huysmans, 1997; Chelune, Naugle, Luders, Sedlak, & Awad, 1993; McCaffrey, Duff, & Westervelt, 2000; Wilson, Watson, Baddeley, Emslie, & Evans, 2000).

Given that the purpose of repeated assessments is to monitor changes in cognitive function over time, information concerning the psychometric properties of the tests used is of paramount importance (e.g. Bird, Papadopoulou, Ricciardelli, Rossor, & Cipolotti, 2003; Lowe & Rabbitt, 1998; McCaffrey, *et al.*, 2000). These properties include their test–retest reliability, practice effects and measures to establish whether significant change has occurred, such as RC indices corrected for practice. Test–retest reliability gives an indication of how much variability can be expected between assessments. Practice effects give an indication of whether scores at reassessment are likely to increase and, if so, by how much. Finally, RC indices provide a measure of how large an individual's change in scores between two assessments must be to exceed normal variation and therefore be indicative of significant improvement or decline. It should be noted that there has recently been considerable debate about which statistical methods are best at detecting significant or 'real' neuropsychological changes in individuals (e.g. Temkin, Heaton, Grant, & Dikmen, 1999). In addition to RC indices corrected for practice, regression models have been proposed that correct for the effects of regression towards the mean, as well as practice effects. Despite this, a recent study concluded that RC indices corrected for practice were no less accurate than regression models (Heaton *et al.*, 2001). Whilst such debate is important, a factor that is often understressed is the need for normative data to be easily interpretable and simple to use. Normative data (whether for changes in scores or for one-off assessments) can only be used to guide interpretation of the results from a neuropsychological assessment, rather than provide categorical information regarding the impairment of a cognitive domain. Therefore, data that are quick and simple to understand and use are preferable to more laborious techniques for interpreting change which may be only marginally more accurate. For this reason, RC indices corrected for practice are used in the present study.

It is somewhat surprising that for a large number of cognitive tasks which are used both in the clinical, as well as the research context, important psychometric properties are largely unknown. For example, it is only very recently that the psychometric properties of the Recognition Memory Test, widely used to assess verbal and visual memory functions, both for clinical as well as research purposes have been documented (Bird *et al.*, 2003). The present study aimed to document the test–retest reliabilities, practice effects and RC indices for six cognitive tests in a large sample of healthy adults over a 1-month interval. The cognitive tests are: the Graded Naming Test (McKenna & Warrington, 1983), and the Silhouettes Test (Warrington & James, 1991), two verbal fluency tests, the Modified Card Sorting Test (Nelson, 1976) and a new test of speed and attention – the new Symbol Digit Test.

The *Graded Naming Test* (GNT; McKenna & Warrington, 1983) is a stringent test of nominal functions. This test involves the oral naming of 30 pictures of objects and has

been widely used both in clinical practice and in clinical research (e.g. Garrard *et al.*, 2001; Kapur, Ironside, Abbott, Warner, & Turner 2001; Langdon & Thompson, 1999). Despite its extensive use, its test–retest reliability, practice effects and RC indices are as yet unknown.

The *Silhouettes Test* is a stringent visual perception subtest of the Visual Object and Space Perception battery (VOSP; Warrington & James, 1991). It involves the identification of silhouettes of animals and objects. Similarly to the GNT, the Silhouettes Test is widely used in clinical practice and in clinically oriented research (e.g. Binetti *et al.*, 1996; Holdstock *et al.*, 2000; Langdon & Thompson, 1999; Ross & Hodges, 1997). Again, somewhat surprisingly, its test–retest reliability, practice effects and RC indices are unknown.

Verbal fluency tests are commonly used to assess 'executive' functions. They have proved to be particularly useful in the detection and differentiation of different types of degenerative disorders (e.g. Mathuranath, Nestor, Berrios, Rakowicz, & Hodges, 2000; Monsch *et al.*, 1992). Verbal fluency tests are usually either tests of phonological or semantic fluency. The tests involve producing as many words as possible which belong either to the same phonological category (e.g. starting with the letter 'S') or semantic category (e.g. 'animals').

A recent and comprehensive study has documented the test–retest reliability and practice effects for various verbal fluency tests in the British population (Harrison, Buxton, Husain, & Wise, 2000). In the phonological fluency tests, four categories were used (the letters 'F', 'A', 'S' and 'B'). In the semantic fluency test, the category was 'animals'. The study enrolled 90 participants and used a variable interval of 1–8 weeks, but no information was provided of the numbers of people assessed at the different intervals. The reliabilities reported were reasonably good, being .82 for the long phonological fluency test (letters F, A and S), .73 for the shortened phonological fluency test (letter B) and .68 for the semantic fluency test (animals). Only around 60% of participants showed improvements in performance at second assessment, indicating that practice effects are not inevitable. No RC indices were provided by this study nor, more generally, have they been reported elsewhere in the literature.

