*Journal of the International Neuropsychological Society* (2003), **9**, 419–428.
Copyright © 2003 INS. Published by Cambridge University Press. Printed in the USA.
DOI: 10.1017/S1355617703930074

# The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test–retest intervals

ALEXANDER COLLIE,[1,2] PAUL MARUFF,[2,3] DAVID G. DARBY,[1,2]
AND MICHAEL MCSTEPHEN[2]

[1]Centre for Neuroscience, The University of Melbourne, Parkville, Victoria, Australia
[2]CogState Ltd, Carlton South, Victoria, Australia
[3]School of Psychological Science, La Trobe University, Bundoora, Victoria, Australia

(Received October 24, 2001; Revised April 15, 2002; Accepted May 13, 2002)

**Abstract**

Performance on many cognitive and neuropsychological tests may be improved by prior exposure to testing stimuli and procedures. These beneficial practice effects can have a significant impact on test performance when conventional neuropsychological tests are administered at test–retest intervals of weeks, months or years. Many recent investigations have sought to determine changes in cognitive function over periods of minutes or hours (e.g., before and after anesthesia) using computerized tests. However, the effects of practice at such brief test–retest intervals has not been reported. The current study sought to determine the magnitude of practice effects in a group of 113 individuals assessed with an automated cognitive test battery on 4 occasions in 1 day. Practice effects were evident both between and within assessments, and also within individual tests. However, these effects occurred mostly between the 1st and 2nd administration of the test battery, with smaller, nonsignificant improvements observed between the 2nd, 3rd, and 4th administrations. On the basis of these results, methodological and statistical strategies that may aid in the differentiation of practice effects from drug-induced cognitive changes are proposed. (*JINS*, 2003, *9*, 419–428.)

**Keywords:** Practice effects, Test–retest, Cognition, Serial assessment, CogState

## INTRODUCTION

Serial assessment of cognition has been employed conventionally to aid clinical decisions regarding the onset or course of a disorder or disease, recovery of function following a pharmacological or therapeutic intervention, or changes in cognitive status caused by medical or surgical procedures. An important issue in serial neuropsychological investigations is when a change in performance from test to retest is meaningful clinically. For many neuropsychological tests, decisions regarding the significance of any cognitive change observed may be obscured by practice effects, which act to enhance test performance following repeated exposure to testing procedures and stimuli. Accordingly, many studies have sought to determine the effects of practice on neuropsychological test performance at test–retest intervals of weeks, months or years (e.g., Benedict & Zgaljardic, 1998; McCaffrey et al., 2000). These studies have also sought to determine the extent to which practice effects operate on tests of different cognitive functions. Some authors suggest that practice effects operate equally across different cognitive tests (McCaffrey et al., 1992; Mitrushina & Satz, 1991). However, others show that tests requiring complex cognitive processing, and tests where formulation of a strategy may aid performance, display greater practice effects than tests that measure more simple cognitive functions (e.g., Wisconsin Card Sorting Test, Stroop Test; Basso et al., 1999).

Neuropsychologists are now required to design trials or assess patients in serial investigations where the expected changes in cognition occur in minutes or hours. For example, following anesthesia or sedation (Ibrahim et al., 2001), the administration of a fast-acting CNS-active substance

Reprint requests to: Alex Collie, Centre for Neuroscience, The University of Melbourne, C/O 51 Leicester Street, Carlton South, Victoria, 3053, Australia. E-mail: acollie@cogstate.com

https://doi.org/10.1017/S1355617703930074 Published online by Cambridge University Press

(e.g., alcohol) or in studies of other physiological perturbations such as fatigue (Dawson & Reid, 1997) or stress (Deijen & Orlebeke, 1994). Computerized cognitive tests are often employed in these situations as they overcome many of the limitations associated with the administration of paper-and-pencil neuropsychological tests at brief intervals. For example, stimulus presentation and contingency onset are controlled by the software (thereby reducing any inter- or intra-assessor unreliability), and data collection and analysis may be automated. In addition, automated cognitive tests often have brief administration times yet still provide reliable data for comparisons of performance between treatment and nontreatment or baseline conditions within individuals. These tests may also allow the presentation of multiple and equivalent alternate forms of a test within a relatively brief time period (Collie et al., 2001a). Provided enough responses are collected, automated cognitive tests can also measure practice effects within a single testing session or even within different stages of a single test. Despite the widespread use and acceptance of computerized cognitive tests in such studies, there remains little published data describing the effects of practice at these brief test–retest intervals.

Inferences about the presence and magnitude of practice effects are also affected by the statistic used to quantify change in test performance. The subtraction of pre- from post-test performance, or the calculation of percentage change scores, are simple and commonly used methods for the determination of change in performance (e.g., Benedict & Zgaljardic, 1998). However, these methods do not take into account the normal variability in test performance or the reliability of the cognitive tests administered. Such techniques are therefore limited in the extent to which they can differentiate true change from measurement error. This limitation is exacerbated as test–retest reliability decreases (Collie et al., 2002). It has been proposed previously that the magnitude of any change in test score should be determined with a statistical technique that takes into account test specific performance variability (Jacobson & Traux, 1991).

We recently reviewed the statistical techniques used commonly to quantify change in cognitive test performance (Collie et al., 2002). Many of the statistical procedures reviewed attempted to control the effects of practice by including some *post-hoc* correction. Application of methodological strategies to minimize practice effects *a priori* may allow more accurate differentiation of the effects of practice from the effects of an independent variable than any statistical manipulation of the data once it has been collected. However, the principled development of procedural and statistical methodologies to deal with practice effects requires that appropriate data be available.

The current study aimed to determine the effects of practice at very brief test-retest intervals, in a group of test-naive individuals of similar age to those commonly referred for day surgery requiring anesthesia or sedation. Further, the current study was designed to mirror those employed in investigations of the effects of anesthetic sedation on cognitive function (e.g., Ibrahim et al., 2001). A second aim was then to develop methodological strategies for (1) minimizing practice effects; and (2) differentiating practice effects from the effects of an independent variable.

