Psychometric evaluation of The Twelve Elements Test and other commonly used measures of Executive Function

by

Claire Surinder Sira
B.Sc., University of Victoria, 1994
M.A., Queen’s University, 1997

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Psychology

© Claire Sira, 2007
University of Victoria

All rights reserved. This thesis may not be reproduced in whole or in part, by photocopy or other means, without the permission of the author.
Psychometric evaluation of The Twelve Elements Test and other commonly used measures of executive function

by

Claire Surinder Sira
B.Sc., Honours, University of Victoria, 1994
M.A., Queen’s University, 1997

Supervisory Committee

Dr. Catherine Mateer, (Department of Psychology)
Supervisor

Dr. Holly Tuokko, (Department of Psychology)
Departmental Member

Dr. Kimberly Kerns, (Department of Psychology)
Departmental Member

Dr. Jillian Roberts, (Department of Educational Psychology and Leadership Studies)
Outside Member
Abstract

Objective: The Six Elements Task (SET; Shallice and Burgess, 1991; Burgess et al., 1996) measures examinees’ ability to plan and organize their behaviour, form strategies for novel problem solving, and self-monitor. The task has adequate specificity (Wilson et al., 1996), but questionable sensitivity to mild impairments in executive function (Jelicic, et al., 2001). The SET is vulnerable to practice effects. There is a limited range in possible scores, and ceiling effects are observed. This dissertation sought to evaluate the validity and clinical utility of a modification of the SET by increasing the difficulty of the test, and expanding the range of possible scores in order to make it more suitable for serial assessments.

Participants and Methods: The sample included 26 individuals with mixed acquired brain injury, and 26 healthy matched controls (20 – 65 years). Participants completed a battery of neuropsychological tests on two occasions eight weeks apart. To control for confounding variables in executive function test performance, measures of memory, working memory, intelligence, substance abuse, pain, mood and personality were included. Self and informant reports of executive dysfunction were also completed. The two groups’ performances on the various measures were compared, and the external
validity of the 12ET was examined. In addition, normative data and information for reliable change calculations were tabulated.

Results: The ABI group exhibited very mild executive function deficits on established measures. The matched control group attempted more tasks on the 12ET, but the difference was non significant. Neither group tended to break the rule of the task. The 12ET showed convergent validity with significant correlations with measures of cognitive flexibility (Trailmaking B and Ruff Figural Fluency), and a measure of planning (Tower of London). The 12ET and published measures were also significantly correlated with intelligence in the brain-injured group. The 12ET did not show divergent validity with a test of visual scanning speed (Trailmaking A). No demographic variables were found to be significant predictors of 12ET performance at Time 2 over and above performance at Time 1, and both participant groups obtained the same benefit from practice. The 12ET did not suffer from ceiling effects on the second administration, and the test-retest reliability of the 12ET variables ranged from low (r = .22 for Rule Breaks in the brain-injured group) to high (r = .78 for Number of Tasks Attempted in the control group).

Conclusions: Despite their (often severe) brain injuries, this sample of brain injured participants did not demonstrate executive impairments on many published tests and their scores were not significantly different from the control group’s scores. Therefore, it was not possible to determine if the 12ET was a more sensitive measure of mild executive deficits than the SET. However, the increase in range did reduce the tendency for participants to perform at ceiling levels. The 12ET showed a number of significant correlations with other executive measures, particularly for the brain-injured group, though these correlations may have been moderated by general intelligence. Two variables of the 12ET, deviation from the optimal amount of time per task and Number of Tasks Completed, showed promise as measures of reliable change in this sample over an 8-week interval.
TABLE OF CONTENTS

SUPERVISORY COMMITTEE ........................................................................................................... ii
Abstract ........................................................................................................................................ iii
Table of Contents ......................................................................................................................... v
List of Tables ................................................................................................................................. viii
List of Figures ................................................................................................................................. ix
Acknowledgments ........................................................................................................................ x
Introduction .................................................................................................................................... 1
  Intelligence and Executive Function ......................................................................................... 4
  Measurement Error ....................................................................................................................... 8
    Issues in Test-Retest Reliability ................................................................................................. 16
    Methodological Procedures to Control for Practice Effects .................................................... 17
  Measuring Change on Neuropsychological Tests .................................................................. 18
  Limitations of Reliable Change Formulae .................................................................................. 22
    Standardized Regression Based Change Scores ................................................................. 22
    Limitations of SRBs ................................................................................................................ 24
  Rationale of the Current Study ................................................................................................. 25
  The Six Elements Task .............................................................................................................. 25
    Limitations of the Six Elements Test ...................................................................................... 32
  Personality and the Twelve Elements Test .............................................................................. 34
  Assessment of Executive Function Through Questionnaires ............................................... 36
The Current Study ...................................................................................................................... 37
  Part 1. Evaluating the Validity of the Twelve Elements Test .................................................. 37
  Part 2. Convergent and Divergent Validity of the Twelve Elements Test ............................ 37
  Convergent Validity ................................................................................................................ 37
    The 12ET and Intelligence ....................................................................................................... 37
    The 12ET and a Measure of Time Estimation ...................................................................... 38
    The 12ET and Measures of Cognitive Flexibility .................................................................. 38
    The 12ET and a Measure of Planning ................................................................................... 39
    The 12ET and a Measure of Self Monitoring ..................................................................... 40
    The 12ET and Self and Informant Reports ......................................................................... 40
    The 12ET and Personality ...................................................................................................... 41
  Divergent Validity ..................................................................................................................... 41
    The 12ET and Visual Scanning Speed ................................................................................... 41
  Part 3 Reliability and Normative Data of the 12ET ................................................................. 41
Method ........................................................................................................................................ 43
  Participants ............................................................................................................................... 43
  Measures ................................................................................................................................. 47
  Procedure ................................................................................................................................... 52
Results .......................................................................................................................................... 64
  Data Cleaning ........................................................................................................................... 64
Part 1 – Evaluating the Validity of the 12 Elements Test.......................................................... 66
Appendix C  Severity of Dependence Scale

The 12ET and Memory ................................................................. 121
The 12ET and Tests of Cognitive Flexibility ................................ 122
The 12ET and a Test of Planning in the Control Group ............ 122
Divergent Validity for the Control Participants ...................... 123
The 12ET and Visual Scanning Speed ......................................... 123
12ET Intercorrelations for the Control Group ......................... 123
Intercorrelations between Published Neuropsychological Tests .... 123
Differences in Performance on Published Executive Function Tests ... 124
Part 3. Normative Data and Reliability of the 12ET ...................... 126
Normative Data ...................................................................... 126
Practice Effects and Reliable Change on the 12ET .................. 126
Limitations ............................................................................. 129
Conclusion ............................................................................. 132
Future Directions ................................................................. 133
References ............................................................................. 138
Appendix A Instructions for The Twelve Elements Task .......... 149
Appendix  B Consent Forms ............................................... 151
Appendix C  Severity of Dependence Scale ................................. 163
LIST OF TABLES

Table 1. Range of demographic and IQ scores .......................................................... 54
Table 2. Neuropsychological test variables collected and calculated ............................. 68
Table 3. 12ET Intercorrelation table ........................................................................... 70
Table 4. Independent t tests between the brain-injured and control participant scores on the published neuropsychological tests (N=52) ................................................................... 74
Table 5. Spearman Correlation Matrix of Dependent variables with Criterion Variables for all Participants ................................................................. 74
Table 6. Spearman Correlations between the Neuropsychological Tests in all participants (N=52) Time 1 ............................................................................. 75
Table 7. Spearman Correlations of Dependent variables with Criterion Variables for Brain-injured Participants at Time 1 .................................................................. 82
Table 8. Spearman Correlations between Neuropsychological Tests in the Brain-injured Sample (N=26 except for the DEX where N=22) at Time 1 ........................................... 83
Table 9. Spearman Correlations of Dependent variables with Criterion Variables for Control Participants at Time 1 ........................................................................ 88
Table 10. Selected Spearman Intercorrelations between Neuropsychological Tests for the Control Group at Time 1 (N=26) ................................................................. 90
Table 11. Crosstabulation by Group for 12ET Rule Breaks ........................................... 93
Table 12. Chi-Square Tests for 12ET Rule Breaks ....................................................... 94
Table 13. Crosstabulation by Group for 12ET for > minutes on one task .................. 94
Table 14. Chi Square Tests for Spending >7 Minutes on One Task .............................. 94
Table 15. Spearman’s Correlations of Attribute Variables for all participants ............ 97
Table 16. Normative Data for the 12ET at Time 1 ...................................................... 100
Table 17. Test-retest Reliability Coefficients for the 12ET over an 8 week interval ..... 101
Table 18. Normative Data for the 12ET at Time 2 ...................................................... 102
Table 19. 12ET Scores Required for Reliable Change after an 8 Week Interval .......... 104
LIST OF FIGURES

Figure 1 The Twelve Elements Test .......................................................... 34
Figure 2. Bar Graph of 12ET Number of Tasks Attempted for at Time 1 .......... 70
Figure 3. Bar Graph of 12ET Rule Breaks for both groups at Time 1 .......... 71
Figure 4. Scatterplot of 12ET # Tasks Completed and WTAR Standard Scores for all Participants at Time 1 ................................................................. 77
Figure 5. Scatterplot of 12ET # Time Requests and WTAR Standard Scores for all Participants ................................................................. 78
Figure 6. Scatterplot of Trailmaking Test Part A and 12ET Tasks Attempted for both groups at Time 1 ................................................................. 81
Figure 7. Trailmaking A T Score and Whether or not participants spent >7 Minutes on one 12ET task at Time 1 ................................................................. 84
Figure 8. Scatterplot of WTAR with 12ET Number of Tasks Attempted for Brain-injured Participants at Time 1 ................................................................. 85
Figure 9. Scatterplot of 12ET Tasks Completed and Tower of London Total Move score for Brain-injured participants at Time 1 ................................................................. 86
Figure 10. Scatterplot 12ET Rule Break and TOL Rule Violation for the Control Group at Time 1 ................................................................. 86
Figure 11. Scatterplot of Number of Time Requests with NEO Conscientiousness T Scores for Control Participants ................................................................. 87
Figure 12. Bar Graph of 12ET Rule Breaks at Time 1 and Time 2 .................... 95
Figure 13. Scatterplot of 12ET Rule Break at Time 1 and Time 2 with WTAR SS for the Brain-injured Group ................................................................. 98
Figure 14. Scatterplot for 12ET Rule Break at Time 1 and Time 2 for the Control Group ................................................................. 98
ACKNOWLEDGMENTS

The idea for this dissertation came out of psychometrics discussions with Dr. Roger Graves, and I am indebted to my committee for helping me clarify the proposal. Upon Roger’s retirement midway through my research, Kim Kerns stepped in to take his place, which was very generous of her. In addition, I would like to thank the Thames Valley Test Company and Dr. Paul Burgess for permitting me to modify the Six Elements Test for research purposes.

I am also very grateful to Braxton Suffield and the managers of Columbia Health Centres and LifeMark Health for allowing me their space to collect my last pieces of data. Of course, I am indebted to all of my participants in Victoria and Calgary.

A person who has shared many moments with me throughout my doctoral degree, and deserves special mention is Lori-Ann Larsen. I’m grateful to have such a good friend.

Though I can’t adequately thank Katy Mateer in these few lines, I will say that she has had a tremendous impact on my career and personal development. Through offering me both clinical and academic opportunities, she has quietly supported and shaped me. I look forward to the next stage of our friendship.

This degree has been the hardest thing I have ever done. It’s true that they don’t give these degrees away. The support of my parents, friends and (most of all) my husband helped me to persist over the years. They knew I’d get there, even when I wasn’t so certain.
Executive function deficits are common after acquired brain injury. They affect all facets of an individual’s life including work, leisure and home life (Sohlberg & Mateer, 2001), and greatly reduce an individual’s ability to live independently in the community (Lezak, 1995). Executive dysfunction typically involves problems in goal-directed behaviour in novel contexts, where there may be competing response alternatives. People with executive dysfunction often demonstrate problems with novel problem solving, planning and judgment. They may understand what is required of them, but fail because of perseveration, impersistence, lack of initiative or intrusions of task-irrelevant behaviour. Activity, as a result, often seems purposeless, disorganized and ineffective. The difficulties are usually most evident when the person needs to do something novel or non-routine. Routine behaviours appear to be less affected because they are in some way already programmed, and can be completed with little overt attention or effort. To do something more novel, it becomes necessary to pay attention, think the problem through, monitor what one is doing in order to suppress inappropriate automatic actions (interference), and be flexible enough to change the plan if things are not working as they should (Mesulam, 2002; Mateer & Sira, 2003a; Stuss & Benson, 1984). To add to the complexity, all the preceding variables interact to varying degrees. An example of such interdependence is where good strategic planning may reduce the potency of interference (Burgess, 1997).

The various components of executive function can fractionate and interact in complex ways. Given the complexities of the system, it is challenging to develop
effective assessment strategies. Assessment is often complicated in that the person may function quite well in a structured environment in which the goals of activity are made explicit. In a testing situation, this imposition of external structure is often exactly what occurs. Distractions are minimized and there are many cues to prompt, support, and direct behaviour. For these reasons, individuals with executive function problems often perform quite well on structured tests while simultaneously demonstrating an inability to function independently in the community. Tests which are more open-ended and which require active problem solving and novel, generative behaviours are most effective in detecting executive function difficulties (Lezak, 1995; Mateer & Sira, 2003b). Some authors state that executive function tests must not only be novel (in content and format; Denkla, 1994) and effortful, but must also involve working memory demands (the ability to maintain information in mind and manipulate it; Denkla, 1994; Phillips, 1997). Another difficulty in the study of executive functioning is the disparity between tasks used for assessment purposes and situations normally encountered by people in their daily lives (Burgess & Robertson, 2002). When assessing planning and organisation, for example, the client may be asked to complete puzzles that require multiple steps, or to “multi-task” (i.e. do more than one task simultaneously). These behaviours may well require certain executive functions to be intact for success, but they do not closely correspond to any real life behaviour with which the client may report problems. Adequate assessment of executive functioning, therefore, requires laboratory based tests, behavioural or functional information obtained through observation of the client’s functioning in daily life, as well as collateral information and questionnaires completed by relatives (Wilson, Evans, Alderman, Burgess & Emslie, 1997).
Assessment of executive functioning is further complicated by the finding that putative executive function tests are not highly intercorrelated (i.e. they have low convergent validity; Basso, Bornstein & Lang, 1999; Baddeley, 1997). This is due, in part, to the fact that executive function tests are impure measures, meaning they require many intact processes for success, including non-executive skills (e.g. posterior functions such as language and perception; Baddeley, 1997; Phillips, 1997). It is therefore important to attempt to rule out alternate hypotheses for executive test failure (such as non-executive memory deficits). In addition, the high level of measurement error makes it difficult to demonstrate divergent validity of executive tests with other task performances. For example, one would not expect a test of planning to be strongly correlated with a test of visual scanning speed, but a high level of measurement error may obscure any observed differences in performance. These factors may also add to the difficulty in demonstrating an association between an individual’s performance on two executive function tests, that intuitively have similar cognitive demands. However, it is important to note that two tasks that appear logically to have similar cognitive demands may in fact be extremely specific and rely on different neuroanatomical areas (e.g. Kimberg and Farrah, 1993). For example different prefrontal cortex areas may perform the same task on differing inputs (Goldman-Rakic, 1987) though most cognitive processes are not sub-served by a single neuro-anatomical area. Finally, one must also be careful to distinguish between anatomical localisation of damage and functional deficit when evaluating executive function deficits of brain-injured individuals (Baddeley, 1997) as executive deficits can be observed with brain injuries that do not appear to involve the frontal lobes.
Intelligence and Executive Function

Another factor that adds to the complexity of assessment of executive function is the intricate relationship between executive function and general intelligence. In order to understand this complexity, a review of the theory of intelligence may be helpful. Spearman (1927) suggested that all intellectual abilities can be represented by a single construct: \( g \). He claimed that cognitive tasks intercorrelate to the extent that they measure \( g \). Based on a factor analysis of existing psychological tests Cattell (1943) introduced the hypothesis that intelligence was made up of two components; Crystallized Intelligence, which includes learned information such as vocabulary knowledge, and Fluid Intelligence, which involves novel problem solving. This hypothesis changed how intelligence was measured, and is reflected in the current properties of the commonly used Wechsler Intelligence Scales, which include both aspects.

Duncan, Burgess and Emslie (1995) demonstrated a relationship between executive function measures and overall intelligence, but they stressed that crystallized and fluid intelligence were differentially related to executive function. These authors asserted that past non significant findings were due to the manner in which intelligence was tested, rather than a lack of relationship between executive function and intelligence. They suggested that while measures of crystallized intelligence, such as vocabulary knowledge, may be unrelated to executive function, measures of fluid intelligence, such as novel problem solving, are strongly related to executive function. In fact, these authors said that \( g \) is, in fact, largely a reflection of executive functioning. They suggested that crystallized intelligence may actually reflect \( g \) at the time of learning the information, but not necessarily at the time of assessment, making measures of crystallized intelligence
relatively insensitive to a subsequent change in $g$. They also noted that the average of several crystallized intelligence scores, such as the Verbal IQ score of the Wechsler tests, correlated better with $g$ than individual “crystallized” subtests. In contrast, measures of fluid intelligence (novel problem solving or spatial tasks) had strong correlations with $g$ even without averaging which they suggested supports the notion that $g$ is akin to fluid intelligence. To test this theory, they looked at the relationship between test scores in five brain-injured participants with high premorbid IQ. They found that in this small sample, fluid reasoning deficits as measured by Cattell’s Culture Fair Test (The Institute for Personality and Ability Testing, 1973) and associated functional impairment, were especially conspicuous despite preserved WAIS IQ scores. Duncan et al., (1995) argued that tests of intelligence and executive tests both assess the ability to formulate the appropriate goal-directed behaviours in novel situations. In a further study, Duncan, Emslie, Williams, Johnson and Freer (1996) found that lesions to the frontal lobe impaired performance on a goal neglect task as well as $g$, and that healthy controls who scored at the low end of the $g$ distribution performed similarly to the brain-injured participants on the goal neglect task.

Others have also found a relationship between conventional executive function measures and intelligence tests (WAIS-R; Wechsler Adult Intelligence Scale Revised, Wechsler, 1981) for neurologically intact individuals (Obonsawin, et al., 2002). The executive tests this group administered were measures of verbal fluency (Benton & Hamsher,1976), the Modified Card Sorting Test (Milner, 1963), the Stroop test (Stroop, 1935), the Tower of London test (Shallice, 1982), Cognitive Estimation (Shoqerat, Mayes, MacDonald, Meudell & Pickering 1990), and the Paced Auditory Serial Addition
task (Gronwall & Wrightson, 1974). They found that performance on all of the executive function tests correlated significantly with performance on the WAIS-R, but that the executive tests correlated only weakly (r=.21-.46) with each other. Interestingly, in healthy controls, the correlations between the executive function tests (except the MCST which was not related to intelligence) were equivalent for Verbal IQ (i.e. crystallized intelligence) and Performance IQ (i.e. fluid intelligence). Obonsawin and colleagues (2002) suggested that one possibility was that some of the shared variance between the executive tests represented shared executive functions, but they said that at least some of the shared variance was accounted for by $g$ given that most of the correlations between the executive tests disappeared when WAIS-R performance was covaried out. However, they did note that $g$ did not account for entire shared variance between the executive tests as some significant correlations remained, and the MCST did not exhibit strong correlation with the WAIS-R.

The discrepant findings for the relationship between different executive tests and intelligence tests is consistent with a recent report by Friedman, Miyake, Corley, Young, DeFries and Hewitt (2006) who argued that not only is intelligence broken into difference facets, but that executive functioning is not a unitary construct. They reported that in healthy participants, fluid reasoning ($G_f$ as measured by the WAIS III Performance IQ; Wechsler, 1997) was significantly correlated to updating (working memory), but was not significantly related to inhibition of a prepotent response or shifting mental set. However, crystallized intelligence ($G_c$ as measured by the WAIS III Verbal IQ) was significantly related to inhibition, updating, and shifting. Therefore they concluded that executive function differentially relates to the different types of intelligence. These findings suggest
that in healthy controls, crystallized intelligence is related to IQ more than fluid intelligence, but the authors noted that in individuals with compromised frontal lobe integrity, they might expect more Gf involvement as the Gc score may be insensitive to the changes in executive function (as suggested above). It is also relevant to note that in the different research articles cited here, the investigators have used different versions of the WAIS. The WAIS III has an additional measure of fluid reasoning (i.e. Matrix Reasoning that is similar to Cattell’s Culture Fair Test) not found on the WAIS-R. As a result of this additional subtest, the WAIS III Performance IQ score is arguably a better measure of Gf than the WAIS-R Performance IQ score.

There are many tests that purport to measure executive function, but some have more ecological validity than others. Wood and Liossi (2007) evaluated the relationship between conventional tests of executive function (Trailmaking B and verbal fluency), ecologically valid tests of executive function (Hayling and Brixton, Zoo Map and Key Search sub-tests from the Behavioural Assessment of the Dysexecutive Syndrome battery; Wilson, Evans, Alderman, Burgess, & Emslie, 1997) and general intelligence in severely brain-injured individuals. Again, these authors reported that the shared variance of executive tests was mostly accounted for by performance on the WAIS III (Full Scale IQ and Performance IQ) as the relationships disappeared when IQ was covaried out. In addition, they found that the various executive tests had low to moderate correlations with each other, and the intercorrelations were similar for conventional and ecologically valid tests (though verbal fluency was related to more ecologically valid test variables than Trailmaking B). They concluded that their findings support Duncan’s hypothesis (Duncan 1995) that aspects of general intelligence are measured by executive function.
tests. A factor analysis showed that performance on executive function tests loaded on two factors (one large factor reflecting \( g \), and one smaller factor reflecting unique executive function variance), which showed the variance on executive tests was not entirely accounted for by intellectual functioning. They suggested the unique executive component of \( g \) was well described by Duncan’s (1995) ”goal neglect”, which as above, has also been observed in healthy individuals under conditions of novelty and weak error feedback.