The *Wisconsin Card Sorting Test* (Heaton, 1989; Heaton, Chelune, Talley, Kay, & Curtiss, 1993) is widely considered as the 'frontal' or 'executive' test *par excellence*. The classic version of this test was altered by Nelson in 1976 to lower the demands made on the patient as well as simplify the scoring procedure. Nelson's revised version is termed the *Modified Card Sorting Test* (MCST; Nelson, 1976) and is widely used both in clinical practice and in research (e.g. Gotham, Brown, & Marsden, 1988; Joyce & Robbins, 1991; Mathias & Coats, 1999).

To the best of our knowledge three studies have documented test–retest reliability and practice effects on the MCST (de Zubicaray, Smith, Chalk, & Semple, 1998; Lineweaver, Bondi, Thomas, & Salmon, 1999, Wilson, Alderman, Burgess, Emslie, & Evans, 1996). All these studies documented low to modest test–retest reliabilities and modest practice effects. The study by Lineweaver *et al.*, (1999) is by far the most comprehensive, enrolling 142 older adults (mean age = 69 years, SD = 8.58). The authors document test–retest reliabilities ranging from .56 to .64 on different indices of MCST performance. Rather surprisingly, the other two previous studies document poorer test–retest reliabilities despite using shorter retest intervals (6–12 months in both studies). However, these studies enrolled a limited number of participants (36 in de Zubicaray *et al.*'s study and 29 in Wilson *et al.*'s study), and therefore the test–retest reliability over intervals of less than a year remains somewhat unclear. In addition, it is

known that performance on the MCST is correlated with age and IQ (e.g. de Zubicaray *et al.*, 1998; Obonsawin *et al.*, 1999). Age and IQ have been reported to affect practice effects on several cognitive tests (e.g. Horton, 1992; Lowe & Rabbitt, 1998; Rabbitt, Diggle, Smith, Holland, & McInnes, 2001; Rapport *et al.*, 1997). Despite this, neither of the two studies investigated their potential influence on practice effects on the MCST.

The new *Symbol Digit Test* is a test of speed and attention developed in the Neuropsychology department of the National Hospital for Neurology and Neurosurgery (see Methods section for a more detailed description). This test is similar to the *Symbol Digit Modalities Test* (SDMT; Smith, 1991) insofar as it assesses visual scanning, tracking and motoric speed. The SDMT has been used extensively in clinical practice and clinically oriented research (see Spreen & Strauss, 1998). It has been shown that the SDMT is sensitive to brain damage and has good test–retest reliability although it does show practice effects (see Spreen & Strauss, 1998). As far as we are aware, RC indices have not been documented for the SDMT.

The new Symbol Digit Test differs from the SDMT in that the test items are perceptually easier. Moreover, this test has fewer items than the SDMT and no time limit. Thus, participants always complete the test. We were interested in investigating the psychometric properties of our test and comparing them with the known test–retest reliability and practice effects of the SDMT.

As reported above, the majority of these cognitive tests have been used extensively in clinical practice and research. One of the main reasons for this is that is they all allow for a range of scores to be obtained. This is because they are either graded in difficulty (GNT, Silhouettes Test), or 'open ended' (verbal fluency, to a certain extent, MCST and the new Symbol Digit Test). This makes them suitable for monitoring changes in cognitive functioning over time as scores at repeat assessments can be compared with a baseline assessment. However, in most cases it has been tacitly assumed that performance on the tests is stable over time. This assumption has rarely been directly tested. Indeed, should this assumption prove to be unfounded then the tests would be of limited usage for monitoring change. In this paper, we explore whether performance on these tests is stable over time and whether age and/or performance on the National Adult Reading Test (Nelson & Willison, 1991) influence practice effects.

Method

Participants

Participants (188 healthy volunteers) were recruited through posters placed in the National Hospital as well as in local churches, community centres and at an engineering company. They were aged between 39 and 75 years, (mean age = 57.0, *SD* = 8.3) and had 13.1 years of education (*SD* = 3.7). None of the participants had a history of alcoholism, head injury, stroke or other neurological condition. There were 71 males and 117 females. Of this sample; 106 were administered the GNT, 99 were administered the Silhouettes Test, 99 were administered the two verbal fluency tests, 90 were administered the MCST and 188 were administered the new Symbol Digit Test.

