Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals

Marina G. Falleti[1], Paul Maruff[1,2], Alexander Collie[2, 3], & David G. Darby[3]

From
1. School of Psychological Science
   La Trobe University, Bundoora
   Victoria, Australia

2. CogState Ltd. Melbourne
   Victoria, Australia

3. Centre for Neuroscience
   The University of Melbourne,
   Parkville, Victoria, Australia

Running head: Computerized assessment of cognitive change

All correspondence should be addressed to:
Marina G. Falleti
C/O 51 Leicester Street
Carlton South, Victoria, 3053, Australia.

Phone: 61 3 9349 1300
Fax: 61 3 9348 2689
Email: mfalleti@hotmail.com

Abstract

There are many situations in which cognitive tests need to be administered on more than two occasions and at very brief test-retest intervals to detect change in group performance. However, previous literature has not specifically addressed these important issues. The main aim of the current study was to examine these two factors by using a computerized cognitive battery designed specifically for the repeated assessment of cognition (i.e., CogState) in healthy young adult individuals. A further aim of the study was to examine how many times the battery needed to be completed before performance, as measured by the battery, stabilized. Forty-five adults (Age range: 18 - 40 years) completed the battery four times at ten-minute test-retest intervals, a fifth time at an interval of one week. The results illustrated that when brief test-retest intervals were used (i.e., 10 minutes), performance stabilized after the second assessment, as significant practice effects were generally observed between the first and the second assessments. Practice effects were also observed on some of the tasks at a one-week test-retest interval. Due to these findings, fifty-five adults (Age range: 18 - 40 years) completed the battery twice at ten-minute test-retest intervals (i.e., to eliminate the initial practice effect), and a third time at an interval of one month. No practice effects were observed. The implications of the results are discussed in terms of methods that can be adopted in order to minimize practice effects when this particular cognitive battery is used.

Keywords: Cognitive assessment, repeated measurements, practice effects, cognition, CogState.

Introduction

In many areas of cognitive assessment, there is a growing need for tests that can be administered to the same individual on more than one occasion. For example, it is common to investigate short-term cognitive change in individuals associated with illicit or licit drugs, surgery or disease. However, there are relatively few cognitive tests designed explicitly for the repeated assessment of cognitive function (Wilson, Watson, Baddeley, Emslie, & Evans, 2000). Furthermore, repeated assessment with the same test often leads to an improvement in performance (i.e., a practice effect; Catron, 1978; Temkin, Heaton, Grant, & Dikmen, 1999), which can obscure true change in central nervous system function (Chelune, Naugle, Lüders, Sedlak, & Awad, 1993). Therefore, the development of cognitive tests specifically for repeated assessment is warranted. Part of this development requires examination and characterization of the magnitude of any practice-related improvement that occurs on such tests.

One method for reducing the effects of practice has been to develop alternative forms or stimulus sets for neuropsychological tests (Benedict & Zgaljardic, 1998). However, where cognitive tests require individuals to learn a strategy or rule (e.g., the Winconsin Card Sorting Test and the Intra/Extradimensional Set Shifting Task from the Cambridge Neuropsychological Test Automated Battery), even alternative forms may not protect against practice-related performance improvement (Kay & Kane, 1991), as once the rules are learned on these tests, they are generally not forgotten (Basso, Bornstein, & Lang, 1999; Lowe & Rabbitt, 1998).

Another approach to countering the effects of practice on performance has been to develop an adjustment for repeated administration (Bruggemans, Van de Vijver, & Huysmans, 1997; Temkin et al., 1999). True change is inferred to have occurred if the observed change is greater than this adjustment. For example, Bruggemans and colleagues (1997) proposed a Reliability Stability Index that corrected for both the effects of repeated testing and the unreliability of test scores, and used this index to determine whether cognitive deterioration after cardiac surgery was observed. However, formal studies of the effects of repeated administration on cognitive performance have shown that the magnitude of practice effects can be related to the difficulty of the task itself (Basso et al., 1999), the length of the test-retest interval used (Benedict & Zgaljardic, 1998), the individual's age and general ability level at the time of testing (Dikmen, Heaton, Grant, & Temkin, 1999), and with the disorder or illness being assessed (Basso et al., 1999; Wilson et al., 2000). There is also the distinct possibility that these factors would interact to modify the magnitude of practice effects. Even without considering these interactions, such findings challenge the validity of general correction factors.

Recently, it has been argued that an alternative approach to assessing cognitive change is to develop tests specifically for the purpose of repeated assessment (Collie, Darby, Falleti, Silbert, & Maruff, 2002). Such tests should be standardized, relatively quick to administer, contain multiple alternative forms, yield data appropriate for prospective statistical designs, not require any strategy formation or rule learning, and show no improvement in performance with repeated assessment. CogState, a computerized cognitive battery, appears to be one such test (Westerman, Darby, Maruff, & Collie, 2001). CogState has been developed specifically for situations that require repeated testing of individuals and validly measures a wide range of cognitive functions (e.g., psychomotor speed, reaction time, working memory, divided attention, learning; Collie, Maruff, & Darby, 2003). CogState is currently being used in many areas of research, where it has been found to be sensitive in detecting mild cognitive impairment (Darby, Maruff, Collie, 2004; Maruff et al., 2004), concussion (Collie et al., 2004a; Collie et al., 2004b; Moriarity et al., 2004), cognitive decline following coronary artery bypass grafting (Silbert et al., 2004), attention deficit hyperactivity disorder (Mollica, Maruff, & Vance, 2004), and fatigue and alcohol (Falleti, Maruff, Collie, Darby, McStephen, 2003). There are other neuropsychological tests that are available, which emphasize repeatability and measure some of the same constructs as CogState (e.g., ImPACT, Iverson, Lovell, & Collins, 2003; and RBANS, Gold, Queern, Iannone, & Buchanan, 1999; Hobart, Goldberg, Bartko, & Gold, 1999); however, it is practice effects that are associated with repeated CogState testing that were the main focus of this paper, as this has not previously been investigated, particularly in a group of young healthy adults.

We have previously reported that a series of computerized tests of psychomotor function, attention, executive function and associative learning, developed to meet these criteria, showed stable performance in older people (Mean age = 63.38 ± 7.58) who were assessed four times in 2 hours (Collie et al., 2003). Once individuals became familiar with the requirements of the different tests, no practice effects occurred even though the very brief test-retest intervals were optimal for these to develop. As previous research has indicated that practice effects are different in younger adults than older adults (Horton, 1992; Mitrushina & Satz, 1991; Shatz, 1981), the current study aimed to extend this research by assessing healthy younger adults.