## METHODS

### Research Participants

One hundred and thirteen neurologically normal older people took part in this study. The sample had a mean age of 63.68 ± 7.58 years (range = 46–82 years), mean education of 13.11 ± 3.67 years (range = 6–22 years) and included 75 females and 38 males. All participants were enrolled in an ongoing study of aging being conducted at an independent research institute in Melbourne, Australia. The procedure through which this cohort was recruited has been described elsewhere (Collie et al., 2001b). All participants spoke English as their first language, and were regarded as having normal cognitive function as measured on a battery of neuropsychological tests sensitive to cognitive impairment in older people (Collie et al., 2001b). Inclusion criteria for all participants included age greater than 45 years, a Mini-Mental State Examination (MMSE; Folstein et al., 1975) score 26 or more and a normal neurological examination. Exclusion criteria included a history of cardiovascular disease, personal history of psychiatric or neurological illness, dementia, head injury, or evidence of cognitive impairment. To establish eligibility for inclusion in the current study, all potential participants were interviewed within the 6-month period immediately prior to study beginning. Informed consent was obtained from all participants prior to their inclusion in the study. Ethics approval was gained from the institutional ethics committee prior to the commencement of the study.

### Materials

All participants completed a battery of cognitive tests on four occasions over a 3-hr period. All tests were computerized adaptations of standard neuropsychological and experimental psychological tests (see below), and were chosen to sample from a range of cognitive domains, including psychomotor speed, attention, decision making, working memory, episodic learning and memory. These tests are also similar to those employed in cognitive studies of the effects of short-lasting CNS-active substances (Dawson & Reid, 1997; Ibrahim et al., 2001).

For each test within this battery, the stimuli consisted of playing cards. The playing cards were selected to minimize the dependence of the tests on specific languages. Pilot data indicated that most people were familiar with playing card stimuli, could differentiate the cards without additional training, and perceived their presentation to represent a game. The test battery required approximately 15 to 20 min to complete depending upon the speed of the individual subject's responses. Responses were indicated by pressing one

https://doi.org/10.1017/S1355617703930074 Published online by Cambridge University Press

of three keys on the computer keyboard (*d*, *k*, and *spacebar*). The *d* key was designated as the button to indicate *left* and the *k* key was designated to indicate *right*. A single button press (*spacebar*) was required for Tests 1 and 5 (see below). A binary decision (left or right) was required in Tests *2*, *3*, *5*, *7*, *8*, and *9*. All three buttons were required to perform Test 6. The dependent variables (DVs) collected for each test included the participant's response times (RTs) and accuracy (i.e., the percentage of correct responses or hit rate). RT was designated as the DV of interest for Tests 1 to 4, while both RT and accuracy were of interest for the remaining tests. Incorrect responses, failures to respond or responses faster than 100 ms were indicated by a buzzer and the data associated with these trials were omitted from analysis. Correct responses received no auditory feedback. The nine tests administered included the following:

1. *Simple reaction time (SRT)*. A single card was presented face-down in the center of the computer screen. Participants were required to press the spacebar whenever the card was turned face-up. Fifteen trials were presented. This test was presented three times: at the beginning (SRT1), in the middle (SRT2) and at the end (SRT3) of the battery.
2. *Choice reaction time (ChRT)*. The stimuli for this test were the same as for the SRT, however the participant was now required to indicate whether the color of the card was black or red.
3. *Complex reaction time (C×RT)*. Two cards were presented simultaneously in the center of the computer screen. The participant had to indicate whether the color of the two cards was the same or different.
4. *Continuous performance*. Participants were required to monitor the simultaneous movement of five cards, and press the spacebar as soon as any part of a card moved outside a predefined area.
5. *One-back working memory (one-back)*. Participants were required to decide whether a new card was the same, or different, to the last card presented.
6. *Divided attention*. This test combined Tests 5 and 6. Participants were required to monitor the simultaneous movement of a line of five cards, while performing the one-back test on the middle, or third, card.
7. *Matching*. A legend of six card pairs was presented at the top of the computer screen. Participants were required to decide whether or not a card pair presented at the bottom of the screen matched any of the pairs in the legend.
8. *Incidental learning*. This test followed immediately on from Test 7, and was identical to Test 7 but that the card pairs in the legend were now turned face down in the display.
9. *Associate learning (learning)*. Similar to the matching test, a legend of five card pairs was presented at the top of the computer screen. Participants were required to decide whether or not a card pair presented at the bottom of the screen matched any of the pairs in the legend. However, upon initial matching, the corresponding legend pair turned face down, and subsequent presentations of that pair had to be judged by memory. This test has 20 trials in which the same five pairs were shown four times each (i.e., 20 repeated pairs) and this was interspersed with 20 never-repeating distractor card pairs.

## Procedure

All participants were contacted by mail and asked to attend the research institute. The test battery was administered on Apple Macintosh iMacs in a large assessment laboratory, in which ten computers were designated for test administration. At least 3, and up to 10 participants performed the test at any one time, allowing rapid data collection. All participants completed the test four times within a period of 3 to 4 hr. As the intention of this study was to determine the magnitude of practice effects in test-naive participants, no practice trials were administered. A 10-min break was given between the first and second, and third and fourth administrations. A longer break (approximately 1.5 hr) was given between the second and third administrations, during which participants ate a light lunch which included caffeinated beverages if requested. Prior to each assessment, participants were required to rate their level of fatigue and anxiety on a scale from zero to 100, where zero indicated the least fatigued or anxious they had ever felt, and 100 indicated the most fatigued or anxious they had ever felt.