Materials and procedures

Six tests, tapping nominal (GNT), visual perceptual (silhouettes), 'executive' (verbal fluency, MCST) and speed and attention (symbol digit) functions were used. In addition, the National Adult Reading Test (NART; Nelson & Willison, 1991) was administered to obtain an estimation of the IQ of our sample.

Graded Naming Test

The published version was used and administered using the instructions given in the test manual (GNT; McKenna & Warrington, 1983). The examiner recorded the total number of correct responses out of a possible 30.

Silhouettes subtest

The published version was used and administered using the instructions given in the test manual (from the VOSP; Warrington & James, 1991). The examiner recorded the total number of correct responses out of a possible 30.

Verbal fluency

Two tests of verbal fluency were administered to each subject; one of phonological fluency (words starting with the letter S) and the other of semantic fluency (animals). The tests were administered using instructions as described in Lezak (1995). In both tests, 1 minute was timed with a stopwatch from when the examiner said, 'Begin'. The responses were recorded by the examiner. The score is the number of words produced in 1 minute.

Modified Wisconsin Card Sorting Test

This test consisted of the 48 test cards and 4 stimuli cards as described by Nelson (1976). It was administered using standard instructions. The number of categories obtained, total errors made and perseverative errors made were all recorded.

New Symbol Digit Test

This new test consists of five rows of test items containing, in all, 50 blank squares, each paired with a randomly assigned abstract symbol. Above these rows is a printed key that pairs each abstract symbol with a single digit (1–9). There is also a short row of five practice items. Subjects are asked to fill in the blank squares with the number that is paired with the particular abstract symbol. After completing the practice items, subjects are asked to fill in all the remaining 50 squares. The time to complete the test is recorded by the examiner with a stopwatch, as is the number of errors made. Performance on the test is assessed by the time taken to complete the test and the number of errors made.

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 above tests were administered to the participants on two different occasions. There was a 1 month interval between the two assessments (M = 30.4 days, SD = 1.4). For

the participants who were administered all of the five tests at first assessment, the following presentation order was used: phonological fluency, MCST, semantic fluency, GNT, Silhouettes Test and the new Symbol Digit Test. This order ensured that participants were assessed on the verbal fluency tests before the naming tests, which might otherwise have aided their performance. For the subjects who were not assessed on all the tests, the same presentation order was employed with the omission of one or more tests. At Assessment 2, the same tests were administered in the same order and the NART was also administered at the end of the assessment.

Statistical analyses

The results of the above tests were analysed using the following statistical procedures. Kolmogorov-Smirnoff Z tests were used to assess whether the data were normally distributed. When the data were not normally distributed, non-parametric analyses were used.

Test–retest reliability was assessed using Pearson's correlations for parametric data and Spearman's rho (ρ) for non-parametric data. Practice effects were assessed using paired t-tests for parametric data and Wilcoxon tests for non-parametric data. RC indices corrected for practice were calculated when the magnitudes of changes in scores between the two assessments were normally distributed. Following the procedure of Chelune *et al.* (1993; see also Bird *et al.*, 2003; Temkin *et al.*, 1999), RC indices were calculated as the standard deviations of the difference between the scores at Assessment 1 and Assessment 2, multiplied by 1.645 (where $1.645 = Z_{0.95}$, from the normal distribution). These indices were then corrected for practice effects by adding the mean change in score from Assessment 1 to Assessment 2. Therefore, in our normal sample, 10% of the controls had a change in score that fell outside the RC indices corrected for practice. Correlations with age and NART IQ were calculated using Pearson's tests.

Results

NART

The NART-estimated IQ of our sample was 113.8 (SD = 11.0). All the analyses subsequently reported were also carried out for a subgroup of participants with lower IQs. This lower IQ subgroup had a mean NART IQ of 100 (SD = 9.4; N = 20). A larger subgroup was analysed for the new Digit Symbol Test, as more participants had carried out this test (mean NART IQ = 100, SD = 7.1; N = 61).

Correlation between IQ, age and performance at Assessment 1

Our data allowed us to investigate the influence of both NART estimated IQ and age on performance of these tests. Table 1 shows the results of our correlational analyses.