Despite the inherent difficulties in accurate assessment of executive functioning, clinicians continue to develop ways to assess these processes because executive deficits often lead to the most functional impairment after a brain injury (Sohlberg & Mateer, 2001). Once one has assessed a client’s executive functioning in various ways, then it may be desirable to reassess the same client’s executive skills at a future date (e.g. following a rehabilitation program) to determine if the interventions have had any impact on functional deficits. As above, the types of tasks that best tap executive functions assess how one adapts to novel situations, and the validity of such tasks decreases with every repeated testing occasion (Burgess, 1997). Despite this major limitation, the need to assess change over time is of great interest to researchers and clinicians. It is up to test developers to introduce creative and effective ways to evaluate change in executive functioning over time. There are several important factors to consider with repeated testing and each will be considered in turn.

**Measurement Error**

When making any psychological measurements, one must always have a means to take error into account to determine if differences in scores are abnormal, or reliable
According to Classical Test Theory (Gulliksen, 1950; Lord & Novick, 1968) an individual’s test score is comprised of that individual’s “true” score on the characteristic being measured plus measurement error. A test that has low reliability will have higher measurement error than a perfectly reliable test, which has no measurement error. With a perfectly reliable test, the observed score equals the true score (Anastassi, 1988). Because no measurement tool is perfectly reliable, even if an individual’s true score does not change, his/her observed score may vary from one testing session to the next because of unreliability of the test, as well as chance fluctuations in test administration or testing conditions such as non-standard administration, or examinee fatigue. In other words, if it were possible to administer the same test on multiple occasions without any systematic bias (such as practice effects; see below), then the distribution of scores would form a normal distribution around the individual’s “true” score.

When conducting serial assessments to evaluate change over time, it is important to distinguish between psychometric reliability of a test, and clinical reliability (Matarazzo, Wiens, Matarazzo, & Goldstein, 1974). Matarazzo et al. (1974) defined clinical reliability as the ability to consistently classify an individual’s performance as normal versus impaired based on cut-off scores. This “sensitivity to brain injury” is also known as validity. While this type of reliability may be useful in some situations (i.e. impaired versus normal classifications), it may be adversely affected by many factors, such as the cognitive ability of the examinees, and practice effects (Dikmen, Heaton, Grant & Temkin, 1999). Psychometric reliability of a test can be defined as the extent to which the test is free from measurement error (McCaffrey, Duff, & James Westervelt,
2000), be it systematic or unsystematic (Franklin, 2003). Henceforth the term “reliability” will refer to psychometric reliability.

There are three common types of reliability coefficients: agreement between two raters (inter-rater reliability), agreement among the test items themselves (internal consistency), agreement between test scores obtained on two different testing occasions (test stability or test-retest reliability). Test-retest reliability of a particular test refers to the relationship between a person’s scores across time (Retzlaff & Gibertini, 1994). The higher the correlation between the test scores on the two occasions, the less likely the test scores are influenced by random changes in the examinee’s state, or by changes in the environment (McCaffrey et al., 2000; Payne & Jones, 1957) provided the domain being measured is stable over time and there are no differential effects of being exposed to the test on a prior occasion (Slick, 2006). Importantly, test-retest reliability reflects the relative rankings of the examinee’s performances between the two testing occasions. There may be a large change in raw scores from one testing occasion to the next, but as long as all examinees improve or reduce their scores by the same amount, the test-retest reliability for that test will be high. If the subsequent performances of a test are considerably different from the first administration for different individuals (i.e. some individuals score higher on the task and some individuals score lower on the task so that their relative rankings are unstable), the test-retest reliability will be low. When test scores from one administration to the next vary in different ways for different individuals, the test is not useful as an outcome measure of change regardless of the characteristic it purports to assess. When evaluating the reliability of a neuropsychological test that is intended as an outcome measure, test-retest reliability is the most appropriate reliability
Knowing the test-retest reliability of one’s measurement tools is informative, because it allows one to know how much test scores are being affected by error (McCaffrey et al., 2000). The test-retest reliability of a test is defined by a correlation coefficient between test scores from two separate testing occasions. There has been debate regarding the level of acceptable test-retest reliability, as measured by correlation (Franklin, 2003). Cohen (1988) recommended a reliability index of .80 or above to be considered acceptable. The square of the correlation is the degree of shared variance, so test scores from two testing occasions that have a correlation of .80, have 64% shared variance. This level of acceptance is defensible, if the use of the data will have lifelong consequences for the individual (Franklin, 2003). When the correlation falls below .71, the degree of shared variance is 50%, which is not acceptable.

Another way in which measurement error can influence test scores is through the statistical phenomenon of regression to the mean. Regression to the mean is the tendency for scores to revert (regress) towards the means of the distributions when the test is re-administered (Kazdin, 1992). If an examinee scores above the mean on one testing occasion, he/she would be expected to regress downward towards the mean on a subsequent testing occasion, and an examinee who scores below the mean on the first testing occasion, would be expected to regress upward towards the mean on the second testing occasion. Regression to the mean occurs because on average, an individual’s “true” score is actually closer to the population mean than his/her observed score (Chelune, 2003). The greatest concern regarding regression to the mean is with individual extreme test scores, as they are most likely to have occurred due to measurement error or
a rare event. Tests that have lower test-retest reliability will have greater regression
towards the mean (Kazdin, 1992). When evaluating group test scores (e.g., group mean
and SD) on subsequent administrations, measurement error would be expected to lead to
the same distribution of spurious extreme scores on both occasions, although for different
individuals. Therefore, any observed changes in group test scores (mean, SD) are not
likely due to regression to the mean or measurement error (McCaffrey et al., 2000).
However, when evaluating an individual’s test-retest performance, regression to the mean
is a potential complicating factor. Exactly how this phenomenon impacts clinical
interpretation of test results for individuals, however, is not well understood (McGlinchey
& Jacobson, 1999; Speer, 1999).

Practice effects (i.e. a constant bias leading to improved scores over time;
Chelune, Naugle, Lüders, Sedlak, & Awad, 1993) can also influence test scores on
subsequent testing occasions due to repeated exposure to the test. This is a form of
systematic measurement error. Practice effects are a factor to consider in all cases of
repeated administrations of measurement tools, but there are special issues with
neuropsychological tests. As noted above, practice effects may not reduce the test-retest
reliability of a neuropsychological test, if individuals taking the test all improve by the
same amount, and decreasing correlations between test scores do not necessarily rule out
the possibility of practice effects. It is when individuals obtain differing levels of benefit
from practice, that practice on a test may reduce the reliability of that test. While practice
effects are seen on most neuropsychological tests, the magnitude of the practice effects
varies with the type of measure, test–retest interval, age of examinee, and overall
competency level of the examinee (Dikmen et al., 1999). In addition, specifics of the test
themselves can affect their susceptibility to practice effects.

Neuropsychological tests that have a speeded component, require an unfamiliar behaviour or infrequently used response style, have an easily conceptualized solution or a single solution are most likely to show significant practice effects (Lezak, 1995). Given Lezak’s caveat, practice effects are likely to be particularly troublesome with tests that rely heavily on novelty, such as executive function tasks (Basso et al., 1999; Dikmen et al., 1999). On some tests that purportedly assess executive function, performance improves dramatically once the examinee learns the “technique” to successful performance. The “technique” is obtained through a sudden insight, or discovery of a successful strategy. Tests that have such a “technique”, and tests that have only one correct response, are in effect “one-shot” tests once they have been solved (Lezak, 1995). Clearly, this type of test will show practice effects (Burgess, 1997). What is measured by executive function tests on subsequent examinations may be quite different from what was measured on the initial examination (Basso, et al., 1999). For example, rather than assessing novel problem solving, the test may be measuring the ability to remember a strategy that was successful in the past. Test-retest reliability of measures with large practice effects may be low, because the examinees may perform at ceiling levels on the second testing occasion, effectively restricting the range of test scores (Slick, 2006). Restricted range has been reported to be a factor that reduces test-retest reliability (Dikmen, et al, 1999). Low test-retest reliability coefficients can result from a lack of systematic effects from prior exposure to the test, a nonlinear effect of prior exposure to the test, or a restriction of range that has led to a limit of potential of practice effects (Slick, 2006).
To mitigate the problems associated with repeated testing occasions, the use of alternate forms to assess the same underlying construct has been suggested. Chelune (2003) described two main difficulties with alternate forms, however. The first difficulty is that while alternate forms may reduce the practice effects due to carry over of explicit content, they do not address any procedural learning that is associated with the test (i.e. the format is the same). This issue is particularly problematic for executive function tests, which need to be novel in both content and format to accurately assess novel problem solving (Denkla, 1994; Phillips, 1997). The second problem with alternate forms is that they have to contend with issues of regression to the mean and measurement error for two scales instead of only one (Chelune, 2003). Goldstein, Materson, Cushman, Reda, Freis, and Ramirez (1990) found significant practice effects at follow-up even when alternate forms were used. Other researchers have found that individuals consistently perform higher on the second assessment, no matter which form was administered first (Franzen, Paul & Iverson, 1996). This evidence may suggest that alternate forms do not necessarily reduce practice effects, and that the use of two separate tests may actually make interpretation more difficult due to different propensities for regression to the mean and measurement error (Chelune, 2003). Practice effects are most evident when examinees are exposed to the same test on repeated testing occasions. However, examinees can obtain benefit from practice on different tests as well. Test taking exposure, termed “test sophistication” (Anastassi, 1988) can lead to improved scores on subsequent testing occasions. Therefore, examinees can obtain benefit from exposure to alternate forms of a test (McCaffrey, et al., 2000) as well as from exposure to different tests (Coutts, Lichstein, Bermudez, Daigle, Mann, Charbonnel, Michaud & Williams, 1987) and even
from exposure to non test related tasks such as games (Dirks, 1982).

Information regarding practice effects is not available for many neuropsychological tests though there have been recent efforts to correct this deficiency (e.g. Basso, et al.1999; McCaffrey et al., 2000). Practicing clinicians generally accept that practice effects diminish with longer test-retest intervals, and this belief has been supported in the literature. Catron (1978) and Catron and Thompson (1979) found that as the test-retest interval increased, the correlations between WAIS IQ scores decreased (though again, this does not necessarily imply that practice effects decreased). It is unknown, however, how much time must pass before practice effects become negligible. There is evidence that in healthy adults, practice effects on executive function tests persist after at least 12 months (Basso et al., 1999) and gains on memory tests are maintained after six years (Zelinski & Burnight, 1997).

Researchers have reported differences in practice effects on neuropsychological measures with different samples and different test-retest intervals. Dikmen et al, (1999) found that for tests where a difference was found, younger, better-educated people with good initial competency, or those with a short interval between testing occasions, had bigger improvements in their scores. It is important to note, however, that no one factor, other than initial competency, was associated with differential practice effects across all neuropsychological measures. In addition, Dikmen et al. (1999) found that individuals who initially scored poorly on a test were much more likely to show a large positive change on the subsequent assessment, and individuals who initially scored well had a much smaller improvement, or even showed deterioration. These findings were likely due to regression to the mean and ceiling effects.
Issues in Test-Retest Reliability

The first issue in evaluating the usefulness of a measure’s test-retest reliability is the normative sample used to collect this data. Neuropsychological tasks are generally normed on healthy individuals. This is helpful to determine if a brain-injured person is performing at a level below what would be expected if they did not have a brain injury. However, this procedure does not allow a clinician or researcher to compare a brain-injured individual’s score to other individuals with a similar brain injury. Moreover, if test-retest reliability data is reported in the test manual, it is typically data taken from the healthy normative sample. This procedure precludes a clinician or researcher from calculating whether test results from two different occasions constitute a reliable change (a change over and above what would be expected due to the imprecision of the test instrument plus practice effects; see below) for a brain-injured individual, as there is no test-retest reliability data for an injured sample (Retzlaff & Gibertini, 1994). One cannot assume that a brain-injured individual would experience the same benefit of previous exposure to a test as a neurologically intact individual (Chelune et al., 1993). Executive function tests may have higher test-retest reliability in brain-injured patients than in healthy controls, possibly because the tests are no longer novel to healthy individuals with no memory impairment. However, brain-injured individuals, like elderly individuals and preschoolers, may also have more unstable performance leading to poor test-retest reliability (Slick, 2006). This is why psychometric measures, such as test-retest reliability, taken from controls may be misleading from a clinical point of view. Therefore it is best to select tests intended for repeated assessment that have reliability data for a brain-injured population. An additional problem with the test-retest reliability
data of many current measures of executive function is that often the two testing sessions used to calculate the reliability are relatively close together in time (McCaffrey, et al., 2000; Retzlaff & Gibertini, 1994). A brief test-retest interval may lead to a bias in interpretation in that the positive carryover effects of learning and prior experience may be maximized. This bias may lead to a larger obtained “practice effect” than would be expected in a follow up administration (e.g. after an 8 week interval) of a test such as following a cognitive rehabilitation program (Chelune et al., 1993). However, as noted above, test intervals had a smaller impact on test-retest reliability than examinee characteristics (Dikman, et. al., 1999).

Methodological Procedures to Control for Practice Effects

One can control for the effects of practice by building procedures into the study design. For example, including a control group in data collection that matched the experimental group on demographic factors and test-retest interval can increase one’s confidence in attributing the change scores in the experimental group to the effects of the intervention. In studies where there was no intervention, such as the present study where the goal was to evaluate the reliability of a test (taking into account measurement error and practice effects), a control group allowed for conclusions to be drawn regarding the benefit a brain-injured sample gained from practice as compared to the non-injured sample. Any benefit the control group obtained on the second administration cannot be attributed to recovery from injury, as may have been assumed with the brain-injured group.
Measuring Change on Neuropsychological Tests

There are several ways of measuring change in test scores from two different testing occasions. The most obvious method of measuring change in test scores from two different testing occasions is to simply subtract the score at Time 1 (T1) from the score at Time 2 (T2; the simple differences method). However, this method provides no way to evaluate whether the difference in scores could have occurred due to chance variation in measurement error. Changes in an individual’s test-retest data can also be evaluated by comparing the individual’s retest score to the group pre test mean score, known as the standard deviation method. For example, if an individual’s retest score falls more than one standard deviation below the group pre test mean score (and a lower score denotes a poorer performance), then one could consider the individual’s score to represent a decline in functioning. While the standard deviation method has been used and reported (e.g. Hermann & Wyler, 1988), it has not been used in recent years. This is because the cut-offs were typically chosen clinically (e.g. 1.0 standard deviation below the mean), rather than psychometrically (e.g. as providing the best specificity) and because it does not adequately take into account measurement error or practice effects.

To address some of the limitations of the simple difference score and standard deviation methods of evaluating neuropsychological change, Payne & Jones (1957) outlined two types of statistical procedures to allow clinicians and researchers to control for measurement error: the simple differences and the regression-based approaches. In addition, an empirical way of determining if the change in an individual’s performance on a test at a subsequent administration was reliable and unusual is to compare it to the observed distribution of change scores from a comparable group of individuals, for
example by looking for changes that are larger than those seen for 95% of the sample (Chelune, 2003). With test-retest information regarding change scores from a sample that represents an individual examinee (such as a similarly brain-injured sample), one can estimate the distribution of expected change. There are several procedures to make this comparison, and these procedures are now commonly known as Reliable Change Indices (RCI’s, Jacobson, Follette & Revenstorf, 1984; Jacobson & Truax, 1991). These calculations require the test-retest means, standard deviations and a reliability coefficient. Reliable Change Index calculations use a fixed alpha level (e.g. $\alpha = .05$). To be considered statistically reliable, the RCI must exceed the $Z$ score associated with the predetermined alpha level. With $\alpha = .05$ (two tailed), an RCI must exceed $\pm 1.96$ to be considered a statistically significant improvement or decrement. Any RCI that does exceed this cutoff would represent a statistically reliable change in test score, which could not be accounted for by measurement error 95% of the time. In clinical practice, an RCI of $\pm 1.645 (\alpha = 0.10$, two-tailed) is commonly used, which is a more lenient test.

Reliable change calculations require some standard error to be included in the formula. Standard error is the amount of error inherent in the reliability of a test. Historically, there are four error terms employed for different applications: Standard Error of Measurement (SEM), Standard Error of Estimate (SEE), Standard Error of Difference (SED) and Standard Error of Prediction (SEP; Lord & Novick, 1968). The SEM, SEE, SED and SEP have all been used to control for measurement error (e.g. Payne & Jones, 1957; Chelune, et al., 1993; Hageman & Arrindell). SED is used in the simple differences methods and SEP in regression based methods. There have been numerous reliable change formulae published in the literature (e.g. Bruggemans, Van de
Vijver, & Huysmans, 1997; Chelune, et al., 1993; Hsu, 1989; Jacobson & Truax, 1991; Payne & Jones, 1957; Speer, 1992. Each formula has its own strengths and weaknesses. The simplest formula, known as the simple differences method, was described by Payne and Jones (1957) and popularized by Jacobson and Truax (1991). This RCI formula divides the observed test-retest difference by the SED. In the case of evaluating change in test scores on neuropsychological tests, where one might expect the systematic influence of practice in addition to measurement error, Chelune (et al., 1993) added a correction to the Payne & Jones (1957) / Jacobson & Truax 1991 formula. In addition to accounting for measurement error, this correction accounts for practice effects and consists of subtracting a constant from the observed difference between test scores. The constant is derived from the mean amount of group improvement (or decrement) over a specified interval in the control sample (this information is typically found in a test manual). The Chelune et al. (1993) formula uses the Standard Error of Difference (SED), as does the Payne and Jones and Jacobson and Truax formula. The SED error term can be estimated statistically by multiplying the SEM, an index of dispersion of the obtained scores around and individual’s true score, by $\sqrt{2}$ or it can be estimated empirically as the standard deviation of the difference between the two observed scores in test-retest data sample. The SEM is estimated statistically as:

$$SD_x \sqrt{(1 - r_{xx})} \quad \text{Equation 1}$$

where $SD_x$ is the standard deviation of the measure, and $r_{xx}$ is the reliability coefficient. The SED is computed as:

$$\sqrt{(SEM_1^2 + SEM_2^2)} \quad \text{Equation 2}$$

Therefore, the Chelune et al. (1993) reliable change formula, which uses the SED, can be...
computed as:

\[ RCI = \frac{(X_2 - X_1) - (M_2 - M_1)}{SED} \]  \hspace{1cm} \text{Equation 3}

A further improvement to this reliable change formula was proposed by Basso et al. (1999) who modified the Chelune et al. (1993) formula to use the SEP. The SED assumes that the measurement error between the two testing occasions is uncorrelated, and is intended for use in situations where practice effects are irrelevant (such as attitude scales; Basso et al., 1999). While some argue that the SEP is meant to be used when parallel versions of the test are administered, rather than the same test administered on two occasions (Slick, 2006), a number of authors consider the SEP to be the most appropriate error term to use when the measurement error is expected to be correlated, as it is in a test-retest situation (Brophy, 1986; Charter, 1996; Dudek, 1979). The modified Chelune et al. (1993) formula, using the SEP as the error term rather than the SED (Basso et al., 1999), then appears to be the most appropriate RCI for use in prospective studies that meet the underlying assumptions.

The standard error of prediction (SEP) is computed as:

\[ SEP = SD_{T2} \sqrt{1- r_{x_1 x_2}^2} \]  \hspace{1cm} \text{Equation 4}

where the standard deviation used is from the scores in the second assessment period and the correlation coefficient is the test-retest coefficient between test scores across the two test periods.

The Basso et al. (1999) RCI formula is defined as:

\[ RCI = \frac{(X_2 - X_1) - (M_2 - M_1)}{SEP} \]  \hspace{1cm} \text{Equation 5}

As with other reliable change calculations, if this RCI exceeds ± 1.96, it indicates that the observed Time 2 score falls outside a 95% confidence interval. In this case it is a
confidence interval for the Time 2 score, which is calculated through multiplying the SEP by 1.96 (allowing for 2.5% intervals at both ends of the sampling distribution of change scores) and adding this result to the Time 1 score plus the (M2-M1) practice effect.

Limitations of Reliable Change Formulae

The RCI is misleading when used in the absence of cutoff points for clinically significant change. When used without cutoffs for clinically significant change, the RCI tells one only if the change was statistically reliable not if it was clinically significant (Jacobson et al., 1999). That is, a statistically significant RCI means that zero change is unlikely, but this statistical significance says nothing about how large the change might be, or whether it is large enough to be clinically significant. Another drawback of the RCI formulae is that one of their assumptions is that practice effects are assumed to be the same for all individuals. Because of differences in test-retest intervals, and individual differences in the characteristics of the sample (e.g. age, level of cognitive impairment) this assumption is not likely to be met. In order to meet the underlying assumptions of RCI formulae, one would need to derive an RCI based on a particular sample, and use it only with individuals who match that sample in characteristics and test-retest interval.