## Data Analysis

For all four assessments, each participant's data was examined and excluded if they had not completed the test battery. Ten participants did not complete all four assessments, and data from the remaining 103 was analyzed. For each participant the accuracy of responses was defined as the number of true positive and true negative responses divided by the total number of trials attempted. All RTs were recorded in milliseconds and RT data was trimmed at the 95% confidence intervals for each individual prior to analysis. Data analysis proceeded in five stages.

First, the test–retest reliability of each dependent variable was compared between each pair of adjacent assessments using the Pearson's product-moment correlation. Second, the extent of any improvement in group performance over the four assessments was determined by submitting each dependent variable to a one-way repeated measures analysis of variance (ANOVA). Fatigue and anxiety data, and years of education, were included in these analyses as covariates. For ANOVAs yielding a significant effect of assessment (i.e., a significant practice effect) *post-hoc* trend analysis was used to compare the goodness of fit of linear, power and logarithmic functions. Further *post-hoc* analysis directly compared performance at each assess-

https://doi.org/10.1017/S1355617703930074 Published online by Cambridge University Press

ment using paired samples *t* tests. Third, for each dependent variable, group mean practice effects were estimated by calculating a percentage change score between Assessments 1 and 4. This metric was compared to Cohen's (1988) *d* statistic calculated for the same data (Table 2). Cohen's *d* statistic is calculated between any two data points as the difference between baseline and follow-up scores divided by the averaged or pooled standard deviation (SD). Although Cohen (1988) recommends the adjusted pooled variance estimate,[1] we employed the standard formula in order to make the results of this study interpretable in SD units, and also to allow comparison with previous studies (e.g., Benedict & Zgaljardic, 1998). The group mean change scores and Cohen's *d* statistic were then calculated for each adjacent pair of assessments (Table 4). To quantitatively compare the magnitude of change between each pair of assessments, Cohen's *d* values were treated as data points in a repeated measures ANOVA. *Post-hoc* planned contrasts were used to directly compare the *d* scores for the following comparisons: *d* 1–2 *versus d* 2–3 and *d* 2–3 *versus d* 3–4.

Fourth, we conducted a more focused analysis of the SRT test data to estimate and compare the effects of practice within a single test, within a single assessment, and between multiple assessments. To do this, we first compared SRT1, SRT2, and SRT3 data from all four assessments using a 3 (SRT test) × 4 (assessment) repeated measures ANOVA (within and between assessment analysis). We then plotted group mean data for each of the 15 trials within the SRT1 test at all four assessments (within test analysis). Power curves were fitted to this trial-by-trial data to gain an estimate of the magnitude of within test practice effects occurring at all four assessments.

Finally, the *t* tests conducted in Stage 2 of the data analysis were also used to determine whether performance had changed during the "lunch break" between Assessments 2 and 3. Similar *t* tests were also conducted on the self-rated fatigue and anxiety data to determine whether the "lunch-break" affected fatigue and anxiety.

## RESULTS

The test–retest reliability coefficients of the cognitive tests administered ranged from .23 to .79, however for the majority of performance measures reliability coefficients were greater than .6 (Table 1). Reliability coefficients were generally lowest for the Test 1–Test 2 retest interval and greatest for the last two intervals on all cognitive tests. These data suggest that the reliability of the cognitive tests administered in the present study was acceptable. Therefore valid inferences about practice effects can be made on the basis of comparisons between serial assessments.

Self-rated fatigue, anxiety and level of education were not found to be significant covariates in any of the repeated measures ANOVAs conducted on the cognitive data. Table 2 shows the group mean data for each cognitive test at all four assessments. Repeated measures ANOVA conducted on this data indicated that performance improved significantly on all tests over this period, with the exception of the accuracy measure on the divided attention and incidental learning tests. The magnitude of these changes are indicated by the mean percentage change scores and the Cohen's *d* statistic, also presented in Table 2. Analysis of the *d* statistic reveals that performance improvements ranged from 0.30 to 1.22 average *SD* units. However, most improvements were of the magnitude of 0.6 to 0.9 *SD* units. Using Cohen's (1988) overlap values, these *d* scores represent between 48.4% and 61.8% overlap between the distribution of first assessment and the distributions of the fourth assessment. When the mean change score for any test was expressed as a percentage of the Assessment 1 score for that test, the amount of change was greater for simpler tests than for more complex tests. Mean changes in RT ranged from 139 ms to 324 ms, while mean changes in accuracy ranged from 6.79% to 22.76%. For all cognitive tests, performance was worse at the first test administration, as indicated by the slower RTs and lower accuracy of responses relative to the second, third, and fourth assessments. *Post-hoc* trend analyses indicated that power functions provided the best description of this practice effect for most tests (Table 3).

**Table 1.** Test–retest reliability of cognitive tests administered

| Cognitive Test | Test 1–Test 2<br>$N = 103$ | Test 2–Test 3<br>$N = 103$ | Test 3–Test 4<br>$N = 103$ |
|---|---|---|---|
| Simple Reaction Time 1 | | | |
| RT | .46 | .76 | .77 |
| Simple Reaction Time 2 | | | |
| RT | .66 | .55 | .71 |
| Simple Reaction Time 3 | | | |
| RT | .59 | .67 | .73 |
| Choice Reaction Time | | | |
| RT | .59 | .61 | .63 |
| Complex Reaction Time | | | |
| RT | .64 | .59 | .66 |
| Continuous Performance | | | |
| RT | .53 | .53 | .48 |
| One-Back | | | |
| RT | .31 | .47 | .64 |
| Accuracy | .35 | .60 | .70 |
| Divided Attention | | | |
| RT | .72 | .38 | .79 |
| Accuracy | .31 | .40 | .48 |
| Matching | | | |
| RT | .39 | .73 | .68 |
| Accuracy | .56 | .75 | .71 |
| Incidental Learning | | | |
| RT | .29 | .33 | .49 |
| Accuracy | .26* | .23 | .45 |
| Associative Learning | | | |
| RT | .65 | .71 | .72 |
| Accuracy | .69 | .68 | .74 |

All data presented are correlation coefficients; RT = response time; Accuracy = percent correct.
$*p > .05$.