While previous research into the area of practice effects has been valuable, two important issues have not been addressed. The first issue concerns the length of the test-retest interval. While practice effects have been studied with various cognitive and psychological batteries, the test-retest intervals that have been used have ranged from one week (Basso, et al., 1999) to two years (McCaffrey & Westervelt, 1995), with shorter test-retest intervals (i.e., less than one week) rarely being examined. Currently, only one study has analyzed the effects of practice at shorter test-retest intervals (daily for 4 weeks, except weekends; Wilson et al., 2000), using mostly paper and pencil tests on people with and without brain injury.

The second important issue relates to the number of reassessments that are conducted. Where practice effects have been investigated, these have generally involved only a single reassessment (see Basso et al., 1999; Heaton et al., 2001; Salinsky, Storzbach, Dodril, & Binder, 2001). This is despite neuropsychologists or psychologists often being required to track recovery over time in individuals or groups using multiple repeated assessments [e.g., after concussion, Johnson, Hertel, Olmsted, Denegor, & Putukian, 2002; following coronary artery bypass grafting (CABG), Mullges, Berg, Schmidtke, Weinacker, & Toyka, 2000; evaluating acute toxic exposure; Weaver et al., 2002]. The investigation of practice effects beyond two repeat testing sessions will aid in the determination of whether it is possible to remove practice effects by allowing individuals to repeat the test until they are familiar with the requirements and their performance has stabilized (i.e., as recommended by Duff, Westervelt, McCaffrey, & Haase, 2001).

There are several benefits in investigating changes in performance over very short test-retest intervals. First, estimates of the magnitude of practice can be derived under conditions most optimal for improvement to occur. Second, the use of very short intervals minimizes the extent to which the individuals will undergo any physical or psychological changes that could give rise to true cognitive change (i.e., changes in sleep patterns, stage in menstrual cycle). Therefore, estimates of improvement in these conditions would be more likely to reflect only measurement-related factors and not include any effects of normal biological variability that are known to cause subtle changes in cognition and which operate over weeks or months (see Bland & Altman, 1996a). Demonstration of the stability or non-stability of performance at very short intervals will have limited application to clinical settings where decisions about cognitive change can be made on the basis of weekly (i.e., CABG, concussion, recovery from head injury) or monthly re-assessments (i.e., CABG, drug effects, dementia, anxiety). Consequently, the effects of practice at these intervals should also be known, as it is possible that performance will continue to improve as a consequence of familiarity with the test, irrespective of the time intervals between assessments.

The current study investigated the presence and magnitude of practice effects at very brief test-retest intervals (i.e., ten minutes and one-week) in a group of healthy young adults. Performance on the cognitive tests was then examined at a longer test-retest interval (i.e., one-month), in a second group of healthy adults.

Method

Participants

Two groups of individuals participated in this study, with the first group being tested at very brief test-retest intervals and the second group being tested at the longer test-retest interval. Group 1 consisted of sixty-three healthy individuals (35 females and 28 males) aged between 18 and 40 years (mean age: 21.64 ± 3.80) who initially agreed to participate in this study. However, seven individuals were excluded because they reported that they were recreational drug users, one was excluded due to a neurological illness and 10 did not complete all five assessments. Therefore, 45 individuals completed the experiment. Group 2 consisted of data obtained from fifty-five healthy individuals (9 females and 46 males) aged between 18 and 40 years (mean age: 32.69 ± 9.62). Even though participants were selected from the same age group, there was a significant difference in age (A Wilcoxon signed ranks test was conducted as there were unequal sample sizes in each group; $Z = -5.121$, $p < 0.05$). There was no significant difference in education levels, were most of the participants in both groups had a tertiary level of education. Participants were screened to exclude smokers, regular to heavy coffee drinkers, recreational drug or alcohol abusers, and those with neurological illnesses. All participants had normal IQ as determined by the National Adult Reading

Test (Group 1: Predicted Full Scale IQ = 113.29; Group 2: 117.12; Nelson & Wilison, 1991) and were not depressed or anxious according to their normal scores on the Center for Epidemiological Studies Depression Scale (Radloff, 1977) and the State-Trait Anxiety Scale (Spielberger, Gorsuch, & Lushene, 1970), respectively. Participants gave informed consent prior to the commencement of the study and were recruited through advertisements placed at La Trobe University. The University's ethics committee granted ethics approval prior to the commencement of the study.

Apparatus

The cognitive battery (CogState; see Westerman et al., 2001) was presented on an Apple Macintosh iMac computer complete with headphones (Computer Processor Information: PowerPC G3; Speed: 350 MHz; Operating System Version 9.2.2.; Apple Inc. Cupertino, CA; USA). All tasks within the battery were adaptations of standard neuropsychological and experimental psychological tests (see below). This battery required approximately 15-20 minutes to complete and consisted of 8 tasks in the form of card games that were presented in succession on a green background (see Figure 1). In order to aid individuals with the task, written instructions were presented to the left of the screen indicating the rule of that particular task. Participants were then given an interactive demonstration of the task, and once they demonstrated their awareness of the rules, the task began.

A grey keyboard resembling that of a computer keyboard appeared in the lower half of the computer screen and the cards associated with each task were presented in the upper half of the screen. Participants were required to respond with three keys: the 'd key, which indicated a 'left' response; the 'k' key, which indicated a 'right' response; and, the spacebar, which indicated the detection or monitoring of cards (see below in the description of each task within the battery). An error beep sounded when an individual pressed the wrong key, and the beginning of each new task was indicated with a shuffling of the cards. The dependent variables (DVs) recorded for each task included reaction times (RTs) and accuracy (i.e., the percentage of correct responses). The 8 tasks of the battery were as follows:

1. Simple reaction time (SRT).

   A card is presented face-down in the centre of the computer screen. When this card turns face-up, participants are required to press the spacebar as quickly as possible. There are 15 trials, and this task is repeated again after the Continuous monitoring task and after the Associative monitoring task.