Despite the relatively high NART IQ of our sample, we documented significant correlations between the NART and performance measures for almost all the tests included in this study; the only exceptions being semantic fluency and the numbers of perseverative errors made on the MCST. In contrast, age only mediated performance on the Silhouettes Test and the Digit Symbol Test. In both these tests, age had an adverse effect on performance. It should be noted that our sample does not include either young (below 40 years) or very old (over 75 years) adults.

Table 1. Effects of NART IQ and age on performance at first assessment

Test	Correlation between NART IQ and performance at Assessment 1	Correlation between age and performance at Assessment 1
GNT	.62***	ns
Silhouettes	.30**	−.28**
Verbal fluency		
'S'	.27*	ns
'Animals'	ns	ns
MCST		
TE	−.24*	ns
PE[a]	ns	ns
TC[a]	.22*	ns
Digit Symbol	−.20**	.46***

*** = $p<.001$, ** = $p<.01$, * = $p<.05$, ns = non-significant.

Notes. TE = total errors, PE = perseverative errors, TC = total categories obtained.
[a] Data analysed non-parametrically.

Test–retest reliability

Main group

The test–retest reliabilities of the neuropsychological tests we investigated were almost all highly significant, the only exception being the number of categories obtained on the MCST. Reliabilities ranged from good to reasonable and to poor. These results are shown in Table 2.

This is the first study to document the test–retest reliabilities of both the GNT and the Silhouettes Test. Both of these tests were found to have good reliabilities. The

Table 2. Test–retest reliability, practice effects and RC indices corrected for practice

Test	Controls (*N*)	Test–retest reliability	Practice effects		RC indices corrected for practice	
			Assessment 1 Mean (*SD*)ψ	Assessment 2 Mean (*SD*)ψ	Lower	Upper
GNT	106	.92***	24.1 (3.7)	25.1 (3.8)***	−1.6	+3.5
Silhouettes	99	.88***	21.6 (4.1)	22.8 (4.1)***	−2.3	+4.5
Verbal fluency						
'S'	99	.63***	18.7 (5.2)	20.8 (5.6)***	−5.5	+9.8
'Animals'	99	.56***	23.4 (5.1)	24.7 (6.3)*	−7.6	+10.5
MCST						
TE	90	.34**	5 (0–22)	3 (0–19)***	−5.6	+10.1
PE[a]	90	.38**	1 (0–9)	0 (0–7)***	N/A	N/A
TC[a]	90	.16[(ns)]	6 (2–6)	6 (3–6)*	N/A	N/A
Digit Symbol	188	.82***	92.0 (23.7)	89.0 (19.5)**	−19.5	+25.5

*** = $p<.001$, ** = $p<.01$, * = $p<.05$, [(ns)] = non-significant.
Notes. SD = standard deviation, *N* = number of participants, TE = total errors, PE = perseverative errors, TC = total categories obtained, N/A = not available (data not normally distributed).
[a] Data analysed non-parametrically.
ψ = for the MCST, the median and range are given.

reliabilities of both verbal fluency tests (phonological and semantic) are reasonable. Using the Fisher transformation and subsequent z tests, we compared our Pearson's correlation coefficients with those reported in a previous study (Harrison *et al.*, 2000). This analysis revealed that our reliabilities are not significantly different from those reported previously. The MCST's test–retest reliability was low, whether considering TE, TPE or TC. These findings are in keeping with two previous studies using similar test–retest intervals but were slightly lower than those reported in a further study using a 1-year interval between assessments (de Zubicaray *et al.*, 1998; Lineweaver *et al.*, 1999; Wilson *et al.*, 1996). The new Symbol Digit Test has good test–retest reliability.

Lower-IQ subgroup

The test–retest reliabilities in the subgroup of participants with lower IQ were similar to those found in the main group (see Table 3). The exceptions to this finding were the verbal fluency tests, where the reliabilities did not reach significance.

Table 3. Test–retest reliability, practice effects and RC indices corrected for practice in the lower-IQ group

Test	Controls (*N*)	Test–retest reliability	Practice effects		RC indices corrected for practice	
			Assessment 1 Mean (*SD*)ψ	Assessment 2 Mean (*SD*)ψ	Lower	Upper
GNT	20	.96***	20.3 (5.3)	21.4 (5.8)***	−1.5	+3.7
Silhouettes	20	.84***	20.5 (4.1)	21.5 (4.0)	−2.7	+4.8
Verbal fluency						
'S'	20	.37(ns)	15.7 (3.6)	15.8 (3.7)	−6.8	+6.8
'Animals'	20	.36(ns)	22.3 (4.6)	23.1 (6.0)	−8.8	+10.8
MCST						
TE	20	.54**	6 (0–22)	3 (0–16)***	−5.7	+11.2
PE[a]	20	.45*	1 (0–8)	0 (0–7)	N/A	N/A
TC[a]	20	.37(ns)	6 (1–6)	6 (3–6)	N/A	N/A
Digit Symbol	61	.90***	95.4 (26.7)	92.8 (24.3)	−17.0	+22.3