[1] The formula for the denominator in the adjusted pooled variance estimate of *d* suggested by Cohen (1988) is $[SD_x + SD_y - 2r_{xy} SD_x SD_y]$.

https://doi.org/10.1017/S1355617703930074 Published online by Cambridge University Press

https://doi.org/10.1017/S1355617703930074 Published online by Cambridge University Press

**Table 2.** Cognitive test data for participants assessed four times in 3 hr

| | Test 1 | Test 2 | Test 3 | Test 4 | | | | |
|---|---|---|---|---|---|---|---|---|
| | *N* = 103 | *N* = 103 | *N* = 103 | *N* = 103 | Mean change score | | | |
| Cognitive test | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | *M* (*SD*) | % change | *d* | *F* statistic |
| Simple Reaction Time 1 | | | | | | | | |
| RT | 755.58 (510.03) | 461.99 (274.57) | 447.55 (225.02) | 433.29 (343.83) | −324.29 (542.17) | −42.92 | 0.75 | 28.66*** |
| Simple Reaction Time 2 | | | | | | | | |
| RT | 681.10 (544.44) | 582.60 (444.17) | 447.36 (258.80) | 455.19 (221.24) | −212.91 (496.66) | −31.26 | 0.59 | 13.56*** |
| Simple Reaction Time 3 | | | | | | | | |
| RT | 623.90 (538.47) | 495.88 (332.86) | 455.61 (317.59) | 428.44 (267.58) | −198.68 (480.12) | −31.84 | 0.48 | 10.89*** |
| Choice Reaction Time | | | | | | | | |
| RT | 872.16 (494.49) | 720.99 (342.47) | 636.73 (206.19) | 613.85 (148.89) | −261.31 (471.92) | −29.96 | 0.80 | 20.88*** |
| Complex Reaction Time | | | | | | | | |
| RT | 1062.61 (479.58) | 899.37 (321.32) | 852.07 (233.22) | 817.18 (209.26) | −246.91 (466.99) | −23.24 | 0.71 | 17.70*** |
| Continuous Performance | | | | | | | | |
| RT | 842.69 (773.86) | 619.33 (321.45) | 576.79 (176.02) | 535.31 (154.82) | −309.97 (739.77) | −36.78 | 0.66 | 13.56*** |
| One Back | | | | | | | | |
| RT | 1197.59 (432.59) | 1086.11 (335.20) | 1042.53 (333.62) | 992.52 (306.62) | −207.97 (401.94) | −17.36 | 0.55 | 10.03*** |
| Accuracy | 62.89 (18.63) | 72.79 (17.94) | 76.29 (18.49) | 78.83 (17.84) | 15.97 (21.18) | 25.41 | 0.87 | 28.67*** |
| Divided Attention | | | | | | | | |
| RT | 1594.37 (624.51) | 1495.63 (557.96) | 1463.08 (545.08) | 1409.33 (425.73) | −185.09 (646.37) | −11.61 | 0.35 | 2.57$^{\mathrm{NS}}$ |
| Accuracy | 71.07 (19.48) | 75.42 (18.22) | 71.58 (18.79) | 77.86 (18.25) | 6.79 (24.49) | 9.55 | 0.36 | 4.32** |
| Matching | | | | | | | | |
| RT | 1947.82 (475.41) | 1749.12 (433.76) | 1732.01 (387.08) | 1762.69 (399.41) | −188.61 (548.57) | −9.68 | 0.42 | 15.44*** |
| Accuracy | 64.98 (22.79) | 83.37 (17.56) | 86.81 (15.07) | 87.74 (14.57) | 22.76 (21.84) | 15.39 | 1.22 | 76.95*** |
| Incidental Learning | | | | | | | | |
| RT | 1391.49 (494.32) | 1350.87 (366.50) | 1299.68 (351.93) | 1253.43 (418.88) | −138.66 (522.15) | −9.96 | 0.30 | 0.58$^{\mathrm{NS}}$ |
| Accuracy | 48.25 (19.26) | 65.41 (21.22) | 65.83 (19.30) | 62.48 (20.69) | 14.13 (25.50) | 29.28 | 0.71 | 19.79*** |
| Associative Learning | | | | | | | | |
| RT | 1669.68 (360.52) | 1569.75 (282.21) | 1537.70 (254.91) | 1477.69 (245.31) | −207.70 (331.75) | −12.44 | 0.63 | 9.14*** |
| Accuracy | 63.81 (15.74) | 71.79 (14.98) | 73.59 (14.83) | 74.49 (13.95) | 10.68 (11.27) | 16.74 | 0.72 | 13.17*** |

All data are presented as mean (±*SD*) unless otherwise stated; RT = reaction time; Accuracy = percent correct; Test 1 = first administration; Test 2 = second administration; Test 3 = third administration; Test 4 = fourth administration.; All *d* statistics were calculated by dividing the difference between the mean scores by the averaged *SD*. * = $p < .05$; ** = $p < .01$; *** = $p < .001$; NS = not significant; mean change score is Test 4 mean minus Test 1 mean. All response times are in millisecond units.