2. Choice reaction time (ChRT).

   This task is similar to the SRT task; however, participants are required to indicate the color of the suit (i.e., black or red) with their response (by pressing the ‘d key or the ‘k’ key). There are 15 trials presented in this task.

3. Complex reaction time (CoRT).Two cards are presented simultaneously in the centre of the computer screen, one above the other. Participants are required to indicate whether the two cards match in color (i.e., are both red or both black) or do not match in color (i.e., one card is black and one card is red). Once again, participants indicate their response by pressing either the ‘k’ key or the ‘d key. There are 15 trials presented in this task.

4. Continuous monitoring task (Mon).

   Five cards are presented beside each other across the centre of the screen. Two horizontal lines are also presented on screen: one above the top edges of the cards and one below the bottom edges of the cards. Each card jiggles up and down at a random and independent speed and direction. Participants are required to monitor the movement of the cards, and press the spacebar when any one of the cards touches one of the horizontal lines. Responses that are made before cards touch the line are classified as errors. There are 15 trials presented in this task.

5. <u>One back working memory task (OBK).</u>

   A single card is presented in the centre of the screen and changes every 2 seconds. Each time the card changes, participants must indicate whether or not the new card is the same or different as the one that was just presented. Participants indicate their response by pressing either the ‘k’ key or the ‘d key. There are 15 trials presented in this task.

6. <u>Matching task (Mat).</u>

   In this task, five pairs of cards (i.e., two rows of five cards) are presented face-up in the top half of the computer screen. These cards remain visible for the duration of the task. Below these cards, another pair of cards is presented in the centre of the computer screen, one above the other. Participants are required to indicate with their response (i.e., the ‘k’ or ‘d keys) whether or not the two cards in this pair are exactly the same of any of the five pairs of cards. This pair of cards in the centre of the screen changes every two seconds. There are 20 trials presented in this task.

7. <u>Incidental learning task (IncL).</u>

   This task immediately follows, and is similar, to task 6. The difference is that the five pairs of cards from the Matching task turn face-down and participants are required to judge whether the pair of cards that is presented face-up in the centre of the computer screen are exactly the same as any of the five pairs of cards. There are 10 trials presented in this task.

8. <u>Associative learning task (AssL).</u>

   This task is similar to task 7. Five new pairs of cards are presented in the top half of the screen. A pair of cards is presented in the bottom half of the screen, and participants are to indicate whether this card matches any of the pairs above it.

Once a match has been made, the pair in the top half of the screen turns face-down and participants are required to remember it so that subsequent presentations of that pair must be judged by memory. Of the five pairs of cards, the only pair that does not turn face-down is the central pair, serving as a control to provide a measure of whether participants are actually trying to match the card pairs. This task consists of 20 trials where the same five pairs of cards are shown 4 times each (i.e., 20 repeated pairs), and 20 trials in which pairs of cards do not match (i.e., 20 never-repeated distractor pairs).

Procedure

Participants were seated at a computer and instructed to place their headphones on. They were informed that instructions for each task would appear to the left of the screen and that they were required to complete the battery as quickly and as accurately as possible. They were also informed that each time they made a mistake an error beep would sound into their headphones. At the beginning of each task in the battery, participants were given three to five trials in which to demonstrate their understanding of the rule of that particular task (i.e., as stated to the left of the screen in a caption; See Figure 1), and this led straight into the experimental trials. Participants in group 1 were required to complete the battery four times within 10 minutes of each other and again one week later. Since the analysis of this data indicated that practice effects mainly operated between the first and second assessments, participants in group 2 were required to complete the battery twice within 10 minutes of each other and again one-month later.

Data Analysis

All RTs were recorded in milliseconds (ms) and transformed to logarithmic 10 values in order to eliminate positive skew (Anastasi & Urbina, 1997). Inspection of the distributions for both RTs and accuracy (i.e., percent correct) indicated one individual who produced responses outside the normal range (Mean + 3SD) on most of the tasks (i.e., outlier); thus, the data from this individual was eliminated from the data analysis. Data analysis then proceeded in three steps.

Test-retest reliability

The test-retest reliability of each performance measure was compared between each assessment using intra-class correlations (Anastasi & Urbina, 1997), which were generated using Social Sciences and Statistical Package for Windows (Version 11). This is a more appropriate measure of association than simpler techniques (such as Pearson's Product Moment Correlation) as it assesses the degree of variation from each assessment (see Wilk et al., 2002). Since more assessments were obtained from Group 1, correlations were calculated for this group only.

Group 1 analysis

To determine the presence and magnitude of any practice effects across time, the speed and accuracy of performance on each task was submitted to a series of one-way repeated measures Analyses of Variance (ANOVA). Where ANOVA indicated a significant effect for assessment, post-hoc comparisons were used to compare group means between adjacent assessments. As this analysis indicated that practice effects occurred between the first and the second assessment for most tasks, performance at the one-week interval (i.e., fifth assessment) was compared to performance at the second assessment. Due to the number of statistical tests that were conducted on the data, the alpha level required for significance was set at $p = 0.01$, although all error rates are presented (Hinkle, Wiersma & Jurs, 1994; Note, there were 16 variables in total).

Since previous researchers have recommended that performance change be expressed as an estimate of effect size since this quantifies any differences between two measures of average performance relative to the variability associated with each (Cohen, 1988; Dunlap, Cortina, Vaslow, & Burke, 1996), Dunlap et al.'s (1996) effect size statistics (i.e., $t_c\ [2(1-r)/n]^{½}$) were calculated for performance change between assessments 1 and 2, and 2 and 5 (one week interval). Raw difference scores were also calculated to provide the amount of change in milliseconds for the speed of responses and in percentages for the accuracy of responses. Finally, the performance of participants at baseline was compared to the magnitude of practice effects using correlation statistics to investigate any moderating effects.

Group 2 analysis

The analysis of data from these three assessments followed the same steps as above; however, effect size estimates and raw difference scores were derived for the one-month test-retest interval only.

Results

Test-retest reliability

Table 1 shows the intraclass correlations between assessments for each performance measure derived from data obtained by Group 1. The majority of the intra-class correlation coefficients were greater than 0.60 and remained consistent across the five assessments. Reliability coefficients were lowest for the speed of responses on the Incidental Learning task between assessments 2 and 3 (i.e., 0.35), and for the accuracy of responses on the Monitoring task between assessments 1 and 2 (i.e., 0.00), and the Simple Reaction Time task between assessments 4 and 5 (i.e., -0.06).