*** = $p<.001$, ** = $p<.01$, * = $p<.05$, (ns) = non-significant.
Notes. *SD* = standard deviation, *N* = number of participants, TE = total errors, PE = perseverative errors, TC = total categories obtained, N/A = not available (data not normally distributed).
[a] Data analysed non-parametrically.
ψ= for the MCST, the median and range are given.

Practice effects

Main group

All the neuropsychological tests investigated in this study showed evidence of significant practice effects at Assessment 2 (see Table 2). However, despite the significant practice effects, the mean increases in performance on all tests are relatively small.

Lower IQ subgroup

Practice effects in the lower-IQ subgroup were almost all of similar magnitude to those found in the main group (see Table 3). However, some of these effects failed to reach

significance. The failure to reach significance is probably a consequence of the reduced number of participants in this subgroup.

Correlation between NART IQ, age and practice effects in the main group

We investigated whether there was a correlation between NART IQ and the magnitude of the observed practice effects in our whole sample. This analysis revealed no significant correlations between NART IQ and practice effects for any of the tests used in this study.

We used the same analysis to investigate whether age influenced practice effects. This analysis revealed no significant correlations between age and practice effects for any of the tests used in this study.

RC indices corrected for practice

Main group

The RC indices corrected for practice are shown in Table 2. These show the upper and lower boundaries beyond which a change in score is significant. For all tests, the normal variability between assessments, as indicated by the RC indices, were larger than the practice effects associated with the test.

The RC indices for both the GNT and the Silhouettes Test are small, as there was little variation in performance between assessments. This is very important information for the clinician, as it means that even quite small changes in scores can represent significant improvement or decline. The RC indices are rather large for the verbal fluency tests and for the Symbol Digit Test.

Lower-IQ subgroup

In the lower-IQ subgroup, the RC indices corrected for practice are very similar to those reported for the main group (see Table 3).

Discussion

This study documented psychometric properties of six neuropsychological tests tapping nominal (GNT), perceptual (Silhouettes Test), executive (two verbal fluency tests and the MCST test), and speed and attention functions (new Symbol Digit Test).

The GNT is a widely used stringent test of nominal functions. Remarkably, there have been no previous studies investigating its psychometric properties over repeated assessments. In the present investigation, we documented very good test–retest reliability for the GNT. The reliability was comparable to that of another commonly used test of nominal functions (the Boston Naming Test; Kaplan, Goodglass, & Weintraub, 1983; Spreen & Strauss, 1998). Significant though small practice effects reflecting mean gains in scores of approximately one word were documented. Neither NART IQ nor age appeared to mediate the practice effects we documented. In keeping with its high reliability, the RC indices are small, indicating that there is only limited variability in performance over time. These very good psychometric properties were documented not only in our whole sample but also the subset with lower NART IQs. Thus, the GNT is a reliable test for monitoring nominal functions over time, irrespective of an individual's premorbid level of ability.

There was a robust correlation between NART IQ and GNT scores. This was identical to the correlation previously reported in a healthy population with NART scores lower than our sample (Warrington, 1997). High correlations between the Boston Naming Test and reading ability have also been documented (Hawkins *et al.*, 1993). Thus, reading ability and nominal functions appear to be highly related. The strong correlation obtained between the GNT and the NART suggests that NART scores should be regarded as a relatively good predictor of GNT performance. Therefore, a discrepancy between performance on the two tests may be indicative of nominal deficits.

Age did not mediate performance on the GNT. A previous study of healthy adults aged 18–77 years documented a very small *positive* correlation between age and GNT scores ($r = .15$; Warrington, 1997). A subsequent study of healthy older adults (aged 64–81 years) found a very small negative correlation ($\rho = -.24$; Clegg & Warrington, 2000) Thus, our results taken together with these previous studies suggest that there is no appreciable age-related decline in performance on the GNT, at least in adults below 70 years. This finding is in accordance with studies of nominal ability using the Boston Naming Test that have documented a small age-related decline in performance only in adults over 70 years of age (see Spreen & Strauss, 1998).