Standardized Regression Based Change Scores

McSweeney, Naugle, Chelune, and Lüders (1993) proposed standardised regression based change scores (SRB), which allow clinicians and researchers to take into account measurement error, regression to the mean, differential practice effects and other demographic variables that may affect test performance. By using this method, one can compare an individual’s observed retest score with his/her regression-predicted retest
score (and divide this difference by the SEP of the regression equation). The regression
equation is created through obtaining baseline scores from a group of control participants
and regressing those scores against retest scores with a simple prediction equation. SRB
equations remove variance from the pre test scores from the post test scores. This
process, in effect, makes all individuals who take the test equivalent, even if their
baseline test scores differ (Frerichs, 2003). What makes SRB methods different from
reliable change calculations is that SRB methods also allow one to compare multiple
predictors of retest scores in addition to just using the observed Time 1 test score as a
predictor. If a multiple regression equation is generated (in contrast to the simple
regression equation that only includes the observed Time 1 test score) one can include
other factors that may affect the Time 2 test score. Potential variables that may influence
the Time 2 test scores include age, education, gender, emotional functioning and
personality type. These variables can be included as predictors in the multiple regression
equation, and an expected Time 2 score can be generated. The observed Time 2 score can
then be compared to the expected Time 2 score to determine not only if statistically
significant change has occurred, but also whether the variables entered into the equation
were significant predictors of this change. It can be noted that when there is only one
predictor, Time 1 score, the SRB method is very similar to the Chelune et al. (1993) RCI
(Frerichs, 2003). Both use the SEP, the difference is that in the Chelune method, the
predicted score is the Time 1 score plus practice effect, while in the SRB method, the
predicted score is the regression to the mean corrected Time 1 score plus practice effect.
Limitations of SRBs

While SRB methods provide distinct advantages, there are several limitations to their use. Because SRB uses multiple regression, the assumptions of multiple regression must not be violated. As such, the relationship between the pre test and post test scores should be linear, and there should be homoscedascicity of variance. Finally, the predictor(s) should be measured without error (Pedhazur, 1982). The assumption that the predictor(s) should be measured without error is at odds with the underlying tenet of Classical Test Theory (that all tests have some measurement error). It has been recommended in the literature that SRB methods should not be employed if the obtained data for change are not normally distributed (McSweeney et al., 1993). Moreover, SRB methods are not appropriate for use when the measures have a tendency towards floor or ceiling effects (Frerichs, 2003). The discussion above, regarding neuropsychological tests of executive function, suggests that these measures may be susceptible to ceiling effects (i.e. when the client has learned the “trick” or solution to the test). This may obviate the use of SRBs when evaluating change scores with measures of executive function. A further caution regarding the use of SRBs is their appropriateness for use with a specific individual. If the SRB is applied to individuals who differ in characteristics or scores than the reference sample from which the SRB equation was derived, then the accuracy of change scores may be compromised (Frerichs, 2003). Of course, some of these cautions are equally applicable to the use of Reliable Change Indices. When change is calculated via both SRB and RCI methods (specifically a multiple regression SRB and the Chelune et al. [1993] reliable change formula), the results are similar (Temkin, Heaton, Grant, & Dikmen, 1999). Because SRB methods are much more complicated to use and require a
higher level of statistical knowledge, they may be less appealing to practicing clinicians than RCI methods. Moreover, RCIs provide information regarding whether a clinically significant change from baseline has occurred, as opposed to an index of how much a score fits with established trends in the reference population as in SRBs (Slick, 2006), which may be more relevant for clinical use.

Rationale of the Current Study

The assessment of executive functioning using neuropsychological tests is fraught with difficulties, which are compounded in the case of serial assessments. There are difficulties stemming from the tests themselves, including the need for novelty, propensity towards ceiling effects and “one shot” tests, as well as difficulties in evaluating psychometrically reliable changes in test scores due to measurement error and practice effects. While there are many neuropsychological measures that are designed to measure executive function, many of them have modest reliabilities suggesting executive functioning is difficult to assess reliably (Slick, 2006). The Six Elements Test is one measure that may be amenable to modifications that might make it more appropriate for serial assessments. The current study is an evaluation of a modification of the Six Elements Test in an attempt to make it more suitable for serial assessments.

The Six Elements Task

It has been reported that individuals with executive function problems, despite sometimes average or above average IQ and other cognitive abilities, have difficulty following rules, monitoring the time, and keeping track of task demands (Mateer & Sira, 2003a; Stuss & Benson, 1984). These deficits would be expected to be especially
apparent when individuals are attempting to multitask. As noted above, organisation and multitasking are executive skills that are particularly difficult to measure in the laboratory setting, despite their ubiquitous nature in everyday functioning.

Shallice and Burgess (1991) developed The Six Elements Test (SET) to measure the ability to plan and organize one’s behaviour to reach an externally determined goal according to a set of rules. The test also measures the ability to multitask, as there are six tasks. The SET emphasizes the capacity to form a strategy, monitor one’s own behavior against the main goal, to be aware of time and to switch tasks as required (Chan & Manly, 2002). When administered the SET, the examinee is asked to complete some portion of six tasks (Dictation Parts A and B, Arithmetic Parts A and B, and Picture Naming Parts A and B) within ten minutes. The completion of all six tasks would take longer than the time available, thereby forcing the examinee to leave tasks unfinished in order to go on to the next task. Parts A and B of each task are intended to be equivalent in difficulty, and a rule is imposed constraining the order in which tasks are attempted. The examinee is free to divide the time as he/she sees fit, but is asked not to perform two tasks of the same kind in immediate succession. For example, if the examinee attempts one of the dictation tasks (Part A or B), he/she would have to attempt either an arithmetic task, or a picture naming task immediately afterwards in order not to break this rule. The SET is proposed to be sensitive to attentional allocation strategies that would be required in order to organize performance on the three sets of tasks.

A version of the SET (Burgess, Alderman, Evans, Wilson, Emslie, & Shallice, 1996a) has been included in the Behavioral Assessment of Dysexecutive Function test battery (BADS; Wilson, et al., 1996). The BADS Six Elements Test is the same as the
original, but the scoring procedures have been slightly modified for ease of clinical use
(P. Burgess, 23/10/2003; personal communication). It has an overall profile score (with a
range of 0-4) calculated from combining the number of tasks attempted, the allocation of
time to the tasks and the number of times that the rule is broken. In addition, individual
parameters including the number of task switches, and the deviation in time allocation
from an optimal duration (per task) can be scored.

Support for the validity of the SET has come from several studies. A group of 78
brain-injured patients (M age = 38.8, SD = 15.7), predominantly with head injury,
reportedly performed significantly below the level of the 216 person (M age = 46.6, SD =
19.8) normative sample (Wilson, et al., 1996). The control participants were significantly
older than the patient sample (t = 3.46, p< 0.0001) but there was no difference in
estimated IQ (NART) between the two groups (t = -0.28, p = 0.782). The mean profile
score on the SET (range from 0-4) of the brain-injured patients was 1.99 (SD=1.18) and
the mean profile score of the control group was 3.56 (SD = .78). These scores were
significantly different (t = 10.60, p < 0.0001). The control group’s mean was close to the
ceiling level on the first administration, and overlapped with the ceiling level in terms of
standard deviation. Chan and Manly (2002) administered the SET to 30 patients with
mild to moderate brain injury (as defined by the Glasgow Coma Scale; Teasdale &
Jennett, 1974) and to 68 normal controls. The brain-injured patients obtained a mean
profile score of 2.73 (SD 1.01) which was significantly lower (t= 23.79, p<.001) than the
normal controls’ mean profile score of 3.58 (SD 0.56; range 2– 4). Chan and Manly
(2002) reported that the basis for the brain-injured group’s low mean profile score was
that they attempted fewer tasks (range 0-6) than the control group (control M Tasks
Attempted = 5.68, SD = 0.68; patient \( M \) Tasks Attempted = 4.47, SD = 1.48; \( t = 4.28, p < 0.001 \). The findings suggested that the SET did discriminate between normal controls and mild to moderately brain-injured participants when looking at mean profile scores or number of tasks attempted. It is important to note, however, that when the variability within the groups is considered, the means actually overlapped. Overlap of mean scores is also noted in the Wilson et al., (1996) data. The considerable variability in test scores for both brain-injured and control participants may decrease the SET’s ability to discriminate between groups. Nevertheless, when included in the BADS, the SET was found to be the most sensitive task to differentiate patients with closed head injury (Wilson et al., 1998) and neurological disorders (Burgess, Alderman, Emslie, Evans, & Wilson, 1998) from a non-clinical population.

Wilson et al. (1998) reported psychometric properties of the SET for the normative sample of the BADS, which included both individuals with brain injury and normal controls. For the number of tasks attempted and number of rule breaks, the interrater reliability was perfect \( (r=1.0, p<0.001; \text{Wilson}, 1998) \). However, the correlation between raters was likely so high for these components of the task, because scoring is completed at a very gross level. Test-retest reliability for the SET was evaluated with 29 normal control subjects at a six to twelve month interval and was found to be moderate \( (r= 0.33, p= 0.78) \) according to Cohen’s criteria (Cohen, 1992). Despite this relatively low correlation between scores, the normal controls' scores from the first to second administrations were not found to significantly differ \( (p = .264) \). The percentage agreement, an alternative method of determining test-retest reliability that reflects the same score being achieved on both occasions in the control sample, was found to be 55%
for the SET. The control participants' mean profile score (range from 0 to 4) at the first administration was 3.41 (SD= 0.91) and their mean profile score on the second administration was 3.62 (SD= 0.78). Taking into account the standard deviation of scores, the controls’ scores were approaching ceiling levels (i.e. a profile score of 4) at the first administration. It is possible that there was little “room for improvement” in scores, resulting in no significant differences in test scores from the two testing occasions (Jelicic, Henquet, Derix & Jolles, 2001). Moreover, the observed small improvement in profile scores on the second testing occasion could be due to measurement error rather than a systematic practice effect. If the SET had greater range, a greater difference in mean scores from one testing occasion to the next may have been observed.

Notwithstanding these caveats, it does appear that the control participants performed slightly, if nonsignificantly, better on the second testing occasion. Because the data that contributed to the profile scores was not provided, the manner in which the controls improved their scores (i.e. attempting more tasks, not breaking the rule of the task, or not spending too much time on one task) is unknown.

Test-retest reliability data for the SET was reported for a small sample (N= 22) of psychiatric patients with a variety of diagnoses (e.g. schizophrenia, affective disorders, anxiety disorders; Jelicic, et al., 2001). The mean age of this sample was 40 years (SD =8.8), and the mean IQ was 101 (SD =16.7). Participants were reported to be “in remission” with respect to their psychiatric symptoms. This sample of psychiatric patients was administered the BADS on two occasions with a three-week interval between them. The SET was found to have moderate test-retest reliability (r= 0.48) according to criteria set by Cohen (1988). The mean profile score (range from 0 to 4) of this sample on the
first administration was 3.00 (SD= 1.31) and their mean profile score on the second administration was 3.27 (SD= 0.94), again demonstrating some improvement potentially due to practice. If one compares this data to the data reported by Wilson et al. (1998) where control participants obtained a profile score of 3.41 (SD= 0.91) on the first testing occasion, then it is obvious that the profile scores obtained by this psychiatric sample who, “as a group, demonstrated poor executive functioning” (Jelicic, et al., 2001, pp.77) had very similar scores to the normal controls. Indeed, the two samples’ scores overlap when the standard deviations are considered. The overlap of the distribution of profile scores for brain-injured and control examinees is troubling. It may be a function of restricted range in scores, but may also suggest that while the profile score of the SET is specific (i.e. unlikely to misclassify a neurologically intact examinee as brain injured), it may not be particularly sensitive, and might not be a useful tool in the assessment of mild impairment in executive function in all cases.

When considering reasons for task success or failure on the SET, it can be helpful to assess qualitative information regarding the examinee’s test taking strategy. Van Beilen, Withaar, van Zomeren, van den Bosch and Bouma (2006) found when individuals with a diagnosis of schizophrenia (single or multiple episode) were administered the SET, they demonstrated a qualitatively “deviant strategy”. Specifically, the participants with schizophrenia tended to complete the subtasks of the test using an item-by-item approach, switching tasks constantly during the ten minutes (dubbed “Continuous Switching”). One might assume, as these authors did, that the Continuous Switching strategy requires less prospective and/or working memory and attentional resources, and therefore, less cognitive effort. To evaluate test taking strategy, the authors administered
the SET and several other cognitive tests to four experimental groups of participants: 60 patients with schizophrenia, 30 healthy controls, 25 patients with a closed head injury, and 25 patients who had sustained peripheral injuries. There was a statistically significant difference between the number of subtasks attempted between the healthy controls and the patients with schizophrenia. However, as the patients with schizophrenia attempted 5.2 (SD=1.3) tasks, and the healthy controls attempted 5.9 (SD=0.3) tasks, it is questionable whether the difference is clinical meaningful despite the statistically significant difference. Again, because of restricted range in test scores, the distributions overlap when the standard deviations are taken into account. Despite the small clinical difference in overall test scores, the patients with schizophrenia were far more likely to use the Continuous Switching strategy than the other three groups. One third of the patients with schizophrenia used this strategy, but it was rare in the other groups. The authors also reported that the patients with schizophrenia scored significantly below the levels of the healthy control group on all other cognitive measures. In addition, those within the psychiatric group who used the Continuous Switching strategy performed significantly more poorly on the verbal memory and perceptual sensitivity tests than those who did not use that strategy. It was suggested that the Continuous Switching strategy may represent a compensatory strategy on the examinee’s part. Unfortunately, these authors did not report the cognitive results for the other tests for the head injured and peripherally injured participants. Therefore it is difficult to determine if this strategy was “schizophrenia specific”. They did suggest that the more typical strategy adopted by the other participants in their sample (i.e. dividing performance on the six tasks over the ten minutes) might reflect a more abstract interpretation of the test instructions, and that
patients who tend to interpret events more concretely might be more likely to use the Continuous Switching strategy. It remains to be seen whether or not other samples of patients with schizophrenia, or different populations evidence this atypical strategy on the SET.

Limitations of the Six Elements Test

The SET is proposed to be a test of planning, self-monitoring and organisational deficits associated with moderate executive dysfunction (Wilson, et al., 1998). While the test may be sensitive to mild executive deficits from a qualitative perspective, the restriction of possible scores may obscure the differences between brain-injured and control groups (Jelicic, et al., 2001). As with all executive function tasks, the SET relies heavily on novelty for accurate assessment and may be especially vulnerable to practice effects. As noted above, the SET has restricted range in possible profile scores (i.e. 0 to 4). As there appear to be small, but apparent, practice effects from one test administration to the next, examinees who were impaired on the first testing occasion may reach ceiling levels by the second or third administration. Once an examinee reaches the ceiling level on a test, it is no longer possible to assess changes in performance with that particular test. For example, a brain-injured examinee may be retested on the same measures before rehabilitation, after completing the rehabilitation program to quantify improvement, and then again at a later date to determine if the gains made through rehabilitation had a lasting effect, or if the examinee had made further gains through generalization of skills in real life settings. If the examinee reached the ceiling level of the test on the second testing occasion as a result of practice, the test would be insensitive to further gains. While this concern may be applied to many neuropsychological tests, it is especially
apparent on tests with restricted range that show a practice effect.

Despite two potential psychometric weaknesses (limited range of possible scores, and questionable sensitivity to mild executive dysfunction) the concept of the SET brings together many useful components of an ecologically valid measure of executive dysfunction. In order to address some of the limitations of the currently published SET, while preserving the concept of a measure of planning and organisation, it may be constructive to modify the SET to increase the difficulty of the test in order to make it more sensitive to mild executive dysfunction, and to increase the range of possible scores to prevent ceiling effects. One way to achieve these goals may be to add more tasks to the SET (to increase the range of possible scores) while simultaneously increasing the time demands. By increasing the range of possible scores and the difficulty of the task (through increased time demands), it may be possible to improve the sensitivity of this task to mild impairment in executive functioning while still maintaining an acceptable level of specificity.

To increase the range of the existing SET, the number of tasks was increased from three tasks with two parts each, to six tasks with two parts each for a total of twelve tasks (with a range of possible scores from 0-12). The two parts of each task are parallel forms equated for difficulty. For simplicity, the new task was called “The Twelve Elements Test” (12ET; see Figure 1). Twelve tasks were determined through pilot work to be suitable to ensure the test would not suffer ceiling effects when administered to non-clinical population. It is important to note that none of the tasks on the SET is difficult and all are well within the cognitive capabilities of most individuals. The primary difficulty associated with the SET is the perception that there is too much to do within the
time limit. In actuality, there is ample time to attempt all the tasks, as only one item on each task must be done for that task to be deemed as “attempted”. The test only becomes difficult if the examinee spends too much time on one or two tasks, to the detriment to attempting any other tasks. The examinee may then not have adequate time to attempt all of the tasks. The time limit for the original SET is ten minutes. To increase the (perceived) difficulty of the 12ET, the time limit was set to 15 minutes. Therefore, despite the fact that the number of tasks was doubled the examinees were told they must attempt as many of the 12 tasks as they could within only 15 minutes.

Figure 1 The Twelve Elements Test

Personality and the Twelve Elements Test

When administering the 12ET to healthy controls in pilot work, it was noted that a
fairly large number of participants attempted relatively few tasks. The participants who attempted few tasks, also tended to be the same participants who completed several tasks. Because the tasks were designed to take longer to complete that the time allotted, completing several tasks would lead to insufficient time for the overall test. Following completion of the 12ET, these participants were queried on their behaviour, given that their performance appeared to “go against” the stated intentions of the task (i.e. to attempt as many tasks as they could). The participants made comments such as “I like to organise things and the card sorting was fun”, and “I had to finish ______ because I couldn’t leave it half done”. These responses suggest that personality factors may have had an unanticipated influence on these participants’ test results. It would have been informative to administer a brief personality inventory to this sample, to control for personality variables that affect test performance. This limitation was rectified in the current study by including the NEO Five Factor Inventory (NEO FFI; Costa & McCrae, 1992), which is a short form of the 240-item NEO Personality Inventory. In particular, the Conscientiousness domain was of interest.

Conscientiousness is one of the “super-traits” of the Revised NEO-PI and is comprised of six primary traits: competence (C1), order (organization and preference for structure) (C2), dutifulness (C3), achievement striving (C4), self-discipline (C5), and deliberation (C6). Costa and McCrae (1992) described the Conscientiousness domain as one that reflects an individual’s ability to manage his/her desires, and to resist impulses and temptations. This domain also reflects self-control as it relates to a more active process of planning, organizing and carrying out tasks towards goal completion. These authors describe the Conscientious individual as one who “is purposeful, strong willed...
and determined, and probably few people become great musicians or athletes without a reasonably high level of this trait” (Costa & McCrae, 1992, pp. 16). Individuals who score high on the Conscientiousness trait, may also be more likely to prefer completing individual subtasks on the Twelve Elements Test than individuals who do not score high on this trait. In addition, individuals who score high on this trait may be more likely to do the Twelve Elements Test “correctly” and achieve the goal of attempting all twelve tasks.

Assessment of Executive Function Through Questionnaires

While traditional neuropsychological measures are often employed to assess executive dysfunctions, they are not the sole way to gather information about these disabilities. An alternate way of measuring executive dysfunction is through the use of questionnaires that list problems commonly associated with executive dysfunction. These questionnaires are typically completed by the brain-injured individual, as well as by another person who knows them well to gather collateral information. The Dysexecutive Questionnaire (DEX; Burgess, Alderman, Wilson, Evans, & Emslie, 1996b) that is part of the BADS (Wilson et al., 1996) is such a questionnaire. It is a 20-item checklist that individuals and informants use to rate the frequency of occurrence of dysexecutive behaviours and characteristics (e.g. impulsivity and difficulty planning). The design of the DEX is based on the common symptoms of executive dysfunction, as classified by Stuss and Benson (1984), and includes emotional, motivational, behavioural, and cognitive aspects. It gives a comprehensive picture of commonly reported symptoms that are described in layman’s terms, and it is relatively easy to complete (Chan & Manly, 2002).
The Current Study

This study has three main sections, and the hypotheses will be described for each section.

Part 1. Evaluating the Validity of the Twelve Elements Test

This study describes the development of the Twelve Elements Test (12ET), and provides data on its sensitivity to brain injury. The sensitivity of the 12ET was evaluated by comparing the performance of the two groups (one group made up of mildly, moderately and severely brain-injured individuals and another group of matched neurologically intact controls). It was hypothesized that the brain-injured group would perform more poorly on the 12ET variables than the control participants.

Part 2. Convergent and Divergent Validity of the Twelve Elements Test

The external (convergent and divergent) validity of the 12ET was evaluated by correlating the 12ET with other neuropsychological and self-report measures to provide further evidence of the clinical utility of the 12ET. The putative executive function tests were expected to correlate with different indices on the 12ET.

Convergent Validity

*The 12ET and Intelligence*

The Wechsler Test of Adult Reading (WTAR) was developed as a brief tool to estimate premorbid intelligence, and was co-normed with other well-known Wechsler scales of intelligence and memory. The test is based on a recognition-reading paradigm where the examinee reads aloud irregular words (i.e. those that cannot be “sounded out”),
but does not have to provide the meaning of the words. In keeping with past findings that
general intelligence is strongly related to performance on executive function tests
(Obonsawin, Crawford, Page, Chalmers, Cochrane, & Lowb, 2002) the Wechsler Test of
Adult Reading (WTAR; Psychological Corporation, 2001) was expected to correlate with
all of the neuropsychological tests. However, the WTAR was expected to have lower
correlations with the putative executive function tests than the executive function tests
did with each other.

The 12ET and a Measure of Time Estimation

Time Estimation from the BADS battery (a Temporal Judgment Test) was
included to evaluate participants’ ability to gauge the passage of time and make educated
guesses. Time Estimation was expected to correlate negatively with the number of times
participants requested the amount of time remaining on the 12ET, as those individuals
who were able to accurately track the passage of time would have less need to request
this information. Time Estimation scores were also expected to be negatively related to
the number of times a participant spent longer than 7 consecutive minutes on one 12ET
subtask (which is considered a prospective memory failure on the SET; P.W. Burgess,
23/10/03, personal communication) because those individuals with good time estimation
skills would be unlikely to spend longer than 7 minutes on one task.