[Table 1 about here]

Group 1 data

The group means and standard deviations for each performance measure at each assessment are shown in Table 2. The results of the ANOVA and post-hoc tests are summarized in Tables 2 and 3. Significant practice effects were observed for nine of the sixteen measures (Table 2). Post-hoc analyses indicated that performance speed improved from the first to the second assessment for the one-back memory and matching tasks (see Table 3). Performance accuracy improved over the same interval for the simple reaction time and incidental learning tasks.

[Tables 2 and 3 here]

At very short assessment intervals (i.e., assessments 2 to 4), no further improvement in performance was observed. However, examination of practice effects at the one week re-retest interval indicates that improvement occurred from the second to the fifth assessment for the speed of performance on the associate learning task and the accuracy of performance on the one-back memory task (see Table 2).

The magnitude of the practice effects as well as the raw difference scores from assessments 1 to 2 (i.e., ten-minute interval), and 2 to 5 (one-week interval) can be seen in Table 3. Magnitudes of practice (i.e., negative effect sizes for the speed of responses and positive effect sizes for the accuracy of responses) ranged from small to moderate, and on tasks where performance was found to be significantly different, magnitudes were moderate in size. Raw difference scores indicate that where there is a significant difference in performance for the speed of responses, this corresponds to a difference of more than 100ms. For the accuracy of responses, this difference is about 10%. Finally, there were no significant correlations between the baseline performance and the improvement in performance from the baseline to the 10 minute interval (or the one-month interval) for any of the measures (correlations ranged from -0.42 to 0.37).

Group 2 data

The group means and standard deviations for each performance measure at each assessment are shown in Table 4. The results of the ANOVA illustrate significant practice effects for six of the sixteen measures (Table 4). Post-hoc analyses indicated no significant improvements in performance from the second to the third assessment (i.e., one-month interval; see Table 5).

[Tables 4 and 5 here]

The magnitude of the practice effects as well as the raw difference scores from assessments 2 to 3 can be seen in Table 5. Magnitudes of practice (i.e., negative effect sizes for the speed of responses and positive effect sizes for the accuracy of responses) were all small. Raw difference scores indicate that change in the speed of performance was under 90ms (except for the matching task) and change in the accuracy of responses was under 15%.

Discussion

The current study illustrated that performance generally improved from the first to the second assessment on the CogState battery. After the second assessment, the performance of the group stabilized and improved no further on any of the cognitive measures. Importantly, test performance did not become worse over the first four assessments, which would have occurred had individuals become fatigued or lost motivation as a consequence of the repeated assessment. When the time between tests was increased to one week, significant practice effects were evident on only two of the measures (i.e., one-back memory accuracy, and speed of associate learning; see Table 3), with moderate improvements also being observed for the accuracy of performance on both the incidental and associate learning tasks. When the test-retest interval increased to one month, no significant practice effects were observed and the amount of change on all of the measures was small in magnitude.

The presence of practice effects

The current study illustrated that practice effects were associated with repeated performance on the CogState battery, although these occurred mainly between the first and second assessments, conducted ten minutes apart. Specifically, moderate improvements were observed for speed of performance on the one-back memory and matching tasks and the accuracy of performance on the simple reaction time and incidental learning tasks (see Table 3). Moderate improvements in performance were also evident in the speed of performance on the incidental learning task and the accuracy of performance on the choice reaction time, one-back learning, and matching tasks although none of these reached statistical significance at the corrected level ($p < 0.01$). In considering the clinical relevance of these improvements, one needs to account for the number of trials in each particular measure. For example, on the matching measure, there was a 5% improvement in the accuracy of performance from the first to the second assessment. When considered against the number of trials in this task (i.e., 20), an improvement of 5% equates to an improvement in accuracy of only one item. As improvement on one trial is not usually indicative of clinically meaningful change, then the effect size index obtained for this measure (i.e., 0.44) does not really represent a moderate improvement in performance (as would be interpreted had only the effect size index been used).

The finding of significant practice effects on some of the tasks are consistent with previous research that has shown the cognitive abilities of detection, memory, matching and learning to be affected by practice when repeated assessments are conducted (Benedict & Zgarljardic, 1998; Feinstein, 1994; McCaffrey, Ortega, & Haase, 1993; Versavel et al., 1997). They are also consistent with observations that practice effects occur for most neuropsychological tests when they are given twice (see McCaffrey et al., 1993; McCaffrey, Westervelt, & Haase, 2001); however, in these previous studies, the intervals between each assessment have varied from one week to two years. When considered with the current results, it suggests that perhaps the presence of practice effects may be dependent upon the number of times an individual performs any particular test battery.

From these results, it appears that the practice effects observed between the first two assessments reflect the extent to which naïve individuals were able to acquire, understand and adhere to the requirements of the different tests rather than reflecting any improvement in the cognitive functions measured. This hypothesis needs to be investigated further before any definite conclusions can be made. If performance on some of tasks in the battery by naïve individuals tested once is not a true reflection of their cognitive abilities, then individuals must be assessed twice before any valid brain-behavior inference can be drawn.

Even though performance on the tasks from the CogState battery stabilized after the second assessment when given over four very short re-test intervals, other practice effects became evident at the one week re-test interval, when compared to the second assessment conducted in the first testing session. For example, a significant and moderate improvement was observed in the accuracy of performance on the one-back memory task (although this was equivalent to obtaining only one more correct answer than in their previous performance) and the speed of performance on the associate learning task (although this was less than 100ms different than in their previous performance). Therefore, these practice effects only reflect small improvements in performance on these particular tasks. Overall, these results are generally consistent with Benedict and Zgaljardic (1998), who state that the magnitude of practice effects usually decreases as the length of time between tests is increased. Furthermore, the results also suggest that practice effects will be observed in fewer tasks as the time between the test-retest interval increases.

The absence of practice effects

Importantly, the current results demonstrated no practice effects for any cognitive measure from the second assessment to the fourth, despite the time difference from the beginning of the second and the end of the fourth assessments being approximately one and a half hours. It is possible that the observation of no further improvement from the second assessment onwards (in the current study) could have been due to fatigue effects. That is, practice effects could have occurred after the second assessment, but these could have been cancelled out by factors such as fatigue. Future studies will need to be conducted to eliminate this potentially confounding factor where, for example, fatigue levels of participants could be subjectively measured and analyzed to see if this is affecting their performance.