**Table 3.** Results of linear, power and logarithmic trend analysis conducted on cognitive data collected over four assessments

| | Linear function equation | $r^2$ | Power function equation | $r^2$ | Logarithmic function equation | $r^2$ |
|---|---|---|---|---|---|---|
| Simple Reaction Time 1 | | | | | | |
| RT | $y = -98.133x + 769.94$ | .67 | $y = 689.29x^{-.3233}$ | .92 | $y = -234.86\text{Ln}(x) + 711.21$ | .84 |
| Simple Reaction Time 2 | | | | | | |
| RT | $y = -81.298x + 744.81$ | .88 | $y = 703.36x^{-0.4044}$ | .85 | $y = -179.66\text{Ln}(x) + 684.3$ | .93 |
| Simple Reaction Time 3 | | | | | | |
| RT | $y = -62.665x + 657.62$ | .87 | $y = 615.24x^{-0.2718}$ | .99 | $y = -141.92\text{Ln}(x) + 613.71$ | .97 |
| Choice Reaction Time | | | | | | |
| RT | $y = -85.921x + 925.74$ | .90 | $y = 867.15x^{-0.2622}$ | .99 | $y = -192.9\text{Ln}(x) + 864.19$ | .98 |
| Complex Reaction Time | | | | | | |
| RT | $y = -78.361x + 1103.7$ | .87 | $y = 1050.1x^{-0.1897}$ | .98 | $y = -177.75\text{Ln}(x) + 1049$ | .97 |
| Continuous Performance | | | | | | |
| RT | $y = -96.467x + 884.7$ | .82 | $y = 820.38x^{-0.3252}$ | .96 | $y = -221.47\text{Ln}(x) + 819.49$ | .94 |
| One Back | | | | | | |
| RT | $y = -65.88x + 1244.4$ | .95 | $y = 1196.4x^{-0.1323}$ | .99 | $y = -145\text{Ln}(x) + 1194.9$ | .99 |
| Accuracy | $y = 6.2854x + 58.281$ | .89 | $y = 62.819x^{0.1993}$ | .97 | $y = 14.154\text{Ln}(x) + 62.749$ | .98 |
| Divided Attention | | | | | | |
| RT | $y = -58.767x + 1637.5$ | .95 | $y = 1593.7x^{-0.0855}$ | .99 | $y = -128.59\text{Ln}(x) + 1592.8$ | .99 |
| Accuracy | $y = 2.0976x + 69.498$ | .45 | $y = 71.223x^{0.0593}$ | .44 | $y = 4.4472\text{Ln}(x) + 71.208$ | .44 |
| Matching | | | | | | |
| RT | $y = -57.249x + 1941$ | .54 | $y = 1909.8x^{-0.0774}$ | .73 | $y = -143\text{Ln}(x) + 1911.5$ | .73 |
| Accuracy | $y = 7.4782x + 64.577$ | .77 | $y = 69.269x^{0.2227}$ | .90 | $y = 17.452\text{Ln}(x) + 69.407$ | .91 |
| Incidental Learning | | | | | | |
| RT | $y = -46.539x + 1440.2$ | .99 | $y = 1402.3x^{-0.0734}$ | .94 | $y = -97.312\text{Ln}(x) + 1401.2$ | .95 |
| Accuracy | $y = 3.368x + 52.819$ | .29 | $y = 53.344x^{0.1651}$ | .52 | $y = 9.2905\text{Ln}(x) + 53.857$ | .49 |
| Associative Learning | | | | | | |
| RT | $y = -60.801x + 1715.7$ | .96 | $y = 1670x^{-0.084}$ | .98 | $y = -132.47\text{Ln}(x) + 1669$ | .98 |
| Accuracy | $y = 4.0864x + 61.045$ | .97 | $y = 64.296x^{0.1267}$ | .99 | $y = 8.8855\text{Ln}(x) + 64.202$ | .99 |

RT = Response time in milliseconds; Accuracy = percent correct.

Table 4 shows the mean percentage change scores and the Cohen's *d* statistic for all cognitive tests at each test–retest interval. The results indicate that significant improvements in performance occurred between Assessment 1 and Assessment 2 for all cognitive tests, with the exception of the divided attention test. In contrast, few significant changes in performance occurred at the later test–retest intervals, although this may be partly due to the relatively large variability in group performance masking the small group mean improvements (Table 2). For most tests, Cohen's *d* was larger for the Assessment 1–Assessment 2 interval than *d* calculated for the later between assessment intervals; however, these differences only reached significance for the SRT1 $[F(1,102) = 14.64, p < .001]$, matching $[F(1,102) = 8.01, p = .001]$ and incidental learning tasks $[F(1,102) = 8.22, p = .001]$.

Statistical analysis of the SRT test data (Table 2) with repeated measures ANOVA indicated significant main effects of assessment $[F(1,102) = 16.99, p < .001]$ and SRT test $[F(1,102) = 5.23, p = .007]$ as well as a significant Assessment × SRT test interaction $[F(2,103) = 3.99, p = .001]$. Further one-way ANOVAs conducted on the SRT test data for each assessment indicated that significant changes in RT occurred within the first assessment [i.e., SRT1 *vs.* SRT2 *vs.* SRT3; $F(2,101) = 7.89$, $p = .001$], and within the second assessment $[F(2,101) = 5.31, p = .006]$, but not within any of the later assessments [Assessment 3: $F(2,101) = .89$, $p = .41$; Assessment 4: $F(2,101) = 1.94$, $p = .15$].

Figure 1 displays power curves fitted to group mean RT data for each of 15 individual trials included in the SRT1 test at all four assessments. At the first assessment, a rapid decrease in mean RT is observed over the first 5 to 6 trials, followed by a stabilization of RTs over the last 9 to 10 trials. This pattern contrasts that observed in later assessments, where faster and more consistent RTs are observed across all trials.

Importantly, very few significant changes were observed between assessments two and three, during which time participants ate lunch. Further, self-rated levels of fatigue and anxiety did not change significantly during this "lunch break".