The 12ET and Measures of Cognitive Flexibility

The Wisconsin Card Sorting Test (WCST; Heaton, Chelune, Talley, Kay, &
Curtiss, 1993) was included as a measure of cognitive flexibility and perseverative
tendency. The WCST appears to measure general reasoning ability that is related to fluid
intelligence (Salthouse, Atkinson, & Berish, 2003: see below for further discussion). The WCST loads onto a separate factor than other executive function tests, and appears to rely on the dorsolateral prefrontal cortex (Macpherson, Phillips & Della Sala, 2002). The Ruff Figural Fluency (RFF; a measure of generative ability, strategic planning and self monitoring) was also included (Ruff, Light & Evans, 1987). The RFF has been shown to be differentially impaired in individuals with right frontal lesions as opposed to left frontal lesions (Ruff, Allen, Farrow, & Niemann, 1994). The RFF was chosen in preference over a verbal fluency task as it requires the generation of novel figures, rather than retrieval from long term memory, so it may be less susceptible to strategies that the participants may have used in the past (Phillips, 1997). In addition, as figural fluency may be a more novel task, it may place greater demands on executive function (Suchy, Sands & Chelune, 2003). The Wisconsin Card Sorting Test Categories achieved, Trailmaking Part B score (a measure of mental flexibility), and Ruff Figural Fluency Number of Unique Designs, were expected to correlate positively with the Number of Tasks Attempted on the 12ET.

The 12ET and a Measure of Planning

The battery also included The Tower of London Drexel version (TOL; Culbertson & Zilmer, 2001) as a test of planning. While TOL does not necessarily elicit planning, as each puzzle can be solved by a trial and error approach (because individual moves can be reversed; Goel & Grafman, 1995), it may reflect planning ability and is a useful measure of impulsive responding. Rule Violations on the TOL have been found to correlate with other executive measures in an adult ADHD population (Riccio, Wolfe, Romine, Davis & Sullivan, 2004). The 12ET Number of Tasks Attempted variable was expected to
significantly correlate with the Tower of London. That is individuals who completed the Tower of London items in fewer moves were expected to attempt more tasks on the 12ET.

*The 12ET and a Measure of Self Monitoring*

CVLT II Repetitions (a proxy measure of self monitoring) was expected to correlate negatively with Number of Tasks Attempted on the 12ET (i.e. an individual who repeated herself frequently was expected to attempt fewer tasks).

*The 12ET and Self and Informant Reports*

For the brain-injured participants, The DEX Self Report and DEX Informant Report were not expected to correlate highly with each other, nor was the DEX Self Report expected to correlate well with performance on the 12ET or other executive function measures. This was expected as previous research has found the predictive value of brain-injured patient's self reports is very poor (Wilson, et al., 1996). This issue is an overarching problem with self-report data from patients with executive dysfunction where they frequently over rate their own abilities, and underestimate their deficits on self-report measures (Prigatano, 1999). The DEX Self Report, while not necessarily a good outcome measure, is a useful instrument as it quantifies the patients’ perception of their disability. When compared to the informant questionnaire, the self report measure quantifies the brain-injured participants’ level of awareness, which is related to the severity of their brain injury. The difference score between the DEX client version and the informant version for the brain-injured group was expected to correlate positively with performance on the neuropsychological tests.
A moderate correlation ($r \geq 0.3$; Cohen, 1988) between the 12ET, the other neuropsychological tests and DEX Informant Report at week one, would suggest that these measures are tapping into similar constructs when the neuropsychological measures are novel tasks.

**The 12ET and Personality**

The Conscientiousness factor of the NEO-FFI was hypothesized to positively correlate with the Number of Tasks Completed on the 12ET.

**Divergent Validity**

**The 12ET and Visual Scanning Speed**

Visual scanning speed from Trailmaking Part A was expected to have low correlations with the 12ET variables, to provide evidence of divergent validity.

**Part 3 Reliability and Normative Data of the 12ET**

In Part 3, the test-retest reliability of the 12ET was obtained by administering it to all participants on two occasions eight weeks apart. Multiple regressions were completed for the brain-injured group and matched control groups to determine if group membership, intelligence, or memory significantly predicted scores on the second testing occasion, over and above the predictive power of scores on the first testing occasion. To add to the clinical utility of the 12ET, normative data from the both groups group were provided. The normative data reported were stratified by any significant predictor demographic variables. This exploratory normative study was intended to determine the clinical utility of the 12ET as an outcome measure for a future executive function
rehabilitation effectiveness study. Through this study, the data (reliability indices, means and standard deviations) needed to calculate whether changes observed in serial assessments are reliable for the 12ET in these populations over an eight week period are presented.
METHOD

This study received ethical approval from the University of Victoria Office of the Vice-President, Research Human Ethics Committee. All participants gave their informed written consent to be in the study (see Appendix C) on the first testing occasion. Verbal consent to continue in the study was obtained from all participants when they were contacted to confirm the second testing occasion. One brain-injured participant declined to participate in the study after reading the consent form on the first testing occasion. Another brain-injured participant declined to complete several tasks on the first testing occasion, and declined to participate on the second testing occasion. This latter participant’s data was excluded from the analysis because of missing data.

Participants

Data were collected from two groups of participants: a mixed acquired brain-injured group, and a matched healthy control group. According to an a priori power analysis using G Power (Buchner, Faul & Erdfelder, 1997) for two groups, to detect a large (i.e. clinically meaningful) effect size $d = .80$ with 80% power, at $\alpha = .05$, 26 participants needed to be included in each group. Data from 28 brain-injured participants, and 32 healthy controls were collected in case of attrition. Participants who were dropped from the analysis for any reason were replaced if the sample size fell below the minimum number required for adequate power. The brain-injured group was composed of 15 men and 11 women with a mean age of 48.6 years (SD = 8.9). The matched control group (see below for matching criteria) was made up of 13 men and 13 women with a mean age of 43.6 years (SD = 12.3). The control group was younger than the brain-injured group, but
the difference was not significant ($t = -1.69, p= .098$). The brain-injured group had an average of 13.3 years of education, whereas the control group had 14.5 years of education on average. As with age, this difference was not significant ($t = 1.39, p= .172$).

The brain-injured participants were recruited through posters and advertisements. They came from several sources: the pool of patients who had completed their neuro-rehabilitation program at the Gorge Road Hospital; individuals associated with the Victoria Head Injury Society; individuals associated with the Southern Alberta Brain Injury Society; and the general community of Victoria BC. Brain-injured individuals with a mild, moderate, or severe brain injury, as defined by their self reported duration of post traumatic amnesia (PTA), were invited to participate.

Post traumatic amnesia is defined as the interval from the onset of the brain injury, until the time when a patient can give a “clear and consecutive account” of what is going on around him / her (Symonds & Russell, 1943). It has been further emphasized that for a patient to be out of PTA, he / she must demonstrate the ability to form continuous memories (Russell & Nathan, 1946). Longer duration of PTA predicts worse outcome after brain injury (Bishara, Partridge, Godfrey, Knight, 1992). Retrospective report of PTA has been found to be strongly correlated ($r = 0.87, p<0.001$) with prospective measures of PTA in a group of closed head injury patients (McMillan, Jongen, & Greenwood, 1996) therefore the brain-injured participants’ self report is considered a reliable estimate regarding the duration of PTA. Duration of PTA was divided into four categories based on Russell and Smith’s (1961) categories. Russell and Smith suggested rating the brain injury associated with a PTA duration of less than 1 hour as “mild”; from 1-24 hours as “moderate”; from 1-7 days as “severe”; and more
than 7 days as “very severe”. Of the 17 traumatically brain-injured participants, two had a mild injury, two had a moderate injury, one had a severe injury, and 12 had a very severe injury. Nine participants in this study had an acquired brain injury (e.g. stroke, tumour or aneurysm) rather than a traumatic brain injury. These participants often did not report a significant duration of post traumatic amnesia, despite an obviously severely brain injury. For these participants, injury severity cannot be accurately classified by duration of PTA.

The following inclusion criteria were used: brain-injured individuals between the ages of 20 and 65 who were at least one year post injury, who were not actively engaged in any formal rehabilitation program, who had no history of drug or alcohol abuse (as defined by the Diagnostic and Statistical Manual IV-TR, APA, 2000), and who had no speech (e.g. aphasia), motor (e.g. paralysis or plegia of both upper limbs) or perceptual deficit (e.g. blindness, neglect, deafness) likely to interfere with completion of the tasks. Participants were screened in the initial phone contact to exclude participants with any of these deficits, and were screened again prior to the first testing session to ensure they were able to complete the tasks. Five potential brain-injured participants were excluded in the initial phone contact because of motor, speech, and/or perceptual deficits that would have interfered with completion of the tasks. These participants would have been over and above the required 26 needed for adequate power.

As employed in past research (e.g. Dikmen et al., 1999), the healthy control group was recruited from friends or family members of the brain-injured participants who matched the participant (or another brain-injured participant) on age, gender, and years of education. Peers or family members were chosen as the control group for several reasons; these participants were likely to be similar to the brain-injured participants on many
variables, including socioeconomic status, education, age, gender, drug use, and risk taking behaviour. Peers or family members were preferable to an orthopaedic control group, for example, because they have not experienced a trauma. While a traumatic event may provide potentially useful experiences to match the groups on (such as dealing with a chronic injury/illness, hospital admission and psychological distress), an orthopaedically injured control group has a number of potentially confounding variables. Othopaedic patients may be engaged in physical rehabilitation and still may experience chronic pain 1 year post injury, which may be different from the brain-injured group (though it is certainly possible that the brain injured participants would also have residual pain from their injuries). Most importantly, by excluding any individuals who have experienced a physical trauma, the possibility of the control participants having had an undiagnosed brain injury is minimized (S. Dikmen, personal communication, 27/04/04).

Peers and family members are an appropriate comparison group for a test evaluation study, as it is necessary to obtain information of how a “normal” group performs on the 12ET when establishing the normative database. Finally, this group is preferable because they are a convenient source of potential participants. If brain-injured participants did not have a peer willing to participate, a participant was recruited from the community to serve as a matched control. Fourteen control participants were recruited from the community, and twelve were volunteers because of their relationship with one of the brain-injured participants. These control participants may be different from the family member volunteers as they are “self selected” rather than selected by nature of their relationship to the brain injured person.

In addition to matching the brain-injured participants on the above variables,
inclusion criteria for the normal controls were: no history of head injury or neurological illness, no history of physical trauma (i.e. an accident that necessitated a hospital admission), no history of drug or alcohol abuse (as defined by the Diagnostic and Statistical Manual IV-TR, American Psychiatric Association, 2000), and no speech, motor or perceptual deficit likely to interfere with completion of the tasks. The control participants were screened in the initial phone interview for any deficits that would interfere with testing.

Measures

Participants were not excluded if they had a history of drug and/or alcohol use, and/or were using drugs and/or alcohol at the time of their participation in the study (though current or past history of drug abuse was an exclusionary criteria). Each participant’s past and current drug and/or alcohol use was assessed by the Severity of Dependence Scale (SDS; Gossop, Darke, Griffiths, Hando, Powis, Hall, & Strang, 1995). The SDS is a short and easily administered 5-item scale that can be used to measure the psychological dependence experienced by users of different drugs (See Appendix D). Each item is rated on a four-point likert scale and the ratings are summed (range 0-15). The psychometric properties of the SDS have been reported in a study of five samples (Total N = 1312). The internal consistency of the SDS was found to range from 0.8 to 0.9 for all of the samples. Test-retest reliability was found to be high (r = 0.89) when the SDS was administered to 100 young adults (mean age 31.9 years) who were in treatment for opiate dependence. The content validity of the SDS was reported by Gossop et al. (1995) for all five samples. The five drug using participant groups of the Gossop et al., (1995) study had an overall mean score of 5.22 on the SDS (range 3.7 – 8.7). Therefore, if the
participants of the present study obtained a score above 6 on the SDS for past or current drug use (they completed the questionnaire twice in the initial assessment in consideration of both time frames), they were excluded from the analysis. Two brain-injured participants and one healthy control participant were excluded from the analysis because of past drug abuse. It could be argued that those with pre-morbid substance abuse histories should be included in the brain-injured group because they represent a substantial proportion of individuals at risk for TBI (Kreutzer & Harris, 1990; Corrigan, 1995). However, the two brain-injured participants who endorsed past substance abuse scored significantly lower than the rest of the brain-injured sample across all measures, including the IQ estimate. Because they did not appear to have come from the same population as the rest of the brain-injured sample, they were excluded from further analysis. One control participant, who scored 12 on the SDS for past drug abuse, was kept in the analysis because a suitable replacement could not be recruited. This participant was not abusing drugs at the time of the study, and performed as well as, or better than, the average scores of the other control participants on most measures. However, his verbal memory was below average, and his performances on Trailmaking A and Trailmaking B were beneath expected levels.

The brain-injured participants had a variety of cognitive deficits, including executive dysfunction, problems in attention, memory, language, and/or visuospatial skills. Brain-injured participants’ cognitive deficits were due to acquired brain injury (ABI) as defined by self and/or informant report. If a participant was described to suffer from cognitive difficulties in daily life by their health care workers and/or family members, for example an inability to perform a sequenced unstructured task such as
making dinner (and they were reportedly able to do so prior to their injury) then they were included in the present study even in the absence of self reported deficits. No brain-injured participants received formal rehabilitative cognitive therapy during the data collection period.

Demographic variables (gender, age, education) were collected for descriptive and matching purposes. The type of brain injury, severity of the brain injury defined by length of post traumatic amnesia, time since injury, whether or not the participants engaged in formal cognitive rehabilitation and current medications were also recorded.

Mood was assessed to exclude potential participants whose scores on the neuropsychological tests may have been reduced by factors other than acquired brain injury. Mood was formally assessed on the initial testing occasion, and assessed informally on the second testing occasion to ensure there were no significant changes in mood in the 8-week interval. On the first testing occasion, the participants’ current mood was assessed with the Beck Depression Inventory II (BDI-II; Beck, Steer, & Brown, 1996). This is a face valid 21-item scale of symptoms associated with depressed mood with higher scores indicating more severely depressed mood. Cut scores have been developed for use with the BDI–II, and are reported in the manual (Beck, et al., 1996). According to the manual, a score between 0-13 is considered minimal depression, a score of 14-19 is considered mild depression, a score of 20-28 is considered moderate depression and a score of 29-63 is considered severe depression. These cut-offs were set so that this test would be as sensitive as possible for use as a screening instrument. To reduce the number of false positives in research studies, the authors recommend using a cut-off score of 17. This score yielded a 93% true positive rate and an 18% false positive
rate for the presence of Major Depression in a clinical sample (N=127). Steer, Brown, Beck and Sanderson (2001) found that 35 individuals diagnosed with a mild DSM-IV Major Depressive Episode had a mean score of 18 on the BDI-II. Therefore, it would be ideal to exclude participants who scored above 17, to control for the negative cognitive effects of depression. Three brain-injured participants had BDI II scores of 19, 20, and 21 respectively. Because of participant attrition for other reasons (described above) these three participants’ data were not excluded so that the minimum sample size could be maintained. These three mildly depressed participants’ average attention and memory scores were not significantly different from the average scores of the rest of the brain-injured sample (Digit Span t = -.794, p = .44, d = .22; CVLT II t = 1.76, p = .09, d = .49). Including the three mildly depressed brain-injured participants, the brain-injured group had an average score of 9.7 (6.1) on the BDI (range 0-21) with a median score of 10. The average BDI II score for the matched control group was 6.7 (5.1) with a range of 0-17 and a median of 6. No matched control participants were excluded because of depressed mood. With inclusion of the three mildly depressed brain-injured participants, the brain-injured and matched control groups did not significantly differ on the BDI II (t = -1.87, p = .07, d = .52) though the brain-injured group did endorse more items overall on the BDI II than the control participants. Further evidence that the two groups did not differ in terms of psychosocial functioning (i.e. anxiety and/or generalized negative affectivity) was obtained through the NEO-FFI. The brain-injured participants did not have significantly higher scores (M = 51.6) on the Neuroticism factor than the matched control participants (M = 47.5; t = -1.36, p = .18, d = .38).

To control for interference on the neuropsychological tests as a result of pain,
participants were asked to rate their pain at the time of the assessment. Several participants reported pain (as measured on a 1 – 7 self report scale with 1= no pain and 7 = the worst pain they can imagine). The mean level of pain reported by the brain-injured group was 1.15 (1.8) with a range of 0-7 and a median of 0. The brain-injured participant who reported pain levels of 7 stated that her pain was well controlled by medication, and her pain did not affect her ability to complete cognitive tasks. This participant, however, was one of the brain-injured participants to score above 17 on the BDI II, and under ideal circumstances her data would have been excluded. Again, because of participant attrition, her data were retained. For the control participants, mean pain levels were 0.5 (.99) with a range of 0-3, and a median of 0. The groups did not significantly differ on self reported pain (t = 1.58, p = .12, d = .44).

Participants in both groups were taking a number of different medications. Participants were instructed to maintain their usual medication regime, in order to maximize ecological validity of the test data. While it is known that some pain medications used by the study participants may affect cognitive functioning, it was reasoned that these individuals would be affected by their medications during the study to a similar degree as they were in their daily life, thereby increasing the generalizability of the results. Three participants in the brain-injured group were taking medications that may have affected their cognitive functioning including Tylenol 3, Dilantin, Gabapentin, diazepam and trazadone. Of the matched control group, two participants were taking medications that may have affected their cognitive functioning including tomazopan and rivatrol.
Procedure

To protect the participants’ privacy, all participants were assigned a number, and only their number was recorded on test forms. All participants were assessed on two occasions. The first administration date was denoted Time 1, and the second administration, completed after approximately eight weeks, was denoted Time 2. For the matched participants who knew each other well, the control participants also completed the Informant Rating of the DEX for the brain-injured participant both testing occasions. Brain-injured participants who were not accompanied by a control participant were asked to designate a person who knew them well who could complete the Informant Rating of the DEX. Most brain-injured participants took a copy of the DEX for their designated informant (usually a spouse) to complete in private, and their informants mailed the scale back to the researcher in a self addressed stamped envelope. For two brain-injured participants, the Informant Rating Scale of the DEX was completed by telephone.

All participants completed the same neuropsychological measures, including the 12ET, on both testing occasions. To minimize the potential effect of bias, a research assistant and the principle investigator each completed approximately half of the assessments in each group. Where possible, the same examiner administered and scored the neuropsychological battery for both testing occasions for individual participants to maximize internal consistency of the assessments. The research assistant was paid through small grants from the Michael Smith Foundation for Health Research and Canadian Institute for Health Research, and the Sara Spencer Foundation.

The neuropsychological battery in this study included measures that purport to assess the hypothesized functions required for successful completion of the 12ET (i.e.
planning, cognitive flexibility, time estimation, self monitoring, working memory and verbal memory) and a measure to assess divergent validity (visual scanning speed). All published measures were administered according to manual based instructions. To evaluate the contribution of general intelligence to successful 12ET completion, the battery included an estimate of intelligence from the Wechsler Test of Adult Reading (WTAR; Psychological Corporation, 2001). This also allowed for examination of whether general intelligence moderated the magnitude of practice effects.

The split half reliability of the WTAR was determined for the standardization sample (N= 1,134) and reported by age group. The reliability coefficients ranged from 0.90 to 0.97 with an average r=0.93. To assess test-retest reliability, the WTAR was administered to 319 examinees on two occasions with intervals ranging from 2 to 12 weeks (with an average interval of 35 days). Across the age groups, the test-retest reliability coefficients ranged from 0.90 to 0.94. The concurrent validity of the WTAR with other measures of reading recognition was high, with correlations ranging from 0.73 to 0.90. The concurrent validity between the WTAR and the Wechsler Adult Intelligence Scale-III over the total standardization sample was 0.73 for Full Scale IQ, and 0.75 for Verbal IQ. Divergent validity was demonstrated by correlating the WTAR with Wechsler Memory Scale III scores (r= 0.49 for General Memory). Finally the performance of individuals with clinical conditions, including a group with traumatic brain injury, was administered the WTAR. In the standardization sample, brain-injured examinees’ performance on the WTAR was not significantly different than that of a non clinical matched control group. This similarity in scores was contrasted with reduced scores for the brain-injured group on other tasks, such as measures that assessed processing speed.
In the current study, the brain-injured group had a mean WTAR Scaled Score of 105, which was significantly lower ($t = -3.69, p=0.001, d = 1.02$) than that of the matched control group ($M = 115$). When the brain-injured participants’ Observed WTAR Standard Scores (Mean IQ=105) were compared to their WTAR Demographic Predicted Standard Scores (Mean IQ = 105), the difference was non significant. When these variables were compared for the matched control participants, again the difference was nonsignificant ($p=.262$). The control participants’ mean Demographic Predicted IQ score was 113, whereas their mean Observed IQ Score was 115. Both groups evidenced a wide range of age and education and a wide range of IQ estimates (see Table 1) though the control participants’ range on the WTAR IQ estimate was smaller.