Finally, no practice effects were observed at the one-month test-retest interval. Once the initial practice effect was minimized (i.e., requiring the individual to complete the test twice and comparing subsequent performance to the second assessment), then no further improvements in performance occurs. This suggests that in designs where individuals need to be tested one-month apart, performance at this interval should reflect actual levels of cognitive ability.

Reliability

The current results illustrated that the reliability for the speed of performance on all tasks was generally high (see Table 1). For the accuracy of performance, reliability levels ranged from moderate to high on all tasks except for monitoring and simple reaction time (final two assessments) tasks. The low to moderate reliability values between tests on some of the tasks could have been affected by a number of factors, such as floor or ceiling effects (Bland & Altman, 1996b; e.g., simple reaction time).

Upon close inspection, it can be seen that for half of the performance measures, the reliability was lower at the second assessment than the first. The lower test-retest reliability observed between baseline and the first re-assessment is likely to result in any extreme performance (i.e., very high or low scores) at baseline to regress to the mean of the sample on subsequent assessments (i.e., regression to the mean; McCaffrey & Westervelt, 1995). Such a phenomenon could obfuscate inferences by making it appear that individuals with very low or very high levels of test performance are more likely to show a significant change in cognition after the intervention than are individuals with average test performance. The higher test-retest reliability of the second and third tests found in the current and also in other studies that have repeated neuropsychological assessments (Duff et al., 2001; Mitrushina & Satz, 1991) suggests that use of a dual baseline, in which the first assessment with the CogState battery serves as a practice and is then excluded from analyses, will maximize test-retest reliability and therefore minimize the possibility of regression to the mean on third and fourth assessments.

Implications

There is a growing awareness within the field of neuropsychology of the importance of repeated assessment for monitoring outcome and progression of disease and for determining the efficacy of pharmacologic or psychological interventions. As conventional neuropsychological assessment is designed to provide a detailed description of patients’ cognitive weaknesses and strengths, it is generally time and resource consuming. Furthermore, because most neuropsychological assessments are based on single test administrations there has been no need to consider how repeated administration can affect clinical decision making. In a number of recent fields there has been a growing need for decisions about the presence of absence of cognitive change rather than cognitive impairment (e.g. concussion management, post-operative cognitive decline) and this increased research into the practical, methodological and statistical issues that underlie repeated assessment. The results of the current study can therefore provide a foundation for further studies in the area. For example, they indicate that repeated assessment can be considered over more than two occasions and at different re-test intervals. They also indicate that different performance measures (e.g. speed rather than accuracy) may be better suited to guide decisions about change than others and that the effects of practice are different for different measures. The current results also illustrate that sensitivity to true cognitive change begins with demonstration of stability in performance in individuals with healthy and unchanging central nervous systems. These aspects of performance as well as the experimental design can be used to assess other novel and existing tests for their potential use in the assessment of cognitive change.

The results have implications for methodological designs of studies that investigate performance change with CogState due to variables such as medical conditions, drugs or treatment programs that act quickly on the central nervous system (e.g., concussion, alcohol, electroconvulsive therapy). The results show the presence of initial practice effects, implying these need to be minimized in order for the appropriate analysis of data. This could be achieved through a dual baseline approach, whereby the individual is assessed twice, with the exclusion of performance on the first test, and the comparison of performance on the second test with any further assessments that are conducted. This is consistent with the recommendations made in a previous study of a single neuropsychological test (the California Verbal Learning Test, CVLT; Duff et al., 2001). Even though this approach (i.e., dual baseline) is optimal for testing healthy adult individuals, it can provide a foundation with which repeated assessments can occur within other groups of individuals (e.g., children, clinical populations). It is evident from the current study that at least 2 assessments need to be conducted before performance stabilized; however, whether this occurs in the specific group that is tested may need further consideration.

The results also highlight that when performance at longer test-retest intervals is assessed (i.e., one-week), practice effects are evident for some of the tasks in the CogState battery. These practice effects did not occur on simple tasks, but rather tasks that were most difficult to perform (e.g., associate learning). One way in which these practice effects can be controlled for methodologically is for studies that are interested in repeatedly assessing cognitive performance, further baseline assessments are taken after the individual is influenced by a particular variable (e.g., alcohol, fatigue; ABA design, or AABA if a dual baseline is used). This would then allow for any differences in baseline performance to be accounted for by the proper application of statistical correction indices (e.g., RSI; Bruggemans et al., 1997).

Finally, the results have implications for the statistical techniques that are used for interpreting performance change and practice effects in groups of individuals when repeated assessments are conducted. Interestingly, while significant practice effects were found for some tasks (being mirrored by moderate magnitudes of change), other tasks where performance also moderately improved (as indicated by the effect size index) did not indicate significant improvements. As it is commonly known, a non-significant difference does not necessarily indicate a change that is not meaningful. Therefore, it is useful in future studies that the magnitude of changes in the form of effect sizes are also calculated to give an indication of the size of improvement that has occurred, rather than just relying on statistical significance tests.

References

Anastasi, A., & Urbina, S. (1997). Psychological Testing. Upper Saddle River, N. J.: Prentice Hall.

Basso, M. R., Bornstein, R. A., & Lang, J. M. (1999). Practice effects of commonly used measures of executive function across twelve months. The Clinical Neuropsychologist, 13, 283-292.

Benedict, R. H. B., & Zgaljardic, D. J. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. Journal of Clinical and Experimental Neuropsychology, 20, 339-353.

Bland, J.M., & Altman, D.G. (1996a). Measurement error. British Medical Journal, 313, 744.

Bland, J.M., & Altman, D.G. (1996b). Measurement error and correlation coefficients. British Medical Journal, 313, 41-42.

Bruggemans, E. F., Van de Vijver, F. J. R., & 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.

Catron, D. W. (1978). Immediate test-retest changes in WAIS scores among college males. Psychological Reports, 43, 279-290.

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.

Cohen, J. (1988). Statistical power for the behavioural sciences (2nd ed.). Hilsdale, NJ: Lawrence Erlbaum.

Collie, A., Darby, D., Falleti, M. G., Silbert, B., & Maruff, P. (2002). Determining the extent of cognitive change following coronary surgery: A review of statistical procedures. Annals of Thoracic Surgery, 73,2005-2011.