## DISCUSSION

When a brief automated cognitive test battery was administered to a large group of neurologically normal individuals four times during a period of 3 to 4 hr, significant improvements in test performance were observed both between and within assessments. This improvement was most

https://doi.org/10.1017/S1355617703930074 Published online by Cambridge University Press

**Table 4.** Mean change scores, percentage change scores and Cohen's *d* on all cognitive tests for Assessments 1 to 4

| | Test 1–Test 2 | | | Test 2–Test 3 | | | Test 3–Test 4 | | |
|---|---|---|---|---|---|---|---|---|---|
| Cognitive test | Mean change score | % change | *d* | Mean change score | % change | *d* | Mean change score | % change | *d* |
| Simple Reaction Time 1 | | | | | | | | | |
| RT | −300.55 (475.22)*** | −39.78 | .75 | −14.45 (225.66) | −3.12 | .06 | −14.26 (270.17) | −3.19 | .05 |
| Simple Reaction Time 2 | | | | | | | | | |
| RT | −102.06 (494.34)** | −14.98 | .20 | −113.38 (324.65)** | −19.46 | .38 | −3.40 (204.01) | −0.76 | −.03 |
| Simple Reaction Time 3 | | | | | | | | | |
| RT | −128.88 (433.97)*** | −20.66 | .29 | −38.09 (228.87) | −7.68 | .12 | −29.27 (214.67) | −6.42 | .09 |
| Choice Reaction Time | | | | | | | | | |
| RT | −159.88 (401.61)*** | −18.33 | .36 | −84.26 (270.22)*** | −11.69 | .31 | −22.89 (161.01) | −3.59 | .13 |
| Complex Reaction Time | | | | | | | | | |
| RT | −160.25 (370.89)*** | −15.08 | .41 | −50.54 (263.73) | −7.01 | .17 | −34.89 (183.63) | −4.09 | .16 |
| Continuous Performance | | | | | | | | | |
| RT | −230.36 (662.59)*** | −27.34 | .41 | −42.54 (272.10) | −3.91 | .17 | −41.49 (169.90)* | −7.19 | .25 |
| One Back | | | | | | | | | |
| RT | −120.47 (458.30)** | −10.06 | .29 | −43.58 (345.78) | −4.01 | .13 | −50.01 (272.23) | −4.80 | .16 |
| Accuracy | 10.99 (21.28)*** | 17.47 | .54 | 3.35 (16.41)* | 4.60 | .19 | 2.49 (13.89) | 3.27 | .14 |
| Divided Attention | | | | | | | | | |
| RT | −86.33 (749.58) | −5.41 | .17 | −42.50 (719.49) | −2.84 | .06 | −53.75 (604.53) | −3.67 | .11 |
| Accuracy | 4.18 (22.53) | 5.88 | .13 | −3.59 (20.32) | −4.76 | −.20 | 7.78 (19.19)*** | 10.87 | .34 |
| Matching | | | | | | | | | |
| RT | −225.42 (485.65)*** | −11.57 | .44 | −33.59 (299.18) | −1.92 | .04 | 30.68 (313.39) | 1.77 | −.08 |
| Accuracy | 18.56 (19.28)*** | 28.56 | .91 | 3.42 (11.59)** | 4.10 | .21 | 1.39 (11.55) | 1.60 | .06 |
| Incidental Matching | | | | | | | | | |
| RT | −73.97 (491.25)* | −5.32 | .09 | −36.51 (392.63) | −2.70 | .14 | −20.86 (386.08) | −1.61 | .12 |
| Accuracy | 16.67 (24.59)*** | 34.55 | .85 | 1.93 (25.71) | 2.93 | .02 | −4.09 (21.83) | 6.23 | −.17 |
| Associative Learning | | | | | | | | | |
| RT | −138.26 (284.09)** | −8.28 | .31 | −69.55 (211.76) | −4.43 | .11 | −29.78 (167.14) | −1.94 | .24 |
| Accuracy | 6.85 (10.82)*** | 10.73 | .52 | 2.52 (10.61) | 3.51 | .12 | 1.60 (9.15) | 2.17 | .06 |

RT = response time; Accuracy = percent correct. All change score data is presented as mean (±*SD*). All *d* statistics were calculated by dividing the difference between the mean scores by the averaged *SD*; * = $p < .05$; ** = $p < .01$; *** = $p < .001$. All response times are in millisecond units.