Similarly to the GNT, no formal investigation of the psychometric properties of the Silhouettes Test had previously been carried out. We documented good test–retest reliability, small practice effects and the rather small RC indices corrected for practice. These findings occurred both in the main group and in the lower-IQ subgroup. Neither age nor IQ appeared to mediate practice effects on the Silhouettes Test. Therefore, the Silhouettes Test is a useful test for monitoring changes in visual perceptual functions over time. A significant correlation between performance on the Silhouettes Test and NART-estimated IQ was found, which has not been reported previously. Thus, it appears that IQ impacts even on what one might regard as a simple perceptual task with no obvious cultural bias. Since, there are relatively few well-standardized perceptual tests, it is unclear whether the correlation with the NART is specific to the Silhouettes Test or could more generally be found on all perceptual tasks. As in a previous study, we documented a slight negative effect of age on performance (Warrington & James, 1991). This could be accounted for by the finding that older subjects are poorer than younger subjects at recognizing non-canonical pictures (Dror & Kosslyn, 1998). Thus, tasks such as the Silhouettes Test are unsurprisingly susceptible to an age-related decline in performance.

We documented reasonable test–retest reliabilities for two tests of verbal fluency (phonological and semantic fluency). These reliabilities were slightly, although not significantly, lower than those previously reported in a younger population (Harrison *et al.*, 2000). There were small but significant practice effects on both the verbal fluency tests, representing increases of one or two words at the second assessment. However, neither NART IQ nor age appeared to mediate practice effects. The RC indices were large, due to the considerable variability in performance on both tests between assessments. Thus, only large changes in scores represent significant change. Considering the rather large RC indices we documented, the small practice effects at retest are unlikely to confound changes in scores on both fluency tests. Thus, our data are in accordance with Harrison *et al.*'s (2000) claim that 'improvement on these tasks is by no means inevitable' (p. 186). These findings are important, since verbal fluency tests are frequently used in repeated clinical assessments and in the monitoring of

cognitive profiles associated with dementia (e.g. Garrard *et al.*, 2001; Perry & Hodges, 2000).

We documented a small correlation between performance on the phonological verbal fluency test and NART IQ. This finding is in keeping with previous research (Crawford, Moore, & Cameron, 1992; Harrison *et al.*, 2000). However, there was no significant correlation between NART IQ and the semantic fluency test. It has been suggested that semantic fluency is easier than phonological fluency. This is because semantic fluency tests require the generation of items from one category, while phonological fluency requires the generation of items from many different categories (Rosen, 1980). Thus, semantic fluency may be a less intellectually demanding task and therefore be less likely to correlate with NART IQ than phonological fluency. Previous research has only documented small correlations between four subtests of the WAIS-R and semantic fluency (Harrison *et al.*, 2000). Age did not correlate with either fluency test, an effect that has been reported previously (Crawford *et al.*, 1992; Miller, 1984). Altogether, both these tests have psychometric properties that allow the monitoring of fairly large changes in performance.

We found poor reliabilities for the MCST. This finding replicates and extends previous studies which document comparable poor reliabilities in smaller samples of healthy adults over similar test–retest intervals (de Zubicaray *et al.*, 1998, Wilson *et al.*, 1996). However, slightly better test–retest reliabilities were reported over a 1-year interval (Lineweaver *et al.*, 1999). It has been argued that tests of 'executive' function requiring the discovery of novel solutions do not lend themselves to repeated administrations (Burgess, 1997; Lowe & Rabbitt, 1998; Wilson *et al.*, 2000). Thus, the MCST's low test–retest reliabilities are in accordance with researchers who consider the MCST to be a 'one-shot test' (Lezak, 1995). Significant practice effects were found on all performance measures. This indicates that participants do benefit from previous exposure to the test. However, inspection of the data revealed that improvement at retest is by no means guaranteed. Indeed, approximately 40% of participants made more errors at retest. This is a consequence of the poor reliability of the test obscuring the small practice effects. These small practice effects were not mediated by either NART IQ or age. We documented no effect of age, but a fairly small, though statistically significant, effect of NART IQ on performance on the MCST. This finding is in line with previous studies (de Zubicaray *et al.*, 1998). In the light of these findings, we conclude that the MCST is unsuited to monitoring changes in 'executive' function over time.