Collie, A., Maruff, P., & Darby, D. G. (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. Journal of the International Neuropsychological Society, 9, 419-428.

Collie, A., Maruff, P., Makdissi, M., McCrory, P., McStephen, M., Darby, D. G. (2004a). CogSport: Reliability and correlation with conventional cognitive tests used in post-concussion medical examinations. Clinical Journal of Sports Medicine, 13, 28-32.

Collie, A., Maruff, P., Makdissi, M., McStephen, M., Darby, D. G., McCrory, P. (2004a). Statistical procedures for determining the extent of cognitive change following concussion. British Journal of Sports Medicine, 28, 273-278.

Darby, D., Maruff, P., Makdissi, M., McStephen, M., Darby, D. G., & McCrory, P. (2004). Statistical procedures for determining the extent of cognitive change following concussion. British Journal of Sports Medicine, 38, 273-278.

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.

Duff, K., Westervelt, H. J., McCaffrey, R. J., & Haase, R. F. (2001). Practice effects, test-retest stability, and dual baseline assessments with the California Verbal Learning Test in an HIV sample. Archives of Clinical Neuropsychology, 16, 461-476.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. Psychological Methods, 1, 170-177.

Falleti, M. G., Maruff, P., Collie, A., Darby, D. G., McStephen, M. (2003). Qualitative similarities in cognitive impairment associated with 24 h of sustained wakefulness and a blood alcohol concentration of 0.05%. Journal of Sleep Research, 12, 265-274.

Feinstein, A., Brown, R., & Ron, M. (1994). Effects of practice of serial tests of attention in healthy subjects. Journal of Clinical and Experimental Neuropsychology, 16, 436-447.

Gold, J. M., Queern, C., Iannone, V. N., & Buchanan, R. W. (1999). Repeatable battery for the assessment of neuropsychological status as a screening test in schizophrenia, I: Sensitivity, reliability, and validity. American Journal of Psychiatry, 156, 1944-1950.

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.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1994). Applied statistics for the behavioural sciences (3rd ed.). Boston: Houghton Mifflin.

Hobart, M. P., Goldberg, R., Bartko, J. J., & Gold, J. M. (1999). Repeatable battery for the assessment of neuropsychological status as a screening test in schizophrenia, II: Convergent/discriminant validity and diagnostic group comparisons. American Journal of Psychiatry, 156, 1951-1957.

Horton Jr, A. M. (1992). Neuropsychological practice effects % age: A brief note. Perceptual and Motor Skills, 75, 257-258.

Iverson, G. L., Lovell, M. R., & Collins, M. W. (2003). Interpreting change on ImPACT following sport concussion. The Clinical Neuropsychologist, 17, 460-467.

Johnson, P. D., Hertel, J., Olmsted, L. C., Denegar, C. R., & Putukian, M. (2002). Effect of mild brain injury on an instrumented agility task. Clinical Journal of Sport Medicine, 35, 297-302.

Kay, G. & Kane, R. L. (1991). Repeated measures in neuropsychology: Use of serial testing to measure changes in cognitive functioning. Journal of Clinical and Experimental Neuropsychology, 13, 49.

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

Maruff, P., Collie, A., Darby, D., Weaver-Cargin, J., Masters, C., & Currie, J. (2004). Subtle memory decline over 12 months in mild cognitive impairment. Dementia and Geriatric Cognitive Disorders, 18, 342-348.

McCaffrey, R. J., Ortega, A., Orsillo, S. M., & Haase, R. F. (1993). Effects of repeated neuropsychological assessments. Archives of Clinical Neuropsychology, 8, 519-524.

McCaffrey, R. J., & Westervelt, H. J. (1995). Issues associated with repeated neuropsychological assessments. Neuropsychology Review, 5, 203-221.

McCaffrey, R. J., Westervelt, H. J., & Haase, R. F. (2001). Serial neuropsychological assessment with the National Institute of Mental Health (NIMH) AIDS Abbreviated Neuropsychological Battery. Archives of Clinical Neuropsychology, 16, 9-18.

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

Mollica, C. M., Maruff, P., & Vance, A. (2004). Development of a statistical approach to classifying treatment response in individual children with ADHD. Human Neuropsychopharmacology, 19, 445-456.

Moriarity, J., Collie, A., Olson, D., Buchanan, J., Leary, P., McStephen, M., & McCrory, P. (2004). A prospective controlled study of cognitive function during an amateur boxing tournament. Neurology, 62, 1497-1502.

Mullges, W., Berg, D., Schmidtke, A., Weinacker, B., & Toyka, K.V. (2000). Early natural course of transient encephalopathy after coronary artery bypass grafting. Critical Care Medicine, 28, 1808-1811.

Nelson, H. E., & Willinson, J. R. (1991). The revised National Adult Reading Test – Test manual. Windsor: NFER-Nelson.

Radloff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. Applied Psychological Measurement, 1, 89-99.

Salinsky, M.C., Storzbach, D., Dodrill, C.B., Binder, L.M. (2001). Test-retest bias, reliability, and regression equations for neuropsychological measures repeated over a 12-16-week period. Journal of the International Neuropsychological Society, 7, 597-605.

Shatz, M. W. (1981). WAIS practice effects in clinical neuropsychology. Journal of Clinical Neuropsychology, 3, 171-179.

Silbert, B., Maruff, P., Evered, L., Scott, D., Kalpokas, M., Martin, K. J., Lewis, M., & Myles, P. (2004). Detection of cognitive decline after coronary surgery: a comparison of computerized and conventional tests. British Journal of Anaesthesia, 92, 814-820.

Spielberger, C., Gorsuch, R., & Lushene, R. (1970). Manual for the state-trait anxiety inventory. Palo Alto, CA: Consulting Psychologists 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.

Weaver, L. K. Hopkins, R. O., Chan, K. J., Churchill, S., Elliott, C. G., Clemmer, T. P., Orme, J. F., Thomas, F. O., & Morris, A. H. Hyperbaric oxygen for acute carbon monoxide poisoning. New England Journal of Medicine, 347, 1057-1067.

Westerman, R., Darby, D. G., Maruff, P., Collie, A. (2001). Computer-assisted cognitive function assessment in pilots: How and why? Australian Defence Force Health Journal, 2, 29-36.