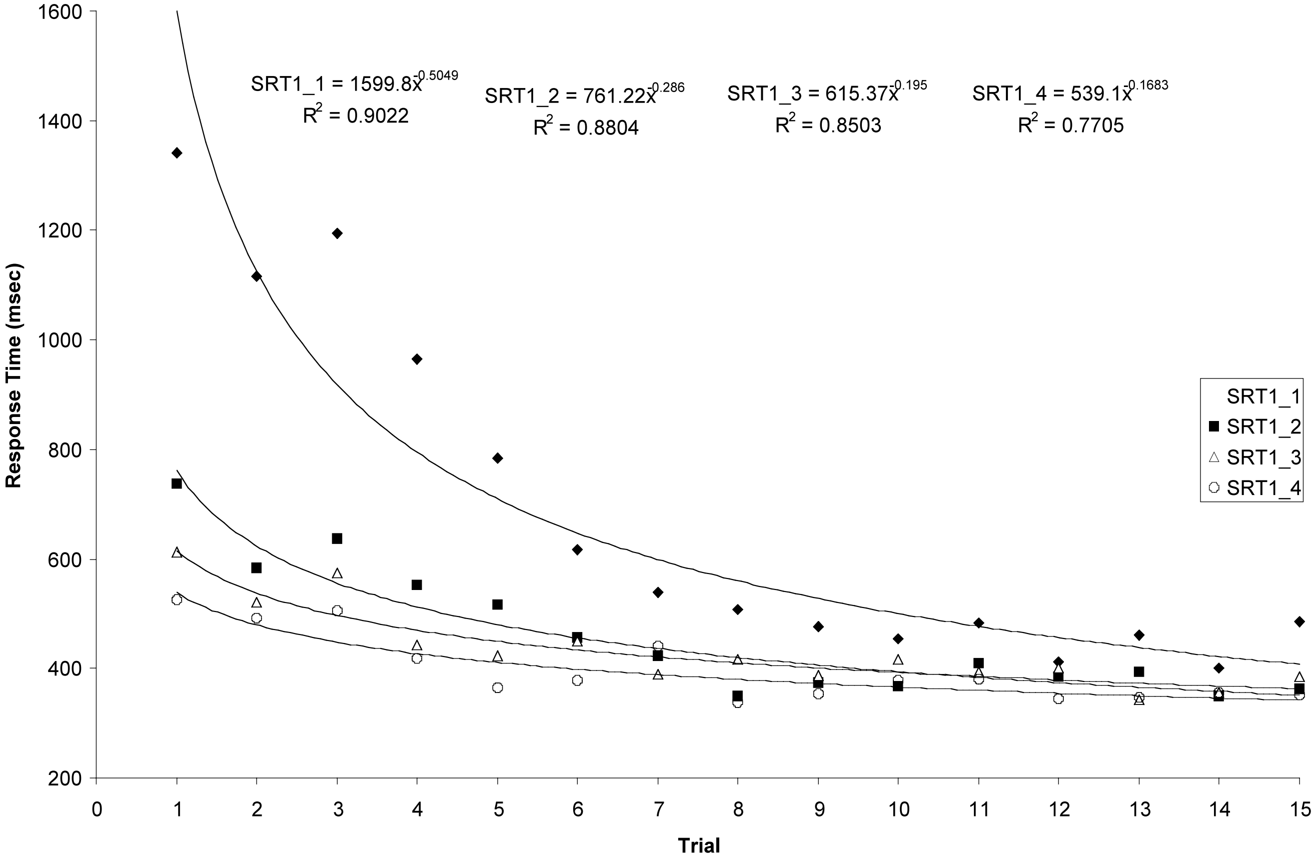

**Fig. 1.** Within and between assessment practice effects revealed by analyses of trial by trial data for the Simple Reaction Time (SRT) test. Power curves are fitted to the group mean data plotted by trial number. Regression equations and $r^2$ values are displayed on the chart. SRT1_1 = SRT Test 1 in Assessment 1; SRT1_2 = SRT Test 1 in Assessment 2; SRT1_3 = SRT Test 1 in Assessment 3; SRT1_4 = SRT Test 1 in Assessment 4.

evident between the first and second assessment, as performance remained more stable between the second, third and fourth assessments (Table 2, Figure 1). These results are consistent with those of recent studies that have employed conventional neuropsychological tests and test–retest intervals of weeks or months (e.g., Benedict & Zgaljardic, 1998; Ivnik et al., 1999; Theisen et al., 1998). On the basis of their findings, these studies propose that the positive effects of practice are most evident between the first and second administration of a cognitive test. While our results generally support this proposal, there are important differences between the present and previous investigations.

In the present study, all assessments were conducted within a brief period of time to reflect the design of studies investigating the effects of anesthesia or sedation on cognitive function. In contrast, prior studies have employed test–retest intervals ranging from days to years to provide meaningful serial normative data for investigations of aging, disease processes, and recovery of cognitive function following medical or pharmacological treatment (McCaffrey et al., 2000). The poor psychometric properties of some commonly employed neuropsychological tests may act to confound accurate interpretation of serially acquired data (e.g., floor and ceiling effects, poor test–retest reliability; Collie et al., 2002). Combined with the amount of time often required to administer a conventional test battery, these properties ensure that paper-and-pencil tests are not commonly used in settings where serial assessments are required at very brief test–retest intervals.

The cognitive test battery employed in the current study is based on standard neuropsychological and experimental psychological tests, has many alternate forms and acceptable test–retest reliability (Table 1), although the equivalence of these alternate forms has not been empirically established. The more complex tests within the present battery are sufficiently difficult that ceiling effects are uncommon, while the output of interest from the less complex tests is in the form of response times, which allows the identification of even minor changes in performance (e.g., in the order of milliseconds). Combined, these factors ensure that the data presented in the current study is interpretable and meaningful, and that very mild changes in cognitive function may be observed reliably.

Our data suggests that a number of methodological strategies may aid in the differentiation of practice from treatment effects in studies employing brief test–retest intervals. First, practice effects appear to operate mainly between first and second assessments on most tests (Tables 2 and 4). An

https://doi.org/10.1017/S1355617703930074 Published online by Cambridge University Press

effective method of minimizing practice effects prior to a treatment condition may therefore be to conduct dual baseline assessments and exclude the results of the first from further analysis. Second, it appears that at least on psychomotor tests, practice effects only operate for the first 5 to 6 trials in a 15-trial test (Figure 1). *Post-hoc* exclusion of the first few trials on these tests may therefore minimize the effects of practice and provide equivalence between assessments. Finally, inclusion of a nontreatment group in whom the magnitude of practice effects can be quantified may aid data interpretation, and allow determination of any interactions between practice and treatment effects (Collie et al., 2002). Adoption of all of these strategies concurrently would provide maximum protection against potentially very large practice effects, and allow more accurate analysis and interpretations of treatment effects.