<table>
<thead>
<tr>
<th>Group</th>
<th>Age</th>
<th>Age SD</th>
<th>Education</th>
<th>Education SD</th>
<th>WTAR Est IQ</th>
<th>WTAR Est IQ SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABI</td>
<td>26-64</td>
<td>9.0</td>
<td>9-25</td>
<td>3.7</td>
<td>89-125</td>
<td>10.7</td>
</tr>
<tr>
<td>NC</td>
<td>24-65</td>
<td>12.2</td>
<td>10-23</td>
<td>2.7</td>
<td>96-126</td>
<td>9.4</td>
</tr>
</tbody>
</table>

SD = Standard Deviation, WTAR Est IQ = WTAR Estimated IQ Score

The battery also included The Tower of London Drexel version (TOL; Culbertson & Zilmer, 2001). The TOL manual provides test-retest reliability information for 35 patients with Parkinson’s Disease ($M$ age = 70.1, SD = 8.6) over an average interval of 140 days. The Total Move Score was reported to have moderate reliability ($r= .75, p = .000$) as did the time related scores (e.g. Initiation Time $r = .75, p = .000$). External validity for the TOL was reported for children, college and adults samples. In the college sample, TOL variables were unrelated to intelligence (Wechsler Adult Intelligence Scale III Information and Picture Completion subtests). Similarly, in a sample of Parkinson’s patients ($M$ age = 73.5, SD = 6.5) the TOL variables were not related to an IQ estimate.
(National Adult Reading Test –2; Nelson & Willison, 1991) but were significantly correlated with the Mini Mental Status Exam (MMSE; Folstein, Folstein & McHugh, 1975).

The Wisconsin Card Sorting Test (WCST; Heaton, Chelune, Talley, Kay, & Curtiss, 1993) variables have been found to have low test-retest reliability with very large confidence intervals when attempting to determine meaningful change, though reliability estimates may be higher for clinical samples (Strauss, Sherman & Spreen, 2006). In the WCST manual, the authors applied generalizability theory to assess the reliability of the WCST. The WCST was administered to 46 children and adolescents on two occasions approximately 1 month apart. Generalizability coefficients for these scores ranged from .39 for percent perseverative responses, to .72 for percent nonperseverative errors. The perseverative errors coefficient was .52, which would be considered moderate. The SEM reported for these data was based on a SD of 15. These scores ranged from 7.94 to 11.91. Because mean scores were not reported, it is not possible to calculate the improvement on the WCST variables that would be required for a reliable change.

The mean score for normal controls on the Temporal Judgment (Time Estimation) test was reported to be 2.15 (SD = .91; Wilson, et. al, 1998). Gillepsie, Evans, Gardner and Bowen (2002) reported similar results with a median score of 2 though they reported there are too few items on the Temporal Judgment task to examine internal consistency. Validity of the Temporal Judgment Test as a measure of neuropsychological constructs has been questioned, as it did not load on to any of the derived factors of the BADS in an Australian sample (Bennet, Ong & Posnford, 2005). It was included in this study because
the specific ability to measure the passage of time was deemed more relevant than generalized cognitive estimation skills.

The Ruff Figural Fluency (RFF) was also included (Ruff, Light & Evans, 1987). The test-retest reliability of the RFF total number of unique designs has been reported to range from .71-.88 over intervals of three weeks to one year. For perseverative errors and ratio scores, however, the reliability coefficients are low (Strauss et al., 2006). The RFF is modestly related to phonemic and semantic fluency (r = .28 - .67; Strauss et. al, 2006) suggesting that the RFF measures similar but unique executive components. The RFF is modestly related to recognition reading as measured by the NART (r= .45) and there is evidence that slowed motor speed reduces RFF Total Unique Design scores.

The Trailmaking Test was included as a measure of visual scanning speed (Trailmaking Part A) and mental flexibility (Trailmaking Part B). Overall, test-retest reliability in healthy controls is higher for Trailmaking Part B than for Trailmaking Part A (Strauss et al, 2006). In clinical samples, test-retest reliability is quite variable across different populations, and cannot be assumed to be high. Part A and Part B correlate moderately well with each other (r = .31 - .6) though it is noteworthy that Trailmaking B is a more difficult visuoperceptual task with greater motor speed demands (Woodruff, Mendoza, Dickson, Blanchard & Christenberry, 1995). In this study, Trailmaking Part A was not expected to correlate with the 12ET, to provide evidence of divergent validity.

It was important to have a measure of working memory in the test battery, as executive functions such as planning, organization and problem solving require the ability to hold different options and potential decisions in mind while one evaluates them. The measure of working memory in this study was Digit Span from the Wechsler Adult
Intelligence Scale III (WAIS III; Wechsler, 1997). The Digit Span scaled score is a combination of scores on Digits forward, a measure of simple span, and Digits backward, a measure of complex span (WAIS III Technical Manual, 1997). While Digits backward is arguably a better measure of working memory, it was not possible to use the raw score from this part of the subtest in the analysis. This was due to the wide range of participant ages in the two samples; the scaled scores associated with individual raw scores were not equivalent for different age groups, and older participants would have been unduly penalized. For this reason, the composite Digit Span scaled score was analysed. The Digit Span subtest has high test-retest reliability for the different age groups (r = .83 -.89) as well as high split half reliability (r=.90; WAIS III Technical Manual, 1997).

The SET, and by extrapolation the 12ET, potentially has the greatest memory load of all the subtests of the BADS (Wilson, Evans, Alderman, Burgess, & Emslie, 1997). In order to control for poor performance on executive measures due to impaired verbal memory (e.g. forgetting the instructions or rules), the California Verbal Learning Test II Standard Form (CVLT II; a measure of verbal memory and self monitoring) was included in the neuropsychological test battery. The authors of the CVLT II reported the internal consistency was high for the five immediate recall trials (Delis, Kramer, Kaplan & Ober, 2000). The authors reported test-retest reliability for a sample of 78 adults over a 3 week interval was high for some variables (r = .8-.89; including Trials 1-5 Correct, Short Delay Free Recall, Long Delay Free Recall) adequate for others (r = .7-.79; including Recognition Hits, False Positives) and low for others (r ≤.59; including Total Repetitions). The relationship between the CVLT II and other neuropsychological measures was not reported. In addition to providing measures of immediate, delayed and
recognition memory, the CVLT II also provides a gross measure of effort through a forced choice paradigm. All but one participant obtained a perfect score on the forced choice part of the CVLT II. The brain-injured participant who did not obtain a perfect score, made one error. This participant also scored poorly on the memory test, and the forced choice error was on a word the participant never learned over the course of the five trials. These data support the view that the participants put forth good effort on the neuropsychological measures completed in this study, though, of course, it is impossible to be certain each participant put forth full effort on every individual test.

Administration instructions of the 12ET were modelled on those of the BADS SET (see Appendix A). The participants were instructed that they were to attempt as many of the twelve tasks as they could within the fifteen minutes, and that it was not expected that they complete any of the twelve tasks during that time. The 12ET requires a relatively large space for ease of administration. To accommodate this task, the neuropsychological assessments were conducted using a table that was at least 158 cm x 60 cm. The examiner sat across from the participants. The examiner had a stopwatch, and the participants were told they could ask how much time was remaining whenever they wanted. The participants were provided with a sheet listing the instructions and rules of the task as well as a sheet of blank paper and pencil. The instruction sheet remained in view throughout the test. The examiner read the rules of the task while the participants followed along on their own copy. These rules were discussed before commencing the task. The examiner assessed the participant's knowledge of the rules by asking the participant to recite the rules from memory. If the participant was unable to recite the
rules from memory, the rules were read again with further explanations as necessary, until the participant was able to report the rules in their own words.

As with many tasks of executive function, the 12 Elements Test requires a type of behaviour for successful completion that may not be an adaptive behaviour in daily life (i.e. doing a small amount of 12 tasks, rather than completing each one sequentially). However, as the rules and purpose of the task are made explicit, this task assesses a client’s ability to monitor and organise his / her own behaviour, despite its somewhat artificial nature. In addition, the 12ET is a novel task, which is a crucial criterion for a measure of executive function (Denkla, 1994; Lezak, 1995; Burgess, 1997). The total number of tasks attempted, number of rule breaks, number of completed tasks, number of time remaining requests, number of task switches, and time spent on each task were recorded. The maximum score on number of tasks attempted is twelve. With twelve tasks to attempt in 15 minutes, the optimal amount of time allocated per task was 75 seconds, Deviation from the optimal amount of time was calculated for each of the tasks, and then summed. Obviously, if a participant spent a long time on one or two tasks, he / she would spend little time on the remaining tasks. To avoid cancellation, both positive and negative deviations were scaled positively. All time spent on a task (in one or more discrete periods) contributed to this measure. Qualitative information, particularly examinee comments regarding the task itself, was recorded to gather further information on the usefulness of the 12ET.

The 12ET has been evaluated in a preliminary study. It was administered to 45 female and 12 male undergraduate students (age range 17–26 years) who had no history of brain injury or neurological condition. These participant’s scores varied widely (Tasks
Attempted range: 3 to 12). The mean Number of Tasks Attempted was moderate ($M = 9.00, SD = 2.74$) with a similar number of subtask switches ($M = 9.02, SD = 3.47$, range 2 to 15). This sample rarely broke the rule of the task ($M = 0.26, SD = 0.67$) and requested the amount of time remaining infrequently ($M = 1.51, SD = 1.24$). Ninety-five percent of the participants did not spend longer than 399 seconds (6 minutes and 39 seconds) on any of the tasks, and therefore it was empirically determined that spending longer than 7 minutes on any one task was unusual. If a participant spent longer than 7 consecutive minutes on one task, then it was considered a prospective memory failure.

The self report executive function assessment measure used in this study was the Dysexecutive Questionnaire (DEX; Burgess et al., 1996b) that is part of the BADS (Wilson et al., 1996). The DEX is a twenty-item questionnaire constructed to sample the range of problems commonly associated with the dysexecutive syndrome. The questions sampled four broad areas of likely changes as described by Stuss and Benson (1984). These areas are emotional or personality changes, motivational changes, behavioural changes and cognitive changes. Each item is scored on a five-point Likert scale, ranging from “never” to “very often”, with a higher score indicating higher frequency of dysexecutive behaviour in everyday life. The DEX comes in two versions, one to be completed by the brain-injured participant and the other by their chosen informant. The items of the informant report version directly parallel those of the self-report version; for example, self-report: *I act without thinking, doing the first thing that comes to mind*, or informant report: *[Name] acts without thinking, doing the first thing that comes to mind.* Scoring of the DEX is done by converting all the responses to number (i.e. never = 0 and very often = 5) and summing them. Scores on the DEX were correlated with scores on
the original SET using the normative sample for the BADS (Wilson, et al., 1996). According to Cohen’s criteria (1988), the DEX Informant Ratings had a moderate relationship with the SET (r= -0.40, p < 0.001) and the correlation with the total profile score of the BADS (made up of six subtests) was large (r=0.62, p < 0.001). This is evidence of moderate convergent validity between the construct measured by the BADS battery, and the DEX. The DEX Self Report Ratings, however, had low or negligible correlations with the subtests of the BADS (r= -0.26 to 0.02, p = ns; Wilson et al., 1996). Moreover, DEX Self Report scores were significantly different from the DEX Informant Report scores (t=2.85, p=0.006). This difference between the client self report and both the informant report and objective test results is attributed to a lack of insight on the clients' part. Consistent with this hypothesis, the clients' scores on the DEX Self Report were five points lower (indicating fewer problems) on average than the DEX Informant Report scores. The reliability of the DEX (internal consistency, split half reliability, or test-retest reliability) was not reported in the manual.

To evaluate whether personality factors predicted performance on the 12ET, the NEO Five Factor Inventory (NEO FFI; Costa & McCrae, 1992) was included. The NEO-FFI is a short form of the 240 item NEO Personality Inventory. In particular, the Conscientiousness domain was of interest. The convergent validity of the NEO-PI and the Myers-Briggs Type Indicator Form G (MBTI; Briggs & Myers, 1987) has been assessed. Furnham (1996) reported that the NEO-PI and the MBTI show clear overlap with the Conscientiousness domain correlating most strongly with the Judging and Perceiving dimensions on the MBTI (r= 0.50 and r=0.41 respectively). However, Furnham (1996) reported that the predictive and construct validity of the NEO-PI are
stronger than those of the MBTI. MacDonald, Anderson, Tsagrakis and Holland (1994) and Costa and McCrae (1989) also reported that the MTBI Judging and Perceiving had the highest correlations with the NEO-PI Conscientiousness dimension. McCrae and Costa (1989) explained this finding as both Conscientiousness and Judging being measures of orderliness and self discipline. Another study (Hahn & Comrey, 1994) reported a factor analysis for the NEO-PI and the Comrey Personality Scales (Comrey, 1994). The Conscientiousness domain of the NEO-PI had the highest loading (0.75) on the factor named Orderliness vs Lack of Compulsion. This domain also loaded on the factor named Attitude vs Lack of Energy (liking physical activity, hard work and striving to excel) to a lesser extent (0.38).

The NEO-FFI is a 60-item paper and pencil self-report measure based on the same five factor model of personality as the NEO-PI, but it only takes 10-15 minutes to complete. It requires a grade six reading level. Like the NEO-PI, each item of the NEO FFI is answered on a five-point likert scale, the answers ranging from “strongly disagree” to “strongly agree”. As a shorter version, the NEO FFI is useful “when time available for testing is limited, and global information on personality is considered sufficient” (Costa & McCrae, 1992, pp. 1). The NEO FFI consists of five 12-item scales that measure each of the following domains: Neuroticism, Openness, Extraversion, Agreeableness and Conscientiousness. Neuroticism refers to a propensity towards negative affectivity and maladaptive coping (Elliott & Umlauf, 1995). Extraversion, in contrast, is a propensity towards positive affectivity and interpersonal interaction (Elliott & Umlauf, 1995). Openness describes intellectual curiosity and an appreciation of experience, whereas Agreeableness involves thoughts and feelings of trust, sympathy, and altruism (Elliott &
Umlauf, 1995). As above, Conscientiousness refers to a preference for structure and self-discipline with respect to goal directed behaviour (Elliott & Umlauf, 1995). The five principal components were derived through validimax factors of the NEO-PI. The correlations between the domain scores on the NEO-FFI and the domain scores of the NEO-PI were 0.92, 0.90, 0.91, 0.77, and 0.97 for the Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness domains respectively. On average, the short form scales account for 85% of the variance of the longer NEO-PI domain scales. The three-month test-retest reliability coefficients of the NEO-FFI sample for the Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness domains were reported to be 0.79, 0.79, 0.80, 0.75, and 0.83 respectively for a sample of 208 college students. The internal consistency (coefficient alpha) of the NEO-FFI domain scales for Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness were 0.86, 0.77, 0.73, 0.68, and 0.81 respectively.
RESULTS

All analyses were completed using SPSS version 11.0.3 for the Mac. For hypothesis driven analyses, family wise alpha was set at 0.05 and a Bonferroni Correction employed to reduce the possibility of Type 1 error. Part 1 had one hypothesis (that the brain-injured participants would perform more poorly on the 12ET variables than the control participants) and Part 2 had eight specific hypotheses regarding significant correlations between 12ET variables and other neuropsychological measures. Therefore, alpha for these nine hypotheses was set to .006. Significant results are highlighted in **bold**. Results for hypotheses driven analyses that are significant at the .05 level, but do not meet the more stringent Bonferroni Correction alpha level have been highlighted in *italics*. For exploratory analyses, such as further concurrent validity, alpha was set to .05. Effect sizes (Cohen’s d) are presented, and Cohen’s criteria (1988, 1992) where 0.2 is indicative of a small effect, 0.5 a medium effect and 0.8 a large effect size have been applied.

Data Cleaning

Neuropsychological test data was scored using published normative data from healthy control samples. Both examiners scored neuropsychological test data, and the principle investigator rescored approximately half of the data to ensure scoring accuracy. Data were checked for entry errors, and if detected, errors were corrected by referring to the original test record forms. An examination of the histograms of each group indicated two outliers in the brain-injured group data across all neuropsychological test measures, including the IQ estimate. Item analysis determined that the outlier data consistently
belonged to two individuals whose SDS scores exceeded the cutoff for past substance abuse. These two individuals’ data were excluded from the analysis to ensure all members of the brain-injured group were sampled from the same population of brain-injured individuals. The data from the two matched control participants for these brain-injured participants were also dropped from the analysis.

Once the outliers were removed, the two samples were tested for normality using the Kolmogorov-Smirnov test and for homogeneity of variance with Levene’s test. Many of the distributions for the neuropsychological test data were negatively skewed for both the brain-injured and matched control groups. Area transformations with Rankit were attempted on 10 variables and for some, but not all, variables the transformations resulted in non significant tests for normality and non significant tests for homogeneity of variance. However, the data were entered in meaningful units (T scores on neuropsychological tests) and transformation of the data would have resulted in interpretation of a different construct, which was undesirable. Therefore the data were analysed in their untransformed state. Where comparing means, t tests are relatively robust to violations of normality. In addition, t tests can be calculated to take into account heterogeneity of variance. Therefore t tests with non-equal variances assumed were reported where Levene’s test for equality of variances was significant. This procedure was deemed to be preferable to using non parametric tests such as the Mann Whitney test because of the relative increase in power provided by the t test. In contrast, Spearman correlations were chosen because, not only were the data non-normally distributed, but the sample size was relatively small, making Pearson r a less viable correlation coefficient.
Part 1 – Evaluating the Validity of the 12 Elements Test.

Six variables can be obtained from the 12ET. The four continuous variables obtained were: Number of Tasks Attempted, Number of Time Requests, Number of Tasks Completed and deviation (in seconds) from the optimal amount of time per task (i.e. 75 seconds). The Number of Tasks Attempted is based on how many tasks the participant begins. The participant only has to spend a moment on the task (completing one math problem for example) for that task to be considered as attempted. The range for this variable was 1-12. While it was possible for participants to return to tasks they had attempted earlier in the test (and thus obtain more than 12 task switches) this behaviour was rare. Typically, once participants had attempted all 12 tasks, they either indicated they were finished (even if there was time remaining) or they worked on the twelfth task for the remainder of the test time. If they indicated they were finished, they were reminded that the purpose of the task was to complete as much of the 12 tasks as possible within 15 minutes. As with the SET, performance on the tasks themselves was not scored, as the relevant variable was whether or not a task was attempted or not. Number of Time Requests were gathered by noting whenever the participant asked for the time remaining and summing the queries at the end of the test. Number of Tasks Completed was obtained through counting how many subtasks the participants completed. The range on this variable was 0-10, as it was considered impossible to complete the two Dictation subtasks. The Restaurant Menus subtasks were considered complete if the participant wrote down all possible combinations of appetizers, entrees and desserts. Time deviation was calculated once the test had been completed by subtracting the obtained amount of time spent on each subtask from 75. If participants returned to a task a second or third
time, then all the time spent on that task was summed and subtracted from 75. The absolute values of the 12 scores for each participant were then summed to obtain Time Deviation for the entire test. The absolute value of the score was used to avoid positive deviations cancelling out negative deviations.

The 12ET also provided two dichotomous variables: Rule Breaks and whether participants spent >7 minutes on one task. These variables were originally counted and the frequencies were summed. However, in this study, they were low frequency events that were not well described by traditional frequencies. When one considers the Number of Rule Breaks variable, for example, one must remember that no feedback was given while participants were completing the 12ET. As a result, if a participant broke the rule once (by completing two parts of the same task in sequential order) then they may well break the rule again because they had not been reminded of the rule. However, conceptually they were breaking the same rule. Therefore the actual Number of Rule Breaks was not as relevant as whether or not participants broke the rule over the course of the test. The same logic was applied to spending longer than 7 minutes on one task.