Wilk, C. M., Gold, J. M., Bartko, J. J., Dickerson, F., Fenton, W. S., Knable, M., Randolph, C., & Buchanan, R. W. (2002). Test-retest stability of the repeatable battery for the assessment of neuropsychological status in schizophrenia. The American Journal of Psychiatry, 159, 838-844.

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

Versavel, M., van Laack, D., Evertz, C., Unger, S., Meier, F., & Kuhlmann, J. (1997). Test-retest reliability and influence of practice effects on performance in a multi-user computerized psychometric test system for use in clinical pharmacological studies. Drug Research, 47, 781-786.

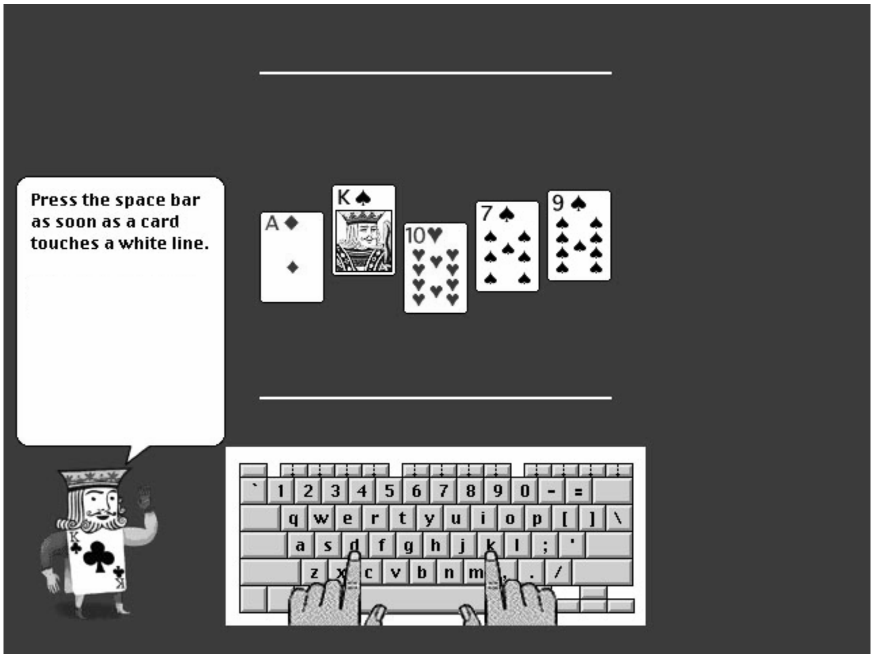

Figure 1. A representation of the continuous monitoring task.

Table 1
Intraclass Correlations of the tasks within the cognitive battery.

Task	Tests 1 to 2	Tests 2 to 3	Tests 3 to 4	Tests 4 to 5
SRT				
RT	0.84*	0.83*	0.94*	0.73*
Acc	0.23	0.39	0.45	-0.06
CHRT				
RT	0.38	0.55*	0.81*	0.71*
Acc	0.32	0.30	0.58*	0.41
CoRT				
RT	0.41	0.65*	0.66*	0.64*
Acc	0.46	0.55*	0.42	0.56*
Mon				
RT	0.54*	0.68*	0.74*	0.51*
Acc	0.00	N/A	N/A	N/A
OBK				
RT	0.62*	0.82*	0.74*	0.70*
Acc	0.25	0.45	0.42	0.47*
Mat				
RT	0.73*	0.85*	0.85*	0.71*
Acc	0.32	0.19	0.34	0.66*
IncL				
RT	0.42	0.35	0.74*	0.60*
Acc	0.51	0.47	0.54*	0.56*
AssL				
RT	0.86*	0.82*	0.77*	0.82*
Acc	0.60*	0.51*	0.46	0.66*

Note: SRT = Simple reaction time; CHRT = Choice reaction time; CoRT = Complex reaction time; Mon = Continuous monitoring; OBK = One-back monitoring; Mat = Matching; IncL = Incidental Learning; AssL = Associative Learning; RT = Reaction time in milliseconds; Acc = Accuracy (i.e., percent correct). * = significant at $p < 0.01$.

Table 2

Group Means (log 10) and Standard Deviations of each task on each testing occasion for Group 1.

Task	**Test 1**	**Test 2**	**Test 3**	**Test 4**	**Test 5**	***F***	***p***
SRT							
RT	2.48 (0.06)	2.48 (0.07)	2.48 (0.07)	2.47 (0.07)	2.47 (0.07)	0.95	0.42
Acc	99.36 (1.41)	99.90 (0.44)	99.86 (0.54)	99.90 (0.47)	99.95 (0.31)	5.11	**0.01**
CHRT							
RT	2.68 (0.06)	2.69 (0.07)	2.70 (0.10)	2.70 (0.08)	2.68 (0.09)	1.52	0.20
Acc	94.56 (6.08)	96.29 (5.12)	95.53 (5.43)	96.37 (5.48)	96.17 (4.86)	1.10	0.36
CoRT							
RT	2.83 (0.09)	2.83 (0.10)	2.83 (0.09)	2.81 (0.10)	2.81 (0.08)	1.41	0.24
Acc	90.73 (10.26)	93.21 (7.05)	94.64 (7.90)	95.71 (5.74)	94.16 (6.23)	3.68	**0.01**
Mon							
RT	2.56 (0.09)	2.58 (0.11)	2.59 (0.09)	2.56 (0.09)	2.59 (0.08)	0.85	0.36
Acc	99.79 (1.37)	100 (0.00)	100 (0.00)	100 (0.00)	100 (0.00)	1.00	0.32
OBK							
RT	2.89 (0.11)	2.83 (0.11)	2.81 (0.09)	2.80 (0.08)	2.81 (0.09)	13.33	**0.00**
Acc	86.66 (10.85)	90.98 (7.54)	91.60 (7.40)	94.42 (6.16)	94.24 (6.39)	9.04	**0.00**
Mat							
RT	3.19 (0.10)	3.13 (0.09)	3.12 (0.10)	3.11 (0.10)	3.11 (0.11)	13.79	**0.00**
Acc	86.69 (11.67)	91.24 (8.89)	94.63 (5.61)	92.65 (6.12)	93.13 (6.60)	7.47	**0.00**
IncL							
RT	3.05 (0.16)	3.00 (0.12)	3.00 (0.14)	3.02 (0.17)	3.01 (0.12)	1.21	0.31
Acc	64.26 (18.19)	74.09 (21.05)	75.94 (17.28)	77.40 (16.04)	80.11 (17.63)	7.01	**0.00**
AssL							
RT	3.10 (0.09)	3.08 (0.09)	3.06 (0.07)	3.05 (0.09)	3.05 (0.08)	10.27	**0.00**
Acc	77.62 (9.45)	80.45 (9.51)	80.13 (10.35)	83.39 (8.52)	83.50 (9.33)	4.53	**0.00**

Note: SRT = Simple reaction time; CHRT = Choice reaction time; CoRT = Complex reaction time; Mon = Continuous monitoring; OBK = One-back monitoring; Mat = Matching; IncL = Incidental Learning; AssL = Associative Learning; RT = Reaction time; Acc = Accuracy (i.e., percent correct).