Many recent research articles have employed a conventional method of calculating group changes in test score, whereby baseline score is subtracted from follow-up score and the degree of change is expressed as a percentage of the baseline score (e.g., Benedict & Zgaljardic, 1998). However, this method considers only the mean level of performance, and does not account for variability in group performance. In the current study, we compared the magnitude of practice effects estimated according to this conventional method with estimates derived from a statistical method that takes into account the variability in test score (Cohen's *d* statistic; Table 2). This statistic has been used previously to examine test–retest differences in neuropsychological tests (Zakzanis et al., 1995). Results from the percentage change method suggested that as test complexity increased the effects of practice decreased. This contrasts previous investigations that observe greater and longer lasting practice effects on more complex cognitive tests (e.g., Wisconsin Card Sorting Test), and propose that this is due to the individual adopting some strategy for test completion or for accurately remembering test stimuli between first and second testing sessions (Basso et al., 1999). For this reason, the results of the percentage change method seem counter-intuitive and unlikely. Analysis of the *d* statistic indicated that practice effects were relatively uniform across all tests. While these results also contrast with previous research, they are not without precedent. For example, we calculated the *d* statistic on the data presented by Ivnik and colleagues (1999; Table 3), who derived factor scores from data gained by administering the Wechsler Adult Intelligence Scale–Revised (WAIS–R) and the Wechsler Memory Scale–Revised (WMS–R) to a group of 50 normal people on four occasions at test–retest intervals of 12 months. A uniform improvement was observed between the first and second testing sessions for factor scores representing varied cognitive domains (Verbal Comprehension: $d = .23$; Perceptual Organization: $d = .30$; Attention–Concentration: $d = .20$; Retention: $d = .34$), with the exception of a learning factor in which a greater level of improvement was observed ($d = 0.62$). The smaller than expected practice effects observed on difficult tasks in the present study may also be due to their temporal position at the end of the test battery, given our observation that practice effects occur within a 15-min testing session (see below).

As noted in the introduction, the automation of cognitive tests allows the examination of changes in performance within a testing session if enough responses are collected to allow statistical comparison. Our results indicate that RTs on the three SRT tests became significantly faster during the first assessment (Table 2). These SRT tests were administered at the beginning (SRT1), in the middle (SRT2), and at the end (SRT3) of the assessment, during which several other more complex cognitive tests were administered. These findings suggest that prior test exposure can affect subsequent performance at test–retest intervals of minutes, and occur even when the individual is performing other tests for the entire test–retest interval. At the second assessment, performance on the SRT tests had reached an asymptote and no further significant improvements were recorded. Similarly, no changes in SRT performance were observed for the third and fourth assessments. This finding has implications for the interpretation of data collected during any cognitive test, and particularly during tests of duration greater than 15–20 min, as it suggests that without prior exposure the individual's level of performance may alter significantly from the beginning to the end of a testing session of this length. This finding also implies that the distribution of data for tests or test batteries with similar or longer durations should be inspected prior to statistical analysis, and that the effects of within-session practice effects should be determined and corrected methodologically. However, this correction may not be necessary for data collected in subsequent assessments of the same individual.

A number of factors may limit any conclusions drawn from the present study. For example, the sample studied were well educated, very healthy and of an older age group. All of these factors may have some influence on the magnitude of practice effects (Horton, 1992; Rapport et al., 1997). Further studies are required in different demographic groups before the current findings may be generalized. Also, the playing-card stimuli employed for the cognitive tests may have been familiar to some participants but not to others. It is possible that in some participants regular exposure to card games may have acted to reduce the magnitude of practice effects, as previous research has observed that mere exposure to test-taking situations may positively influence performance (Anastasi, 1988). Our results and conclusions may be of particular interest to clinicians working in medical fields where patients are exposed to agents or interventions that may have transient or short-lasting effects on cognition.

## REFERENCES

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

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

Benedict, R.H.B. & Zgaljardic, D.J. (1998). Practice effects during repeated administrations of memory test with and without

https://doi.org/10.1017/S1355617703930074 Published online by Cambridge University Press

alternate forms. *Journal of Clinical and Experimental Neuropsychology*, *20*, 339–352.

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

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

Collie, A., Darby, D.G., & Maruff, P. (2001). Computerised cognitive assessment of athletes with sports-related head injury. *British Journal of Sports Medicine*, *35*, 297–302.

Collie, A., Maruff, P., Shafiq-Antonacci, R., Smith, M., Hallup, M., Schofield, P., Masters, C., & Currie, J. (2001). Memory decline in healthy older people: Implications for identifying mild cognitive impairment. *Neurology*, *56*, 1533–1538.

Dawson, D. & Reid, K. (1997). Fatigue, alcohol and performance impairment. *Nature*, *388*, 235.

Deijen, J.B. & Orlebeke, J.F. (1994). Effect of tyrosine on cognitive function and blood pressure under stress. *Brain Research Bulletin*, *33*, 319–323.

Folstein, M.F., Folstein, S.E., & McHugh, P.R. (1975). "Mini-mental state": A practical guide for grading the cognitive status of patients for the clinician. *Journal of Psychiatric Research*, *12*, 189–198.

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

Ibrahim, A.E., Ghoneim, M.M., Kharasch, E.D., Epstein, R.H., Groudine, S.B., Ebert, T.J., Binstock, W.B., Philip, B.K., & the Sevoflurane Sedation Study Group. (2001). Speed of recovery and side-effect profile of sevoflurane sedation compared with midazolam. *Anesthesiology*, *94*, 87–94.

Ivnik, R.J., Smith, G.E., Lucas, J.A., Petersen, R.C., Boeve, B.F., Kokmen, E., & Tangalos, E.G., (1999). Testing normal older people three to four times at 1- to 2-year intervals: Defining normal variance. *Neuropsychology*, *13*, 121–127.

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

McCaffrey, R.J., Duff, K., & Westervelt, H.J. (2000). *Practitioner's guide to evaluating change with neuropsychological assessment instruments*. New York: Kluwer Academic/Plenum Publishers.

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

Mitrushina, M. & Satz, P. (1991). Effect of repeated administration of a neuropsychological battery in the elderly. *Journal of Clinical Psychology*, *47*, 790–800.

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

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

Zakzanis, K.K., Heinrichs, R.W., & Ruttan, L.A. (1995). Three-year test–retest reliability of neurocognitive measures in schizophrenia. *Schizophrenia Research*, *15*, 140–148.

https://doi.org/10.1017/S1355617703930074 Published online by Cambridge University Press