The new Symbol Digit Test has good test–retest reliability, which is very similar to the SDMT (Smith, 1991). We documented small, but significant, practice effects on the test, which represented a time saving of approximately 3 seconds. The RC indices were rather large. Therefore, the small practice effects are unlikely to confound the interpretation of changes in performance over time. Both NART IQ and age influenced performance on the Symbol Digit Test. However, as with the verbal fluency tests, neither of these factors appeared to mediate practice effects. Overall, these data suggest that the new Symbol Digit Test is a useful alternative to the SDMT, although it can detect only rather large changes in performance.

Conclusion

The psychometric properties of the GNT and the Silhouettes Test indicate that they are useful tools for monitoring even small changes in nominal and perceptual functioning.

The verbal fluency tests and the new Digit Symbol Test are suitable for monitoring only rather large changes in performance. The psychometric properties of MCST indicate that it is not a useful test to include in repeated assessments.

It has been argued that it may be the degree of change found in neurological patients which is particularly relevant when assessing certain patient populations (e.g. McCaffrey *et al.*, 2000). Indeed, a recent study has raised concern about whether 'norms for change' (particularly RC indices) obtained from healthy adults can be generalized to patients, particularly if impaired scores are recorded at initial assessments (Heaton *et al.*, 2001). Further research is needed to investigate whether the same psychometric properties we documented in our healthy sample can be generalized to various neurological popuations.

Acknowledgements

We thank Dr Hilary Watt for advice with the statistical analyses. We also particularly thank WS Atkins in Epsom for help in recruiting volunteers and all those who took part in this study. This research was funded in part by programme grant G9626876 from the Medical Research Council (UK).

References

Binetti, G., Cappa, S. F., Magni, E., Pardovani, A., Bianchetti, A., & Trabucchi, M. (1996). Disorders of visual and spatial perception in the early stages of Alzheimer's Disease. *Annals of the New York Academy of Science*, *17*, 221–225.

Bird, C. M., Papadopoulou, K., Ricciardelli, P., Rossor, M. N., & Cipolotti, L. (2003). Test–retest reliability, practice effects and reliable change indices for the recognition memory test. *British Journal of Clinical Psychology*, *42*, 407–425.

Bruggemans, E. F., van de Vijver, F. J., & Huysmans, H. A. (1997). Assessment of cognitive deterioration in individual patients following cardiac surgery: Correcting for measurement error and practice effects. *Journal of Clinical and Experimental Neuropsychology*, *19*, 543–559.

Burgess, P. W., (1997). Theory and methodology in executive function research. In P. Rabbit (Ed.), *Methodology of frontal and executive functions*. Hove: Erlbaum.

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.

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

Clegg, F., & Warrington, E. K., (2000). Psychometric testing of older adults: Provisional normative data for some commonly used tests. *Clinical Neuropsychological Assessment*, *1*, 22–37.

Crawford, J. R., Moore, J. W., & Cameron, I. M. (1992). Verbal fluency: A NART-based equation for the estimation of premorbid performance. *British Journal of Clinical Psychology*, *31*, 327–329.

de Zubicaray, G. I., Smith, G. A., Chalk, J. B., & Semple, J. (1998). The Modified Card Sorting Test: Test–retest stability and relationships with demographic variables in a healthy older adult sample. *British Journal of Clinical Psychology*, *37*, 457–466.

Dror, I. E., & Kosslyn, S. M. (1998). Age degradation in top-down processing: Identifying objects from canonical and noncanonical viewpoints. *Experimental Aging Research*, *24*, 203–216.

Garrard, P., Lambon Ralph, M. A., Watson, P. C., Powis, J., Patterson, K., & Hodges, J. R. (2001).

Longitudinal profiles of semantic impairment for living and nonliving concepts in dementia of Alzheimer's type. *Journal of Cognitive Neuroscience*, *13*, 892–909.

Gotham, A. M., Brown, R. G., & Marsden, C. D. (1988). 'Frontal' cognitive function in patients with Parkinson's disease 'on' and 'off' levodopa. *Brain*, *111*, 299–321.

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*, 181–191.

Hawkins, K. A., Sledge, W. H., Orleans, J. F., Quinland, D. M., Rakfeldt, J., & Hoffman, R. E. (1993). Normative implications of the relationship between reading vocabulary and Boston Naming Test performance. *Archives of Clinical Neuropsychology*, *8*, 525–537.

Heaton, R. (1989). *A manual for the Wisconsin Card Sorting Test*. Odessa, FL: Psychological Assessment Resources.