Table 3

Performance change in each task for Group 1.

	Ten-minute interval				**One week interval**			
	F	***p***	**d**	**Raw diff**	***F***	***p***	**d**	**Raw diff**
SRT								
RT	0.73	0.40	-0.09	-4.37 ms	0.63	0.43	-0.13	-6.05 ms
Acc	6.90	**0.01**	0.49	0.54%	0.36	0.55	0.13	0.05%
CHRT								
RT	0.11	0.74	0.17	12.18 ms	0.18	0.67	-0.07	-6.51 ms
Acc	2.54	0.11	0.31	1.73%	0.01	0.91	-0.03	-0.12%
CoRT								
RT	0.01	0.93	-0.001	-1.56 ms	3.40	0.07	-0.28	-37.63 ms
Acc	2.46	0.12	0.28	2.48%	0.50	0.48	0.14	0.95%
Mon								
RT	0.55	0.46	0.02	14.20 ms	0.21	0.64	0.08	6.41 ms
Acc	1.00	0.32	-	0.21%	-	-	-	-
OBK								
RT	11.55	**0.00**	-0.54	-100.66 ms	2.23	0.14	-0.22	-35.11 ms
Acc	5.22	0.03	0.46	4.32%	6.81	**0.01**	0.47	3.26%
Mat								
RT	19.57	**0.00**	-0.61	-198.75 ms	2.14	0.15	-0.25	-76.64 ms
Acc	4.99	0.03	0.44	4.55%	1.74	0.19	0.24	1.89%
IncL								
RT	2.37	0.13	-0.36	-122.94 ms	0.36	0.55	0.12	32.77 ms
Acc	8.36	**0.00**	0.50	9.83%	2.58	0.12	0.31	6.02%
AssL								
RT	1.32	0.25	-0.26	-65.73 ms	9.93	**0.00**	-0.36	-84.05 ms
Acc	3.33	0.07	0.30	2.83%	3.57	0.07	0.32	3.05%

Note: The 10 minute interval was from Assessment 1 to 2; and the one week interval was from assessment 2 to 5. The d statistic indicates the effect size index; Raw diff indicates the raw difference values (e.g., raw speed of response on SRT at assessment 2 minus assessment 1). For the speed of responses, negative effect size indices and raw difference scores indicates improvement in performance compared to the previous assessment. For the accuracy of responses, positive effect sizes indicates improvement in performance compares to the previous assessment.

Table 4

Group Means (log 10) and Standard Deviations of each task on each testing occasion for Group 2.

Task	Test 1	Test 2	Test 3	*F*	*p*
SRT					
RT	2.52 (0.10)	2.50 (0.09)	2.51 (0.09)	1.85	0.17
Acc	97.54 (5.92)	99.49 (1.48)	99.88 (0.67)	6.82	**0.01**
CHRT					
RT	2.71 (0.11)	2.71 (0.08)	2.73 (0.09)	1.01	0.36
Acc	91.34 (10.09)	94.00 (6.31)	95.27 (6.89)	4.15	0.02
CoRT					
RT	2.85 (0.08)	2.84 (0.08)	2.84 (0.09)	1.01	0.36
Acc	88.69 (12.32)	93.81 (6.07)	94.32 (8.58)	7.84	**0.00**
Mon					
RT	2.54 (0.16)	2.53 (0.13)	2.53 (0.11)	0.55	0.58
Acc	99.64 (2.70)	100.00 (0.00)	100.00 (0.00)	1.00	0.32
OBK					
RT	2.90 (0.14)	2.87 (0.12)	2.85 (0.11)	7.36	**0.00**
Acc	81.94 (13.09)	89.51 (10.56)	91.01 (9.46)	21.62	**0.00**
Mat					
RT	3.22 (0.11)	3.19 (0.12)	3.17 (0.10)	11.32	**0.00**
Acc	72.05 (24.47)	83.29 (15.58)	86.41 (11.61)	20.56	**0.00**
IncL					
RT	3.06 (0.12)	3.04 (0.11)	3.03 (0.14)	1.56	0.22
Acc	68.13 (17.82)	69.65 (17.43)	72.14 (18.03)	1.21	0.30
AssL					
RT	3.08 (0.10)	3.08 (0.09)	3.06 (0.08)	2.58	0.08
Acc	71.20 (14.58)	71.53 (12.78)	74.59 (11.25)	2.94	0.06

Table 5

Performance change in each task for Group 2 for the one-month test-retest interval.

	F	***p***	**d**	**Raw diff**
SRT				
RT	1.32	0.26	-0.08	-6.21 ms
Acc	3.18	0.08	0.19	2.34%
CHRT				
RT	3.15	0.08	-0.11	5.29 ms
Acc	1.23	0.27	0.17	3.93%
CoRT				
RT	0.22	0.64	-0.08	-15.80 ms
Acc	0.16	0.69	0.17	5.64%
Mon				
RT	0.21	0.65	-0.14	-8.70 ms
Acc	-	-	-	0.36%
OBK				
RT	3.05	0.09	-0.11	-87.49 ms
Acc	1.43	0.24	0.13	9.07%
Mat				
RT	2.67	0.11	-0.11	-204.66 ms
Acc	3.59	0.06	0.11	14.36%
IncL				
RT	0.19	0.67	-0.13	-64.64 ms
Acc	1.19	0.28	0.13	4.00%
AssL				
RT	2.51	0.12	0.10	-53.33 ms
Acc	4.40	0.04	0.12	3.39%