The complete data set of neuropsychological test data collected at Time 1 was comprised of 33 variables. In the interest of avoiding multicollinearity and reducing the number of exploratory comparisons, those variables with a significant Spearman correlation (p< .05) with at least one other variable from the same test were not all included in the analysis. Table 2 shows all neuropsychological test variables collected and/or calculated, and the list of the 15 variables that were included in the analyses. Neuropsychological test data was scored using published normative data from healthy control samples.
<table>
<thead>
<tr>
<th>Test Variable Collected</th>
<th>Test Variables Analysed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intelligence</strong></td>
<td></td>
</tr>
<tr>
<td>WTAR Standard Score</td>
<td>WTAR Standard Score</td>
</tr>
<tr>
<td><strong>Working Memory</strong></td>
<td></td>
</tr>
<tr>
<td>Digit Span Scaled Score</td>
<td>Digit Span Scaled Score</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td></td>
</tr>
<tr>
<td>CVLT II Total Immediate Recall</td>
<td>CVLT II Total Immediate Recall</td>
</tr>
<tr>
<td>CVLT II Delayed Recall</td>
<td>CVLT II Repetitions</td>
</tr>
<tr>
<td>CVLT II Repetitions</td>
<td>CVLT II Recognition</td>
</tr>
<tr>
<td><strong>Temp. Judgment</strong></td>
<td></td>
</tr>
<tr>
<td>Time Estimation Time Raw Score</td>
<td>Time Estimation Time Raw Score</td>
</tr>
<tr>
<td><strong>Cognitive Flexibility</strong></td>
<td></td>
</tr>
<tr>
<td>RFF Unique Designs</td>
<td>RFF Unique designs</td>
</tr>
<tr>
<td>RFF Error Ratio</td>
<td>RFF Error Ratio</td>
</tr>
<tr>
<td>RFF Strategy</td>
<td></td>
</tr>
<tr>
<td>WCST Categories Completed</td>
<td>WCST Categories Completed</td>
</tr>
<tr>
<td>WCST Perseverative Errors</td>
<td>WCST Perseverative Errors</td>
</tr>
<tr>
<td>WCST Failure to Maintain Set</td>
<td>WCST Failure to Maintain Set</td>
</tr>
<tr>
<td>Trailmaking B</td>
<td>Trailmaking B</td>
</tr>
<tr>
<td>Trailmaking A – Trailmaking B</td>
<td>Trailmaking A – Trailmaking B</td>
</tr>
<tr>
<td>Trailmaking B Errors</td>
<td></td>
</tr>
<tr>
<td><strong>Inhibition</strong></td>
<td></td>
</tr>
<tr>
<td>Stroop Word</td>
<td></td>
</tr>
<tr>
<td>Stroop Colour</td>
<td></td>
</tr>
<tr>
<td>Stroop Colour Word</td>
<td></td>
</tr>
<tr>
<td>Stroop Interference</td>
<td>Stroop Interference</td>
</tr>
<tr>
<td><strong>Planning</strong></td>
<td></td>
</tr>
<tr>
<td>TOL Total Correct</td>
<td>TOL Total Moves</td>
</tr>
<tr>
<td>TOL Total Moves</td>
<td></td>
</tr>
<tr>
<td>TOL Initiation Time</td>
<td></td>
</tr>
<tr>
<td>TOL Problem Solving Time</td>
<td></td>
</tr>
<tr>
<td>TOL Time Violation</td>
<td>TOL Rule Violations</td>
</tr>
<tr>
<td>TOL Rule Violation</td>
<td></td>
</tr>
<tr>
<td><strong>Visual Scanning</strong></td>
<td></td>
</tr>
<tr>
<td>Trailmaking A</td>
<td>Trailmaking A</td>
</tr>
<tr>
<td><strong>Criterion</strong></td>
<td></td>
</tr>
<tr>
<td>12ET # Tasks Attempted</td>
<td>12ET # Tasks Attempted</td>
</tr>
<tr>
<td>12ET Rule Break?</td>
<td>12ET Rule Break?</td>
</tr>
<tr>
<td>12ET # Time Requests</td>
<td></td>
</tr>
<tr>
<td>12ET Time deviation</td>
<td></td>
</tr>
<tr>
<td>12ET &gt;7 minutes?</td>
<td></td>
</tr>
<tr>
<td>12ET# Tasks Completed</td>
<td></td>
</tr>
</tbody>
</table>

The first analysis was to compare the performance of the two participant groups on the 12ET variables at Time 1 to evaluate content validity. Specifically, it was hypothesized that the control group would perform significantly better on the 12ET test variables, which would provide evidence for the validity and clinical utility of the 12ET as a neuropsychological assessment tool that is sensitive to the effects of a brain injury. As above, to reduce multicollinearity, only two of the six 12ET variables were included: Number of Tasks Attempted, and Number of Rule Breaks. The Number of Tasks Attempted was significantly correlated with the Number of Time Requests, Number of Tasks Completed, whether or not participants spent more than seven minutes on one task, and their total deviation (in seconds) from the optimal amount of time spent on each task (see Table 3). Therefore these variables were not included in the Part 1 analyses. All validity significance tests were one tailed, as it was hypothesised a priori that the matched control group would have superior scores on the neuropsychological tests than the brain-injured group. In addition, these planned comparisons were subject to the Bonferroni correction. For the two analysed variables, Levene’s test for Homogenity of Variance was non significant, therefore equal variances were assumed. On the 12ET, the matched control group attempted more tasks ($M = 8.12$, $SD = 2.9$, Range = 3-12) than the brain-injured group ($M = 6.73$, $SD = 3.1$, Range = 1-12); see Figure 2. This difference was non significant, but the effect size was almost large enough to be considered moderate ($t = -1.661$, $p = .052$, $d = .45$). The matched control group broke the rule of the 12ET infrequently (No = 20, Yes = 6) as did the brain-injured group (No = 19, Yes = 7), and the groups did not differ significantly on this variable ($\chi^2 = .103$, $p = .749$).
Table 3. 12ET Intercorrelation table

<table>
<thead>
<tr>
<th></th>
<th>Tasks Attempted</th>
<th>Rule Break?</th>
<th>Time Requests</th>
<th>Time Deviation</th>
<th>&gt;7 mins on one task?</th>
<th>Tasks Completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tasks Attempted</td>
<td>.</td>
<td>R=.231</td>
<td>r=.299</td>
<td>r=.924</td>
<td>r=.442</td>
<td>r=.344</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P=.100</td>
<td>p=.031</td>
<td>p=.000</td>
<td>p=.001</td>
<td>p=.012</td>
</tr>
<tr>
<td>Rule Break?</td>
<td>.</td>
<td>.</td>
<td>r=.046</td>
<td>r=.197</td>
<td>r=.07</td>
<td>r=.186</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p=.745</td>
<td>p=.162</td>
<td>p=.624</td>
<td>p=.186</td>
</tr>
<tr>
<td>Time Requests</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>r=-.236</td>
<td>r=-.332</td>
<td>r=.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p=.092</td>
<td>p=.016</td>
<td>p=.62</td>
</tr>
<tr>
<td>Time Deviation</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>r=.441</td>
<td>r=.471</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p=.001</td>
<td>p=.000</td>
</tr>
<tr>
<td>&gt; 7 mins on one task?</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>r=.031</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p=.929</td>
</tr>
</tbody>
</table>

p values are two tailed

Figure 2. Bar Graph of 12ET Number of Tasks Attempted for at Time 1
Group Differences on Published Executive Function Tests.

To explore whether the non significant differences between the two participant groups was specific to the 12ET, or generalized to the full battery, performance on the other neuropsychological measures was compared through t tests. The control participants scored significantly better on four conventional executive function test variables than the brain-injured participants (see Table 4 for significant differences). The control participants did not perform significantly better than the brain-injured participants on the WCST, Time Estimation, the Stroop Test, or the Tower of London Total Move Score.

It is possible that the brain-injured group’s diversity masked any differences between the brain-injured group and the matched control group. Seventeen brain-injured participants had a traumatic brain injury (TBI), and nine had an acquired brain injury.
Table 4. Independent t tests between the brain-injured and control participant scores on the published neuropsychological tests (N=52)

<table>
<thead>
<tr>
<th>Test</th>
<th>t</th>
<th>p (1 tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trailmaking B</td>
<td>2.54</td>
<td>.007</td>
</tr>
<tr>
<td>Trailmaking A - Trailmaking B</td>
<td>1.75</td>
<td>.043</td>
</tr>
<tr>
<td>RFF Unique Designs</td>
<td>2.60</td>
<td>.006</td>
</tr>
<tr>
<td>TOL Rule Violations</td>
<td>2.28</td>
<td>.014</td>
</tr>
</tbody>
</table>

(ABI). To explore this possibility, the TBI group was compared to the ABI group on the neuropsychological test variables. There were no significant differences in performance on any of the neuropsychological tests, except the 12ET. The TBI participants were more likely to break the rule of the 12ET (t = -3.35, p = .004, 2 tailed) and completed more 12ET tasks (t = 2.30, p = .031, 2 tailed). Despite completing more tasks than the ABI participants, the TBI participants also showed a trend towards attempting more 12ET tasks (t = -1.91, p = .068, 2 tailed) and showed a trend towards spending closer to the optimal amount of time on each task (t = 1.80, p = .085, 2 tailed). The TBI and ABI participants did not differ on the number of times they requested the time remaining, or whether or not they spent more than seven minutes on one task.

Another potential reason that the two groups did not differ on the 12ET variables is because the ABI group did not evidence executive impairments. To assess whether or not the ABI participants, as a group, demonstrated executive dysfunction, their scores were examined as compared to the normative data (rather than as compared to the matched control group in this sample). Despite their injuries (in some cases very severe injuries), the ABI participants in this sample did not tend to score poorly on the established executive tests. Their lowest score was on Trailmaking B (with their mean score falling in the low average range). In addition, they also performed somewhat slowly on the Stroop Word test, with scores falling in the low average range.
Part Two – Convergent and Divergent Validity

As with Hypothesis One, the large data set of variables collected at week one was reduced to non redundant variables. Despite their significant correlations, however, all of the 12ET variables were included in this part of the analysis as they were the criterion variables of interest. In addition, if a variable from a conventional executive function test was included in an a priori hypothesis or has been reported to be a sensitive measure of impairment in past research, then it was included in this analysis, even if it correlated significantly with another variable of the same test (e.g. Trailmaking A and Trailmaking B scores were significantly correlated, but both had associated specific hypotheses and were therefore included). DEX self and informant questionnaire measures were also included in this analysis (see Table 5).

The 12ET and Intelligence

The WTAR was expected to correlate with all the neuropsychological tests, and it was significantly correlated with three variables of the 12ET (See Table 5), as well as the WCST Categories Completed, Trailmaking B, and Digit Span (see Table 6). The hypothesis that the correlations between the WTAR and the executive function tests would be smaller than the correlations between the executive function tests themselves was not supported, and magnitude of the correlations between the WTAR and the executive function tests were large in some cases. With respect to 12ET variables, those individuals with a higher WTAR Scaled Scores tended to complete more tasks (see
Table 5. Spearman Correlation Matrix of Dependent variables with Criterion Variables for all Participants.

<table>
<thead>
<tr>
<th>Dependent Variables</th>
<th>12ET Tasks Attempted</th>
<th>12ET Tasks Completed</th>
<th>12ET Rule Break*</th>
<th>12ET &gt; 7 Minutes*</th>
<th>Time Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTAR</td>
<td>r=.155</td>
<td>r=.336</td>
<td>r=.265</td>
<td>r=.018</td>
<td>r=-.291</td>
</tr>
<tr>
<td></td>
<td>p=.137</td>
<td>p=.007</td>
<td>p=.029</td>
<td>p=.45</td>
<td>p=.018</td>
</tr>
<tr>
<td>Digit Span</td>
<td>r=.114</td>
<td>r=.414</td>
<td>r=.142</td>
<td>r=-.085</td>
<td>r=-.174</td>
</tr>
<tr>
<td>CVLT II Tot.</td>
<td>r=.094</td>
<td>r=.381</td>
<td>r=.163</td>
<td>r=.001</td>
<td>r=.269</td>
</tr>
<tr>
<td>Imm. Recall</td>
<td>r=.254</td>
<td>p=.003</td>
<td>p=.124</td>
<td>p=.496</td>
<td>p=.027</td>
</tr>
<tr>
<td>CVLT II</td>
<td>r=.090</td>
<td>r=.116</td>
<td>r=.087</td>
<td>r=.213</td>
<td>r=.064</td>
</tr>
<tr>
<td>RFF Total</td>
<td>r=.173</td>
<td>r=.164</td>
<td>r=.280</td>
<td>r=.093</td>
<td>r=.167</td>
</tr>
<tr>
<td>Unique Designs</td>
<td>r=.111</td>
<td>r=.122</td>
<td>p=.022</td>
<td>p=.255</td>
<td>p=.137</td>
</tr>
<tr>
<td>RUFF FF Errors</td>
<td>r=-.026</td>
<td>r=.008</td>
<td>r=.064</td>
<td>r=.016</td>
<td>r=.024</td>
</tr>
<tr>
<td>Time Estimation</td>
<td>r=.114</td>
<td>r=.218</td>
<td>r=.094</td>
<td>r=.284</td>
<td>r=.107</td>
</tr>
<tr>
<td>Stroop</td>
<td>r=-.038</td>
<td>r=-.040</td>
<td>r=-.176</td>
<td>r=-.080</td>
<td>r=.177</td>
</tr>
<tr>
<td>Interference</td>
<td>r=.396</td>
<td>r=.390</td>
<td>r=.105</td>
<td>p=.267</td>
<td>p=.105</td>
</tr>
<tr>
<td>Trailmaking A</td>
<td>r=.360</td>
<td>r=-.149</td>
<td>r=-.176</td>
<td>r=.098</td>
<td>r=.301</td>
</tr>
<tr>
<td></td>
<td>p=.004</td>
<td>r=.145</td>
<td>r=.212</td>
<td>p=.245</td>
<td>p=.015</td>
</tr>
<tr>
<td>Trailmaking B</td>
<td>r=.126</td>
<td>r=.074</td>
<td>r=.277</td>
<td>r=-.092</td>
<td>r=-.283</td>
</tr>
<tr>
<td></td>
<td>p=.186</td>
<td>p=.302</td>
<td>p=.023</td>
<td>p=.286</td>
<td>p=.021</td>
</tr>
<tr>
<td>Trails A – Trails B</td>
<td>r=.179</td>
<td>r=.081</td>
<td>r=.195</td>
<td>r=.067</td>
<td>r=.249</td>
</tr>
<tr>
<td>TOL Total Moves</td>
<td>r=.206</td>
<td>r=.356</td>
<td>r=.154</td>
<td>r=.233</td>
<td>r=.031</td>
</tr>
<tr>
<td>Score</td>
<td>p=.072</td>
<td>p=.005</td>
<td>p=.137</td>
<td>p=.049</td>
<td>p=.410</td>
</tr>
<tr>
<td>TOL Rule</td>
<td>r=-.168</td>
<td>r=.138</td>
<td>r=-.126</td>
<td>r=-.317</td>
<td>r=.041</td>
</tr>
<tr>
<td>WCST</td>
<td>r=.046</td>
<td>r=-.084</td>
<td>r=.132</td>
<td>r=.083</td>
<td>r=-.195</td>
</tr>
<tr>
<td>TOL Rule Violations</td>
<td>p=.072</td>
<td>p=.005</td>
<td>p=.137</td>
<td>p=.049</td>
<td>p=.410</td>
</tr>
<tr>
<td>NEO FFI</td>
<td>r=.206</td>
<td>r=.356</td>
<td>r=.154</td>
<td>r=.233</td>
<td>r=.031</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>p=.072</td>
<td>p=.005</td>
<td>p=.137</td>
<td>p=.049</td>
<td>p=.410</td>
</tr>
<tr>
<td>DEX Self (N=26)</td>
<td>r=.117</td>
<td>r=.221</td>
<td>r=.192</td>
<td>r=.156</td>
<td>r=.043</td>
</tr>
<tr>
<td>Informant (N=22)</td>
<td>r=.179</td>
<td>p=.488</td>
<td>p=.350</td>
<td>p=.177</td>
<td>p=.009</td>
</tr>
</tbody>
</table>

p values are one tailed, * Dichotomous Variable. N=52 except where noted.
Table 6. Spearman Correlations between the Neuropsychological Tests in all participants (N=52) Time 1

<table>
<thead>
<tr>
<th>Test 1</th>
<th>WTAR</th>
<th>Digit</th>
<th>CVLT</th>
<th>CVLT Imm</th>
<th>CVLT Rep</th>
<th>RFF</th>
<th>Unique</th>
<th>Err</th>
<th>Time Est</th>
<th>Int</th>
<th>A</th>
<th>B</th>
<th>A – B</th>
<th>Move</th>
<th>Rule</th>
<th>Cat</th>
<th>Pers</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTAR</td>
<td>1.00</td>
<td>.580</td>
<td>.379</td>
<td>-.230</td>
<td>.397</td>
<td>.147</td>
<td>.106</td>
<td>-.180</td>
<td>.324</td>
<td>.613</td>
<td>.536</td>
<td>-.013</td>
<td>.031</td>
<td>.284</td>
<td>.217</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>.000</td>
<td>.003</td>
<td>.050</td>
<td>.002</td>
<td>.150</td>
<td>.228</td>
<td>.100</td>
<td>.010</td>
<td>.000</td>
<td>.000</td>
<td>.464</td>
<td>.414</td>
<td>.021</td>
<td>.061</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digit</td>
<td>1.00</td>
<td>.389</td>
<td>-.232</td>
<td>.317</td>
<td>.087</td>
<td>.164</td>
<td>.046</td>
<td>.322</td>
<td>.501</td>
<td>.453</td>
<td>-.061</td>
<td>-.041</td>
<td>.295</td>
<td>.334</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>.002</td>
<td>.049</td>
<td>.011</td>
<td>.269</td>
<td>.122</td>
<td>.374</td>
<td>.010</td>
<td>.000</td>
<td>.000</td>
<td>.335</td>
<td>.386</td>
<td>.017</td>
<td>.008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CVLT</td>
<td>1.00</td>
<td>-.208</td>
<td>.236</td>
<td>.219</td>
<td>.156</td>
<td>.141</td>
<td>.269</td>
<td>.350</td>
<td>.277</td>
<td>.101</td>
<td>.052</td>
<td>.470</td>
<td>.360</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imm</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>.069</td>
<td>.046</td>
<td>.059</td>
<td>.134</td>
<td>.160</td>
<td>.027</td>
<td>.006</td>
<td>.023</td>
<td>.238</td>
<td>.358</td>
<td>.000</td>
<td>.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CVLT</td>
<td>1.00</td>
<td>-.358</td>
<td>-.008</td>
<td>.075</td>
<td>.141</td>
<td>-.135</td>
<td>-.121</td>
<td>.025</td>
<td>-.087</td>
<td>-.071</td>
<td>.027</td>
<td>-.114</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rep</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>.005</td>
<td>.479</td>
<td>.300</td>
<td>.160</td>
<td>.170</td>
<td>.197</td>
<td>.431</td>
<td>.270</td>
<td>.308</td>
<td>.426</td>
<td>.027</td>
<td>-.114</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RFF</td>
<td>1.00</td>
<td>-.119</td>
<td>-.152</td>
<td>-.122</td>
<td>.428</td>
<td>.252</td>
<td>.149</td>
<td>.118</td>
<td>.201</td>
<td>-.042</td>
<td>.220</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>.200</td>
<td>.142</td>
<td>.194</td>
<td>.001</td>
<td>.036</td>
<td>.146</td>
<td>.202</td>
<td>.077</td>
<td>.390</td>
<td>.059</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RFF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Err</td>
<td>1</td>
<td>.027</td>
<td>-.082</td>
<td>.130</td>
<td>.000</td>
<td>-.025</td>
<td>-.210</td>
<td>-.147</td>
<td>-.118</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>.424</td>
<td>.281</td>
<td>.178</td>
<td>.499</td>
<td>.429</td>
<td>.067</td>
<td>.149</td>
<td>.202</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stroop</td>
<td>1.00</td>
<td>.041</td>
<td>.216</td>
<td>.112</td>
<td>.038</td>
<td>-.015</td>
<td>.168</td>
<td>.251</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Int</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>.387</td>
<td>.062</td>
<td>.215</td>
<td>.395</td>
<td>.457</td>
<td>.117</td>
<td>.036</td>
<td>.019</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trails</td>
<td>1.00</td>
<td>.029</td>
<td>-.104</td>
<td>.015</td>
<td>.053</td>
<td>-.007</td>
<td>.192</td>
<td>.092</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>.419</td>
<td>.232</td>
<td>.457</td>
<td>.353</td>
<td>.481</td>
<td>.087</td>
<td>.259</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trails</td>
<td>1.00</td>
<td>.604</td>
<td>.368</td>
<td>.002</td>
<td>.058</td>
<td>.129</td>
<td>.294</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>.000</td>
<td>.004</td>
<td>.493</td>
<td>.342</td>
<td>.181</td>
<td>.017</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trails</td>
<td>1.00</td>
<td>.870</td>
<td>.135</td>
<td>-.008</td>
<td>.287</td>
<td>.268</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A – B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>.000</td>
<td>.170</td>
<td>.477</td>
<td>.020</td>
<td>.027</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOL Move</td>
<td>1.00</td>
<td>.129</td>
<td>.007</td>
<td>.428</td>
<td>.298</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rule</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>.182</td>
<td>.481</td>
<td>.001</td>
<td>.016</td>
<td>.162</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WCST Cats</td>
<td>1.00</td>
<td>.424</td>
<td>.046</td>
<td>.099</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

p values are one tailed

Figure 4). Moreover, those with a higher WTAR score also asked how much time was remaining on the 12ET more frequently than those participants with a lower WTAR score (see Figure 5). Breaking the rule of the 12ET was not related to the WTAR, however, whether or not participants spent more than 7 minutes on one task was significantly related to the WTAR.

As shown in Table 5, nine neuropsychological test variables correlated significantly with variables of the 12ET to provide evidence of convergent validity for the 12ET. Several hypotheses were made regarding the convergent and divergent validity of the 12ET based on the entire sample of brain-injured and matched control participants, and these hypothesised comparisons are subject to the Bonferroni correction. Planned comparisons are reported first, followed by exploratory results.

Convergent Validity

*The 12ET and a Measure of Time Estimation*

The hypothesis that scores on the Time Estimation Test would be significantly negatively related to the whether participants spent >7 minutes on one 12ET task was not supported ($r = -.107, p = .226$) though the relationship was in the expected direction (See Table 5). The hypothesis that Time Estimation scores would be significantly negatively related to 12ET Number of Time requests was also not supported, and this relationship showed a trend in the opposite direction than expected ($r = .218, p = .06$). That is, individuals who did better on the Time Estimation Test tended to ask how much time was remaining on the 12ET more frequently than those who did poorly on the Time Estimation Test.
Contrary to expectations, the 12ET Number of Tasks Attempted was not significantly correlated with the WCST or Trailmaking B. However, two 12ET variables were related to Trailmaking B, with significant relationships for Tasks Completed ($r = .277$, $p = .023$) and whether or not participants spent more than 7 minutes on one task ($r = -.283$, $p = .021$). The difference score between speed (in seconds) on Trailmaking A and Trailmaking B was also related to whether or not participants spent more than 7 minutes on one task ($r = -.249$, $p = .038$). Number of Tasks Completed on the 12ET was significantly related to the RFF Number of Unique Designs ($r = .280$, $p = .022$).

Figure 4. Scatterplot of 12ET # Tasks Completed and WTAR Standard Scores for all Participants at Time 1
The 12ET and a Test of Planning

Two 12ET variables (Number of Rule Breaks and Number of Time Remaining Requests) were significantly negatively correlated with the Tower of London Total Move score. Fewer 12ET Rule Breaks ($r = -0.233, p = 0.049$) and fewer time remaining requests ($r = -0.356, p = 0.005$) were associated with fewer moves on the TOL. TOL Rule Violations were also significantly related to 12ET Rule Breaks ($r = -0.317, p = 0.011$). The a priori hypothesis that the 12ET Number of Tasks Attempted would be significantly correlated with the TOL was not supported.