Heaton, R., Chelune, G. J., Talley, J., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card Sorting Test (WCST) manual: Revised and expanded*. Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Temkin, N., Dikmen, S., Avitable, N., Taylor, M. J., Marcotte, T. D., & Grant, I. (2001). Detecting change: A comparison of three neuropsychological methods, using normal and clinical samples. *Archives of Clinical Neuropsychology*, *16*, 75–91.

Holdstock, J. S., Mayes, A. R., Cezayirli, E., Isaac, C. L., Aggleton, J. P., & Roberts, N. (2000). A comparison of egocentric and allocentric spatial memory in a patient with selective hippocampal damage. *Neuropsychologia*, *38*, 410–425.

Horton, A. M. (1992). Neuropsychological practice effects × age: A brief note. *Perceptual and Motor Skills*, *75*, 257–258.

Joyce, E. M., & Robbins, T. W. (1991). Frontal lobe function in Korsakoff and non-Korsakoff alcoholics: Planning and spatial working memory. *Neuropsychologia*, *29*, 709–723.

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test*. Philadelphia: Lea and Feiberger.

Kapur, N., Ironside, J., Abbott, P., Warner, G., & Turner, A. (2001). A neuropsychological-neuropathological case study of variant Creutzfeldt-Jakob disease. *Neurocase*, *7*, 261–267.

Langdon, D. W., & Thompson, A. J. (1999). Multiple sclerosis: A preliminary study of selected variables affecting rehabilitation outcome. *Multiple Sclerosis*, *5*, 94–100.

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

Lineweaver, T. T., Bondi, M. W., Thomas, R. G., & Salmon, D. P. (1999). A normative study of Nelson's (1976). modified version of the Wisconsin Card Sorting Test in healthy older adults. *Clinical Neuropsychologist*, *13*, 328–347.

Lowe, C., & Rabbitt, P. (1998). Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: Theoretical and practical issues. *Neuropsychologia*, *36*, 915–923.

Mathias, J. L., & Coats, J. L. (1999). Emotional and cognitive sequelae to mild traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology*, *21*, 200–215.

Mathuranath, P. S., Nestor, P. J., Berrios, G. E., Rakowicz, W., & Hodges, J. R. (2000). A brief cognitive test battery to differentiate Alzheimer's disease and frontotemporal dementia. *Neurology*, *55*, 1613–1620.

McCaffrey, R. J., Duff, K., & Westervelt, H. J. (2000). Practitioner's guide to evaluating change with intellectual assessment instruments. New York: Plenum.

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

Miller, E. (1984). Verbal fluency as a function of a measure of verbal intelligence and in relation to different types of cerebral pathology. *British Journal of Clinical Psychology*, *23*, 53–57.

Monsch, A. U., Bondi, M. W., Butters, N., Salmon, D. P., Katzman, R., & Thal, L. J. (1992). Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. *Archives of Neurology*, *49*, 1253–1258.

Nelson, H. E. (1976). A modified card sorting test sensitive to frontal lobe defects. *Cortex*, *12*, 313–324.

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

Obonsawin, M. C., Crawford, J. R., Page, J., Chalmers, P., Low, G., & Marsh, P. (1999). Performance on the Modified Card Sorting Test by normal, healthy individuals: Relationship to general intellectual ability and demographic variables. *British Journal of Clinical Psychology*, *38*, 27–41.

Perry, R. J., & 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.

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.

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

Rosen, W. G. (1980). Verbal fluency in aging and dementia. *Journal of Clinical Neuropsychology*, *2*, 135–146.

Ross, S. J., & Hodges, J. R. (1997). Preservation of famous person knowledge in a patient with severe post-anoxic amnesia. *Cortex*, *33*, 733–742.

Smith, A. (1991). *Symbol Digit Modalities Test*. Los Angeles: Western Psychological Services.

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

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

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

Warrington, E. K., & James, M. (1991). *The Visual Object and Space Perception Battery*. Bury St Edmunds, UK: Thames Valley Test Company.

Wilson, B. A., Alderman, N., Burgess, P., Emslie, H., & Evans, J. J. (1996). *Behavioural assessment of the dysexecutive syndrome*. Bury St Edmunds, UK: Thames Valley Test Company.

Wilson, B. A., Watson, P. C., Baddeley, A. D., Emslie, H., & Evans, J. J. (2000). Improvement or simply practice? The effects of twenty repeated assessments on people with and without brain injury. *Journal of the International Neuropsychological Society*, *6*, 469–479.

Received 12 June 2002; revised version received 5 March 2003