The 12ET and a Measure of Self Monitoring
The hypothesis that participants who tended to repeat themselves on the CVLT II would attempt fewer tasks on the 12ET was not supported.

*The 12ET and Self and Informant Reports*

As expected, the brain-injured participants’ DEX Self Report and DEX Informant Report did not have a significant relationship. In addition, the DEX Self Report did not correlate significantly with performance on the 12ET or other executive function measures. In keeping with predicted findings, the difference scores between the DEX client version and the informant version for the brain-injured group was significantly positively correlated with performance on several neuropsychological test variables (CVLT II Immediate Recall, Trailmaking A, 12ET >7 minutes on One Task) though the difference score was not significantly correlated with any other 12ET variables.

*The 12ET and Personality*

The hypothesis that the Number of 12 Tasks Completed would be significantly related to NEO-FFI Conscientiousness was not supported.

Divergent Validity

*The 12ET and Visual Scanning Speed*

Visual scanning speed from Trailmaking Part A was expected to have low correlations with the 12ET variables, and these low correlations were expected to provide evidence of divergent validity for the 12ET. In this study, Trailmaking A was significantly positively correlated with 12ET Number of Tasks Attempted ($r = .360, p = .004$). Trailmaking A was negatively correlated with whether or not participants spent
longer than 7 minutes on one task ($r = -.301, p = .015$) and participants’ deviation from the optimal amount of time per task ($r = -.276, p = .024$). That is, participants who were faster on Trailmaking A tended to attempt more 12ET tasks, were less likely to spend >7 minutes on one task, and had smaller deviations from the optimal amount of time per task on the 12ET. Inspection of the data showed a wide range of variability in scores for both groups (see Figure 6) including an outlier in the brain-injured group who attempted all 12 tasks on the 12ET but who scored very low on Trailmaking A.

As above, the RFF Number of Unique Designs was significantly correlated with Number of Tasks Completed on the 12ET. Further analyses revealed the RFF Unique Design was also significantly related to Trailmaking A ($r = .428, p = .001$) which suggests visual scanning speed (and perhaps processing speed in general) may underlie these various relationships. That is, individuals who were fast tended to complete more tasks on the 12ET, as well as create more unique designs on the RFF.

Convergent Validity by Group

Because one might not expect to see relationships between neuropsychological tests in control participants (who might be expected to perform at ceiling levels on many, if not all of them) the correlations between the neuropsychological tests and the 12ET have been presented separately for each group (see Tables 7 and 9). In addition, the intercorrelations between the neuropsychological tests have been provided for the brain injured group (see Table 8). When the data were analysed separately for each group, a different pattern of relationships emerged.
Figure 6. Scatterplot of Trailmaking Test Part A and 12ET Tasks Attempted for both groups at Time 1

The 12ET and Intelligence

Many of the correlations between the WTAR Standard Score and the neuropsychological test variables disappeared when only the brain-injured group was examined. In contrast, the correlations between the WTAR and the 12ET variables remained significant, and in some cases were even larger than in the whole sample (e.g. between the WTAR and whether or not brain-injured participants spent longer than 7 minutes on one task).

The 12ET and Memory

When the brain-injured participants were analysed alone, the significant correlation between the CVLT II Immediate memory score and the 12ET variables disappeared suggesting poor memory in the brain-injured participants was not a
Table 7. Spearman Correlations of Dependent variables with Criterion Variables for Brain-injured Participants at Time 1

<table>
<thead>
<tr>
<th>Dependent Variables</th>
<th>12ET # Tasks Attempted</th>
<th>12ET # Time Requests</th>
<th>12ET # Tasks Completed</th>
<th>12ET Rule Break?</th>
<th>12ET &gt; 7 mins</th>
<th>12ET Time Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTAR</td>
<td>r= .389</td>
<td>r= .487</td>
<td>r= -.098</td>
<td>r= .116</td>
<td>r= -.506</td>
<td>r= -.340</td>
</tr>
<tr>
<td>CVLT II</td>
<td>r= .183</td>
<td>r= .272</td>
<td>r= .202</td>
<td>r= -.093</td>
<td>r= -.306</td>
<td>r= -.130</td>
</tr>
<tr>
<td>CVLT II</td>
<td>r= .116</td>
<td>r= -.146</td>
<td>r= -.086</td>
<td>r= -.162</td>
<td>r= .030</td>
<td>r= -.177</td>
</tr>
<tr>
<td>RUFF FF</td>
<td>r= .019</td>
<td>r= .159</td>
<td>r= .357</td>
<td>r= -.203</td>
<td>r= -.200</td>
<td>r= .180</td>
</tr>
<tr>
<td>RUFF FF</td>
<td>r= -.134</td>
<td>r= -.256</td>
<td>r= .000</td>
<td>r= -.203</td>
<td>r= .114</td>
<td>r= .168</td>
</tr>
<tr>
<td>Time</td>
<td>r= .012</td>
<td>r= .187</td>
<td>r= .111</td>
<td>r= -.242</td>
<td>r= .031</td>
<td>r= -.132</td>
</tr>
<tr>
<td>Stoop</td>
<td>r= .022</td>
<td>r= -.222</td>
<td>r= -.132</td>
<td>r= .035</td>
<td>r= .171</td>
<td>r= .053</td>
</tr>
<tr>
<td>Trailmaking A</td>
<td>r= .306</td>
<td>r= -.194</td>
<td>r= .323</td>
<td>r= -.093</td>
<td>r= -.513</td>
<td>r= -.183</td>
</tr>
<tr>
<td>Trailmaking B</td>
<td>p= .064</td>
<td>p= .171</td>
<td>p= .054</td>
<td>p= .326</td>
<td>p= .004</td>
<td>p= .186</td>
</tr>
<tr>
<td>Trails A – Trails B</td>
<td>r= .483</td>
<td>r= .020</td>
<td>r= .202</td>
<td>r= .041</td>
<td>r= .407</td>
<td>r= .205</td>
</tr>
<tr>
<td>Tower of B</td>
<td>r= .483</td>
<td>r= .020</td>
<td>r= .202</td>
<td>r= .041</td>
<td>r= .407</td>
<td>r= .205</td>
</tr>
<tr>
<td>London Tot.</td>
<td>r= -.146</td>
<td>r= -.434</td>
<td>r= .122</td>
<td>r= -.128</td>
<td>r= .071</td>
<td>r= .219</td>
</tr>
</tbody>
</table>

* Planned comparisons subject to the Bonferroni correction.  p values are one tailed. N=26 except where noted.

**Table 8. Spearman Correlations between Neuropsychological Tests in the Brain-injured Sample (N=26 except for the DEX where N=22) at Time 1**

<table>
<thead>
<tr>
<th>WTAR Digit</th>
<th>CVLT Imm.</th>
<th>DEX</th>
<th>RFF</th>
<th>Time</th>
<th>Stroop</th>
<th>Trails</th>
<th>Trails</th>
<th>Trails</th>
<th>Trails</th>
<th>TOT</th>
<th>TOL</th>
<th>WCST</th>
<th>WCST</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Span</strong></td>
<td><strong>Im.</strong></td>
<td><strong>rep.</strong></td>
<td><strong>Unique</strong></td>
<td><strong>Est.</strong></td>
<td><strong>Int.</strong></td>
<td><strong>A</strong></td>
<td><strong>B</strong></td>
<td><strong>A – B</strong></td>
<td><strong>Move</strong></td>
<td><strong>Rule</strong></td>
<td><strong>Cat.</strong></td>
<td><strong>Pers.</strong></td>
<td><strong>p</strong></td>
</tr>
<tr>
<td>WTAR</td>
<td>1.00</td>
<td>.700</td>
<td>.194</td>
<td>-.023</td>
<td>.096</td>
<td>.233</td>
<td>.241</td>
<td>-.134</td>
<td>.297</td>
<td>.414</td>
<td>.384</td>
<td>-.371</td>
<td>-203</td>
</tr>
<tr>
<td><strong>p</strong></td>
<td>.000</td>
<td>.171</td>
<td>.455</td>
<td>.336</td>
<td>.126</td>
<td>.118</td>
<td>.257</td>
<td>.070</td>
<td>.018</td>
<td>.207</td>
<td>.293</td>
<td>.202</td>
<td>.159</td>
</tr>
<tr>
<td>Digit</td>
<td>1.00</td>
<td>.265</td>
<td>-.095</td>
<td>-.011</td>
<td>.296</td>
<td>.215</td>
<td>-.181</td>
<td>.367</td>
<td>.552</td>
<td>.468</td>
<td>-.102</td>
<td>-.148</td>
<td>.406</td>
</tr>
<tr>
<td><strong>p</strong></td>
<td>.096</td>
<td>.322</td>
<td>.480</td>
<td>.072</td>
<td>.146</td>
<td>.188</td>
<td>.033</td>
<td>.002</td>
<td>.008</td>
<td>.310</td>
<td>.235</td>
<td>.020</td>
<td>.014</td>
</tr>
<tr>
<td>CVLT</td>
<td>1.00</td>
<td>-.063</td>
<td>-.454</td>
<td>.120</td>
<td>.059</td>
<td>-.019</td>
<td>.323</td>
<td>.354</td>
<td>.304</td>
<td>.124</td>
<td>.055</td>
<td>.497</td>
<td>.419</td>
</tr>
<tr>
<td><strong>p</strong></td>
<td>.380</td>
<td>.017</td>
<td>.280</td>
<td>.388</td>
<td>.464</td>
<td>.054</td>
<td>.038</td>
<td>.066</td>
<td>.274</td>
<td>.395</td>
<td>.005</td>
<td>.005</td>
<td></td>
</tr>
<tr>
<td>DEX</td>
<td>1.00</td>
<td>.424</td>
<td>.150</td>
<td>.080</td>
<td>.085</td>
<td>.098</td>
<td>.181</td>
<td>.443</td>
<td>.213</td>
<td>.269</td>
<td>.101</td>
<td>.003</td>
<td></td>
</tr>
<tr>
<td><strong>p</strong></td>
<td>.424</td>
<td>.195</td>
<td>.054</td>
<td>-.111</td>
<td>-.364</td>
<td>-.240</td>
<td>-.117</td>
<td>-.257</td>
<td>-.131</td>
<td>-.110</td>
<td>-.200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RFF</td>
<td>1.00</td>
<td>.192</td>
<td>.407</td>
<td>.311</td>
<td>.048</td>
<td>.124</td>
<td>.302</td>
<td>.124</td>
<td>.281</td>
<td>.323</td>
<td>.187</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>p</strong></td>
<td>.141</td>
<td>-.272</td>
<td>.388</td>
<td>.214</td>
<td>.080</td>
<td>.075</td>
<td>.166</td>
<td>.140</td>
<td>.367</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>p</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time Est</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>p</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stroop Int</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>p</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trails A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>p</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trails B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>p</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trails A – B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>p</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOL Moves</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>p</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOL Rule</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>p</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WCST Cats</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>p</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

contributor to 12ET scores. Rather, control participants with superior memory scores were more likely to make Time Requests than those with average memory scores.

The 12ET and Tests of Cognitive Flexibility

In contrast to the entire sample where Trailmaking B had a non significant relationship with 12ET Number of Tasks Attempted, this correlation was significant (p<.05) in the brain-injured group. However, this was a planned comparison, and this relationship was not significant at the more stringent level required by the Bonferroni correction ($r = .483, p = .044$). In addition, the difference score between Trailmaking A and Trailmaking B (a gross index of switching time; Salthouse 1998) was significantly

Figure 7. Trailmaking A T Score and Whether or not participants spent >7 Minutes on one 12ET task at Time 1

> 7 minutes on one task?

Brain Injured Group
Control Group
correlated with 12ET Number of Tasks Attempted, spending more than 7 minutes on one task, and Time Deviation for the brain-injured group. Finally, in the brain-injured group, the WCST was significantly correlated with Number of Time Requests. The WCST Categories Completed was also significantly correlated (p<.05) with 12ET Number of Tasks Attempted, but because this was a planned comparison, this relationship was not significant at the more stringent level required by the Bonferroni correction ($r = .429, p = .014$).

*The 12ET and a Test of Planning*

In the entire sample the TOL Total Move Score, was significantly related to the 12ET Number of Time Requests and whether or not participants broke the rule of the task. When only the brain-injured participants were included in the correlation, the
Figure 9. Scatterplot of 12ET Tasks Completed and Tower of London Total Move score for Brain-injured participants at Time 1

Figure 10. Scatterplot 12ET Rule Break and TOL Rule Violation for the Control Group at Time 1
correlation between 12ET Rule Breaks and TOL Rule Violations observed in the entire sample was no longer significant. However, a new relationship emerged between the brain-injured participants’ TOL Rule Violations and 12ET Time Deviations \( (r = .338, p = .046) \).

The 12ET and Personality

Conscientiousness on the NEO-FFI was hypothesized to be significantly related to 12ET Number of Tasks Completed for the entire sample, but this expectation was not supported. However, Conscientiousness was significantly related to 12ET Number of Time Requests \( (r = -.402, p = .021) \), with control participants who scored higher on Conscientiousness tending to ask about the time remaining less frequently (see Figure 11).
Table 9. Spearman Correlations of Dependent variables with Criterion Variables for Control Participants at Time 1

<table>
<thead>
<tr>
<th>Dependent Variables</th>
<th>12ET # Tasks Attempted</th>
<th>12ET # Tasks Requests</th>
<th>12ET # Tasks Completed</th>
<th>12ET Rule Breaks?</th>
<th>12ET &gt; 7 mins</th>
<th>12ET Time Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTAR</td>
<td>r = -.260</td>
<td>r = -.007</td>
<td>r = .397</td>
<td>r = -.147</td>
<td>r = -.029</td>
<td>r = .307</td>
</tr>
<tr>
<td>Digit Span</td>
<td>r = -.164</td>
<td>r = .464</td>
<td>r = -.008</td>
<td>r = -.012</td>
<td>r = -.039</td>
<td>r = .126</td>
</tr>
<tr>
<td>CVLT II Total Recall</td>
<td>r = -.231</td>
<td>r = .378</td>
<td>r = -.103</td>
<td>r = .226</td>
<td>r = -.135</td>
<td>r = .198</td>
</tr>
<tr>
<td>RFF Unique</td>
<td>r = .202</td>
<td>r = .042</td>
<td>r = .063</td>
<td>r = -.324</td>
<td>r = -.069</td>
<td>r = -.263</td>
</tr>
<tr>
<td>RFF Errors</td>
<td>r = .243</td>
<td>r = -.020</td>
<td>r = -.094</td>
<td>r = .031</td>
<td>r = -.096</td>
<td>r = .209</td>
</tr>
<tr>
<td>Time Estimation</td>
<td>r = .410</td>
<td>r = -.174</td>
<td>r = .274</td>
<td>r = .178</td>
<td>p = .320</td>
<td>r = .381</td>
</tr>
<tr>
<td></td>
<td>p = .149</td>
<td>r = .204</td>
<td>r = -.059</td>
<td>r = -.324</td>
<td>r = .226</td>
<td>r = -.196</td>
</tr>
<tr>
<td>Stroop</td>
<td>r = -.106</td>
<td>r = .364</td>
<td>r = -.256</td>
<td>r = -.153</td>
<td>r = -.183</td>
<td>r = .019</td>
</tr>
<tr>
<td>Interference</td>
<td>p = .303</td>
<td>r = .034</td>
<td>r = .103</td>
<td>p = .228</td>
<td>p = .185</td>
<td>p = .463</td>
</tr>
<tr>
<td>Trailmaking A</td>
<td>r = .414</td>
<td>r = -.059</td>
<td>r = .241</td>
<td>r = -.018</td>
<td>r = -.019</td>
<td>r = .430</td>
</tr>
<tr>
<td>Trailmaking B</td>
<td>r = -.188</td>
<td>r = .038</td>
<td>r = .159</td>
<td>r = -.226</td>
<td>r = -.058</td>
<td>r = .170</td>
</tr>
<tr>
<td>Trails A – Trails B</td>
<td>r = -.253</td>
<td>r = .156</td>
<td>r = -.164</td>
<td>r = -.171</td>
<td>r = -.077</td>
<td>r = .248</td>
</tr>
<tr>
<td>Tower of London</td>
<td>r = -.315</td>
<td>r = -.348</td>
<td>r = .084</td>
<td>r = -.391</td>
<td>r = -.010</td>
<td>r = .232</td>
</tr>
<tr>
<td>Tower of London</td>
<td>r = -.191</td>
<td>r = .150</td>
<td>r = .203</td>
<td>r = -.345</td>
<td>r = -.023</td>
<td>r = .185</td>
</tr>
<tr>
<td>WCST Cats.</td>
<td>r = -.279</td>
<td>r = .243</td>
<td>r = .187</td>
<td>r = -.179</td>
<td>r = -.113</td>
<td>r = .210</td>
</tr>
<tr>
<td>NEO FFI</td>
<td>r = -.181</td>
<td>r = -.402</td>
<td>r = .037</td>
<td>r = .012</td>
<td>r = .251</td>
<td>r = .212</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>p = .188</td>
<td>p = .021</td>
<td>p = .429</td>
<td>p = .476</td>
<td>p = .108</td>
<td>p = .149</td>
</tr>
</tbody>
</table>

**Bold** - Correlation is significant at the .05 level (p values are 1-tailed). N=26


Divergent Validity by Group

*The 12ET and Visual Scanning Speed*
When the groups were analysed separately, the correlation between Trailmaking A and 12ET Number of Tasks Attempted for the control group remained significant \( r = .414, p = .02 \), but the correlation for the brain-injured group was no longer significant \( r = 0.306, p = .064 \). However, with the outlier from the brain-injured group removed (see Figure 6), the correlation was again significant \( r = 0.45, p = .01 \).

Intercorrelations between Conventional Tests

With respect to the other neuropsychological test correlations (see Tables 6 and 8), the significant correlations between the CVLT II and the RFF, Trailmaking A, and Trailmaking A-Trailmaking B observed in the entire sample were no longer significant in the brain-injured group. Repetitions on the CVLT II, which were significantly related to RFF Unique Designs in the entire sample, were instead significantly related to perseverative errors on the WCST in the brain-injured group. In addition, Ruff Unique designs were also significantly related to WCST Perseverative Errors in the brain-injured sample. The other difference of note was the reduction of the correlation between Time Estimation and the WCST variables; these relationships were not significant when the brain-injured group was analysed alone.

Given the discrepancies between the correlations for the entire sample and the brain-injured group, clearly the control participants accounted for some of the significant relationships between the conventional executive function tests and the criterion variables (See Table 9). The most notable difference between the control and brain-injured participants’ intercorrelations for the neuropsychological measures was the relative lack of correlation between the executive function tests and the WTAR for the control group (see Table 10).
**Intercorrelations on the 12ET by group**

The 12ET test variables were significantly correlated, and therefore only select variables have been chosen for exploratory analysis in this section. Post hoc exploration of the intercorrelations (1 tailed) for each group found some interesting differences. For the brain-injured group, as might be expected, Number of Time Requests was significantly negatively related to both Time Deviation ($r = -.436, p = .013$), and whether or not participants spent > 7 minutes on one task ($r = -.399, p = .022$). More interestingly however, 12ET Number of Tasks Attempted was positively correlated with Number of Time Requests ($r = .437, p = .013$) in the brain-injured group. For the control participants, Number of Time Requests was not significantly correlated with any other 12ET variables.

### Table 10. Selected Spearman Intercorrelations between Neuropsychological Tests for the Control Group at Time 1 (N=26)

<table>
<thead>
<tr>
<th></th>
<th>WCST Pers.</th>
<th>WCST Cats</th>
<th>RFF Unique</th>
<th>Stroop Int</th>
<th>Trailmaking B</th>
<th>Trails A - B</th>
<th>TOL. Move</th>
<th>Time Est.</th>
<th>WTAR SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>WCST Pers.</td>
<td>-1</td>
<td>$r=.609$</td>
<td>$p=.000$</td>
<td>$r=.205$</td>
<td>$r=.040$</td>
<td>$r=.131$</td>
<td>$r=.205$</td>
<td>$r=.310$</td>
<td>$r=.075$</td>
</tr>
<tr>
<td>WCST Cats</td>
<td>.</td>
<td>1-</td>
<td>$r=-.481$</td>
<td>$p=.006$</td>
<td>$r=.319$</td>
<td>$r=.079$</td>
<td>$r=.239$</td>
<td>$r=.088$</td>
<td>$r=.234$</td>
</tr>
<tr>
<td>RFF Unique</td>
<td>.</td>
<td>.</td>
<td>1-</td>
<td>$r=.025$</td>
<td>$r=.041$</td>
<td>$r=.006$</td>
<td>$r=.022$</td>
<td>$r=.026$</td>
<td>$r=.306$</td>
</tr>
<tr>
<td>Stroop Int.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1-</td>
<td>$r=-.170$</td>
<td>$r=.157$</td>
<td>$r=.-.103$</td>
<td>$r=.197$</td>
<td>$r=.-.276$</td>
</tr>
<tr>
<td>Trailmaking B</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1-</td>
<td>$r=.902$</td>
<td>$r=.173$</td>
<td>$r=.-.196$</td>
<td>$r=.699$</td>
</tr>
<tr>
<td>Trials A - B</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1-</td>
<td>$r=.062$</td>
<td>$r=.-.256$</td>
<td>$r=.629$</td>
</tr>
<tr>
<td>TOL Move</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1-</td>
<td>$r=.-.103$</td>
<td>$r=.000$</td>
</tr>
<tr>
<td>Time Est.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1-</td>
<td>$r=.278$</td>
</tr>
</tbody>
</table>

*Bold* – significant at $p<.05$ (1 tailed, no Bonferroni correction).

WTAR SS = Wechsler Test of Adult Reading Standard Score, RFF Unique = Ruff Figural Fluency Total Unique
Designs, Time Est = Temporal Judgment battery from the Behavioural Assessment of the Dysexecutive Syndrome,
Stroop Int = Stroop Test Interference Score (Golden version), Trails A – B = Trailmaking Test Part A - Trails B =
Trailmaking Test Part B, TOL TMove = Tower of London Total Move Score, WCST Cats = Wisconsin Card Sorting
Test Categories Completed, WCST Pers. = Wisconsin Card Sorting Test Perseverative Errors
Part 3. Reliability and Normative Data of the 12ET

To determine if the 12ET showed practice effects, a multiple regression was employed for all 52 participants for each of the continuous 12ET test variables (Number of Tasks Attempted, Number of Tasks Completed, Number of Time Requests, Time Deviation) with scores at Time 2 as the outcome variable, and group membership, and scores at Time 1 as the predictors. The dichotomous variables (Rule Breaks and whether or not participants spent >7 minutes on one task) were analysed with Chi Squares, and are reported below. Group membership was included in the regression to determine if brain-injured participants obtained the same benefit from practice as the control participants. If there was a significant group effect at this point, it meant that the brain-injured group and the control group were demonstrating significantly different practice effects at Time 2.

The general equation for the multiple regression was as follows:

\[ \text{Outcome} = a + b_1(\text{pretest}) + b_2(\text{group}) + b_3(\text{pretest X interaction}) + \text{error} \]

Using the 12ET variable of Number of Tasks Attempted as an example, the specific equation would be as follows:

\[ 12ET \text{ Number of Tasks Attempted at Time 2} = a + b_1(12ET \text{ Number of Tasks Attempted at Time 1}) + b_2(\text{group}) + b_3 (12ET \text{ Number of Tasks Attempted at Time 1 X Group}) + \text{error} \]
Group Differences

In the multiple regression, The Number of Tasks Attempted at Time 1 predicted 50.3% of the Number of Tasks Attempted at Time 2 ($F = 50.53, p = .000, d = 1.97$). When group membership was added, the model only predicted 51% of the variance (R change = .007) so the two groups obtained the same benefit from the pretest practice on this variable. There was no significant interaction effect (R change .00).

With respect to the Number of Time Requests, the score at Time 1 predicted 33.2% of the variance in the Number of Time Requests at Time 2 ($F = 24.7, p = .000, d = 1.38$). Group membership did not significantly change the model (R change = .000) and the interaction between group membership and Number of Time Requests was non significant (R change = 0.29).

The participants’ deviation from the optimal time on each task at Time 1 predicted 58.1% of the Time Deviation at Time 2 ($F = 69.5, p = .000, d = 2.31$). Adding group membership to the model was non significant (R change = .021) as was the interaction (R change = .006).

The Number of Tasks Completed at Time 1 was also a significant predictor (28.3%) of the Number of Tasks Completed at Time 2 ($F = 19.7, p = .000, d = 1.23$). Group membership did not add significantly to the model (R change =.01) nor was the interaction between group membership and Number of Tasks Completed at Time 1 significant (R change = .000).

The two dichotomous variables were low frequency occurrences and were analysed with the Crosstabs procedure by group. The Likelihood Ratio is reported because of the small sample size. For Rule Breaks (see Figure 12), there was a significant
relationship between the Time 1 and Time 2 scores for the control participants (see Table 11; Likelihood Ratio Statistic = 6.46, p = .011) but not for the brain-injured participants (Likelihood Ratio Statistic = 1.17, p = .279). Because some cells had lower frequencies than the minimum of 5 required for this procedure (see Table 11) Fisher’s Exact Test was also checked (see Table 12). Again, this test was significant for the control participants (p = .046) but not for the brain-injured participants (p = .263). The analysis of whether or not participants spent longer than 7 minutes on one task was incomplete because no control participants spent longer than 7 minutes on one task at Time 2 (see Table 13). Therefore, the Likelihood Ratio Test was not reported for that group. For the brain-injured group, the scores at Time 1 and Time 2 were not significantly related (Likelihood Ratio Statistic = 2.39, p = .122). Again, because of too few observed frequencies, Fisher’s Exact Test was checked, but it was also non significant for the brain-injured participants (see Table 14).

Table 11. Crosstabulation by Group for 12ET Rule Breaks

<table>
<thead>
<tr>
<th>Group</th>
<th>Rule Break Time 1?</th>
<th>Rule Break at Time 2?</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>no</td>
<td>Expected Count</td>
<td>18.5</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>Expected Count</td>
<td>5.5</td>
</tr>
<tr>
<td>Total</td>
<td>Expected Count</td>
<td></td>
<td>24.0</td>
</tr>
<tr>
<td>ABI</td>
<td>no</td>
<td>Expected Count</td>
<td>13.9</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>Expected Count</td>
<td>5.1</td>
</tr>
<tr>
<td>Total</td>
<td>Expected Count</td>
<td></td>
<td>19.0</td>
</tr>
</tbody>
</table>
Table 12. Chi-Square Tests for 12ET Rule Breaks

<table>
<thead>
<tr>
<th>Group</th>
<th>Value</th>
<th>df</th>
<th>Asymptote Sig. (2-sided)</th>
<th>Exact Sig. (2-sided)</th>
<th>Exact Sig. (1-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>Pearson Chi-Square</td>
<td>7.222</td>
<td>1</td>
<td>.007</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td>Continuity Correction</td>
<td>3.291</td>
<td>1</td>
<td>.070</td>
<td>0.070</td>
</tr>
<tr>
<td></td>
<td>Likelihood Ratio</td>
<td>6.464</td>
<td>1</td>
<td>.011</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fisher's Exact Test</td>
<td></td>
<td></td>
<td>0.046</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linear-by-Linear Association</td>
<td>6.944</td>
<td>1</td>
<td>.008</td>
<td></td>
</tr>
<tr>
<td>Brain-injured</td>
<td>Pearson Chi-Square</td>
<td>1.236</td>
<td>1</td>
<td>.266</td>
<td>0.263</td>
</tr>
<tr>
<td></td>
<td>Continuity Correction</td>
<td>0.376</td>
<td>1</td>
<td>.540</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Likelihood Ratio</td>
<td>1.172</td>
<td>1</td>
<td>.279</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fisher's Exact Test</td>
<td></td>
<td></td>
<td>0.263</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linear-by-Linear Association</td>
<td>1.189</td>
<td>1</td>
<td>.276</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N of Valid Cases</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bold – p value is significant at the .05 level. ABI = Brain-injured Group

Table 13. Crosstabulation by Group for 12ET for > minutes on one task

<table>
<thead>
<tr>
<th>Group</th>
<th>&gt; 7 minutes on one task at Time 1?</th>
<th>No</th>
<th>Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC</td>
<td>No</td>
<td>24</td>
<td>24</td>
<td>24.0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>26</td>
<td>5.0</td>
<td>26.0</td>
</tr>
<tr>
<td>ABI</td>
<td>No</td>
<td>17.8</td>
<td>4.2</td>
<td>22.0</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>3.2</td>
<td>.8</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>21.0</td>
<td>5.0</td>
<td>26.0</td>
</tr>
</tbody>
</table>

Table 14. Chi Square Tests for Spending >7 Minutes on One Task

<table>
<thead>
<tr>
<th>Group</th>
<th>Value</th>
<th>df</th>
<th>Asymptote Sig. (2-sided)</th>
<th>Exact Sig. (2-sided)</th>
<th>Exact Sig. (1-sided)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain-injured</td>
<td>Pearson Chi-Square</td>
<td>2.881</td>
<td>1</td>
<td>.090</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Continuity Correction</td>
<td>1.016</td>
<td>1</td>
<td>.314</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Likelihood Ratio</td>
<td>2.386</td>
<td>1</td>
<td>.122</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fisher's Exact Test</td>
<td></td>
<td></td>
<td>.155</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Linear-by-Linear Association</td>
<td>2.771</td>
<td>1</td>
<td>.096</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N of Valid Cases</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 12. Bar Graph of 12ET Rule Breaks at Time 1 and Time 2
Practice Effects on the 12ET vs Stability of the DEX

Variability in concurrent validity between the DEX and 12ET was expected at week eight due to practice effects on the 12ET while the DEX scores were expected to remain stable in the absence of a rehabilitation intervention. The significant correlation between the DEX difference score (DEX Informant Report score – DEX Self Report score) and whether or not the brain-injured participants spent more than 7 minutes on one 12ET task at Time 1 disappeared at Time 2. However, the DEX Self Report at Time 2 was significantly correlated with 12ET Rule Breaks at Time 2 ($r = .557$, $p = .002$ 1 tailed), despite the lack of a relationship between these variables at Time 1.

The DEX Self Report at Time 1 ($M = 25.9$, range = 3-59) versus Time 2 ($M = 26.6$, range = 6-64) were significantly correlated ($r = .703$, $p = .000$) and the DEX difference score at Time 1 ($M = 2.09$, range = -23 to 31) and at Time 2 ($M = 3.9$, range = -12 to 31) were also significantly correlated ($r = .742$, $p = .000$). Therefore, the changes in correlations between the DEX variables and the 12ET variables were due to improved scores on the 12ET variables from Time 1 to Time 2 (see normative data tables below) rather than changes on the DEX. In fact, the scores on the 12ET variables did increase from Time 1 to Time 2. As a group, the participants of this study attempted more tasks at Time 2 (Number of Tasks Attempted; $t = -3.185$, $p = 0.001$, $d = 0.88$), completed more 12ET tasks at Time 2 (Number of Tasks Completed; $t = 2.173$, $p = 0.016$, $d = .60$), and tended to spend closer to the optimal amount of time on each task at Time 2 (Time Deviation; $t = 2.603$, $p = 0.005$, $d = .72$).
Are Practice Effects Mediated by Attribute Variables?

To assess whether relevant attribute variables predict improvement differentially in test scores, general intelligence (WTAR estimated IQ), working memory (Digit Span) and memory (CVLT II delayed recall) were entered in a sequential (stepwise) manner into the regression equation after Time 1 scores were forced in. These variables were expected to be correlated, and the correlations are listed in an intercorrelation matrix (See Table 15). As the correlations are moderate to large, no correction has been used as there is actually only one attribute variable, and alpha was set at 0.05 for the regression analyses.

None of the attribute variables significantly contributed to the model’s ability to predict Tasks Attempted at Time 2 over and above the predictive power of scores at Time 1. Similarly, the regression for Number of Time Requests at Time 2 was non significant for the attribute variables. The attribute variables did not add to the predictive power of the model for time deviation from the optimal time per task. Finally the attribute variables did not significantly contribute to the Number of Tasks Completed at Time 2.

Table 15. Spearman’s Correlations of Attribute Variables for all participants

<table>
<thead>
<tr>
<th></th>
<th>Digit Span</th>
<th>CVLT Delayed Recall</th>
<th>WTAR Standard Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Digit Span</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>1.000</td>
<td><strong>.441</strong></td>
<td><strong>.580</strong></td>
</tr>
<tr>
<td>Sig. (1-tailed)</td>
<td>.</td>
<td><strong>.001</strong></td>
<td><strong>.000</strong></td>
</tr>
<tr>
<td><strong>CVLT Delayed Recall</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>1.000</td>
<td>.</td>
<td><strong>.333</strong></td>
</tr>
<tr>
<td>Sig. (1-tailed)</td>
<td>.</td>
<td>.</td>
<td><strong>.008</strong></td>
</tr>
</tbody>
</table>

Bold – p value is significant at .05
Figure 13. Scatterplot of 12ET Rule Break at Time 1 and Time 2 with WTAR SS for the Brain-injured Group

Figure 14. Scatterplot for 12ET Rule Break at Time 1 and Time 2 for the Control Group

As above, whether or not participants made a 12ET Rule Break at Time 1 significantly predicted Rule Breaks at Time 2 for the control participants (Likelihood
**Ratio Statistic = 6.46, p= .011** but not for the brain-injured participants (Likelihood Ratio Statistic = 1.17, p= .279). Inspection of the scatterplots (see Figures 13 and 14) showed there was little difference between the two groups with respect to the relationship between WTAR standard scores and whether or not participant broke the 12ET rule, though there were several outliers that would affect the median. Inspection of the participant data scores revealed that two different control participants broke the rule at Time 1 than did at Time 2, but three of the same brain-injured participants who broke the rule at Time 1 broke it again at Time 2.

Normative Data for the 12ET

As none of the attribute variables were related to the 12ET in a clinically meaningful way, unstratified normative data have been reported for 12ET variables based on the performance of the participants at Time 1 (see Table 16). The table includes the range of obtained scores and associated percentiles (or percentages for dichotomous variables) for brain-injured and control participants on the 12ET variables.

The test-retest reliability data for the 12ET variables is presented in Table 17. For Number of Tasks Attempted, 60% of the variance of control participants’ scores at Time 2 is accounted for by scores at Time 1. Whether or not a control participant broke the rule of the 12ET at Time 1 accounted for only 28% of Rule Breaks at Time 2. Similarly, the Number of Time Requests made by control participants at Time 1 predicted only 23% of the same variable at Time 2.

To compare the reliability of the 12ET with established neuropsychological tests, control participants’ Total Move Score on the TOL at Time 1 predicted 12% of their score at Time 2. Control participants’ predictive scores on the WCST were 44% for
Table 16. Normative Data for the 12ET at Time 1

<table>
<thead>
<tr>
<th>Continuous Variables</th>
<th>Brain-injured</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Score</td>
<td>Frequency</td>
</tr>
<tr>
<td>12ET Tasks Attempted</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>12ET Time Requests</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12ET Tasks Completed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12ET Time Deviation (seconds)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequencies are grouped</td>
<td>293</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>477</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>618</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>830</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>1180</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>1282</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1493</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1548</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dichotomous Variables</th>
<th>Frequency</th>
<th>Percent in each category</th>
<th>Frequency</th>
<th>Percent in each category</th>
</tr>
</thead>
<tbody>
<tr>
<td>12ET Rule Break?</td>
<td>NO</td>
<td>19</td>
<td>73%</td>
<td>NO</td>
</tr>
<tr>
<td></td>
<td>YES</td>
<td>7</td>
<td>27%</td>
<td>YES</td>
</tr>
<tr>
<td>12ET &gt;7 Minutes on one task?</td>
<td>NO</td>
<td>22</td>
<td>85%</td>
<td>NO</td>
</tr>
<tr>
<td></td>
<td>YES</td>
<td>4</td>
<td>15%</td>
<td>YES</td>
</tr>
</tbody>
</table>
Table 17. Test-retest Reliability Coefficients for the 12ET over an 8 week interval

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>r = 0.78</td>
<td>r = 0.53</td>
<td>r = 0.48</td>
<td>r = 0.36</td>
<td>NA</td>
<td>r = 0.75</td>
</tr>
<tr>
<td></td>
<td>SEP=1.57</td>
<td>SEP=0.230</td>
<td>SEP=1.37</td>
<td>SEP=1.14</td>
<td></td>
<td>SEP=166.2</td>
</tr>
<tr>
<td>Brain-injured</td>
<td>r = 0.68</td>
<td>r = 0.22</td>
<td>r = 0.49</td>
<td>r = 0.56</td>
<td>r = 0.33</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEP=2.24</td>
<td>SEP=0.441</td>
<td>SEP=2.27</td>
<td>SEP=0.779</td>
<td>r = 0.74</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SEP=0.379</td>
<td>SEP=238.8</td>
<td></td>
</tr>
</tbody>
</table>

Tasks Att. = Number of Tasks Attempted, Time Req. = Number of Time Requests, > 7 Mins? = Whether or not participants spent longer than 7 minutes on one task, Time Dev. = Time Deviation

Categories achieved, and 69% for perseverative errors. The RFF Number of Unique Designs obtained by the control participants at Time 1 predicted 44% of their score on that task at time 2. Similarly, the control participants’ Time Estimations score at Time 1 predicted 32% of those scores at Time 2. The control participants’ Trailmaking A-Trailmaking B difference score at Time 1 was a poor predictor of Time 2 scores, accounting for only 13% of the variance.

Recommendations for Reliable Change Index

The RCI recommended by this author is the modification of the Chelune (1993) index used by Basso et al. (1999) as defined above in Equation 5. It is repeated here to aid the reader.

\[
RCI = \frac{[(X2 - X1) - (M2 - M1)]}{SEP} \quad \text{Equation 5}
\]

The SEP is as defined above in Equation 4. In the RCI formula, X1 and X2 are the individual’s scores at Time 1 and Time 2 and M1 and M2 are the group mean scores from the individual’s respective group (brain-injured or control).

Reliable Change on the 12ET

The 12ET was administered to both groups on two occasions, eight weeks apart, to determine the reliability of the test over that interval. Through regression, it was
Table 18. Normative Data for the 12ET at Time 2

<table>
<thead>
<tr>
<th>Continuous Variables</th>
<th>Brain-injured</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Score</td>
<td>Frequency</td>
</tr>
<tr>
<td>12ET Tasks Attempted</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>12ET Time Requests</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4-12</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12ET Tasks Completed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12ET Time Deviation (seconds) Frequencies are grouped</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12ET Rule Break?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO</td>
<td>171</td>
<td>1</td>
</tr>
<tr>
<td>YES</td>
<td>440</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>486</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>743</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>1058</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>1079</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>1322</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>1640</td>
<td>1</td>
</tr>
<tr>
<td>12ET &gt;7 Minutes?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>YES</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>
determined that the brain-injured group and control group obtained the same benefit from practice on all the 12ET variables. In addition, it did not appear that the attribute variables (in this case general intelligence, working memory or memory) mediated learning on the second testing occasion. These findings allow one to use the listed reliabilities to determine whether or not an individual’s score has improved reliably in the absence of any other variables. The reliability index in the SEP is the test-retest reliability index for each of the measures from the individual’s group (see Table 17).

To illustrate an example, the RCI formulae for 12ET Number of Tasks Attempted in a brain-injured examinee (using the less stringent test of \( p = .10 \)) would be as follows:

\[
RCI = \frac{[(X_2 - X_1) - (M_2 - M_1)]}{SEP}
\]

1.645 = \[
\frac{[(X_2-X_1) - (8.58-6.73)]}{2.24}
\]

1.645 = \[
\frac{[(X_2-X_1) - 1.85]}{2.24}
\]

3.696 = (X_2-X_1) – 1.85

5.54 = X_2-X_1

Therefore, a brain-injured examinee would have to improve her 12ET Number of Tasks Attempted score by 5.54 tasks (rounded to 6 tasks as it is impossible to attempt 0.5 of a task) in order for her performance to be considered a significant change over and above that expected by practice and measurement error.

The reliable change scores for the 12ET continuous variables are listed below (see Table 19). All decimals have been rounded. Tasks Attempted and Time Requests are expected to increase with practice, however Tasks Completed and Time Deviation are expected to decrease with practice. Because dichotomous variables do not have a mean score, they do not lend themselves to reliable change calculations.
Table 19. 12ET Scores Required for Reliable Change after an 8 Week Interval.

<table>
<thead>
<tr>
<th></th>
<th>Tasks Attempted</th>
<th>Time Requests</th>
<th>Tasks Completed</th>
<th>Time Deviation (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>4</td>
<td>2</td>
<td>-1</td>
<td>-75</td>
</tr>
<tr>
<td>Brain-injured</td>
<td>6</td>
<td>5</td>
<td>-1</td>
<td>-394</td>
</tr>
</tbody>
</table>
DISCUSSION

This dissertation evaluated the psychometric properties of the Twelve Elements Test (12ET) in brain-injured and matched control participants. Participants completed the 12ET along with published neuropsychological measures on two testing occasions. In addition, brain-injured participants completed a self-report executive dysfunction questionnaire and had a spouse or other informant who knew them well complete the informant version of the questionnaire on two occasions. Other neuropsychological measures were included in the battery to assess convergent and divergent validity. It was hypothesized that matched control participants would perform better on the 12ET test variables than brain-injured participants. In addition, it was hypothesized that scores on variables of the 12ET would be significantly correlated with performance on certain published neuropsychological test variables. Finally, to add to the usefulness of the test, normative data and reliable change indices for the 12ET were provided.

Participant Matching

As can be seen in the results, the matched control group’s WTAR estimated IQ was significantly higher than the brain-injured group’s WTAR estimated IQ. This finding was unexpected, as a great deal of attention was paid to optimizing the age and education match between the two groups. A possible explanation for this finding could be that the brain-injured and control participants were not drawn from the same population, and despite their apparent match on age and education variables, the control group was drawn from a population with a higher average IQ than the brain-injured participants. The greater number of brain-injured participants with WTAR standard scores below 100 as
compared to the control group, despite apparently similar education levels between the groups, may have led to lower IQ scores on average. While confrontational reading ability is reported to be robust to brain injury (WTAR Test Manual; Psychological Corporation, 2001), it is also possible that the brain-injured participants, particularly those who were injured early in life, had lower confrontational reading performance than would be expected based on their age and education. It may be that the small difference in education between the two groups coupled with the fact that some of the brain-injured participants were injured early in life, led to a large difference between estimated and observed confrontation reading skills. Support for this notion comes from the TBI Model Systems data set, which found evidence to suggest that a small difference in premorbid education in brain-injured samples led to significant differences in post injury employability ratings. From the TBI Model Systems data, individuals with 13 years of education were rated as 1.55 times more employable than those with 10 years of education (Sherer, Hart, Nick, Whyte, Thompson, & Yablon, 2003).

Past research on the relationship between executive function tests and intelligence informs the findings of the present study. In this analysis, the WTAR, a measure of crystallized intelligence, was significantly correlated with performance on some 12ET variables for the brain-injured participants, but not for the control participants. This finding was unexpected as it is discrepant from past research (Duncan et al, 1995; Friedman et al. 2006) where performance on executive tasks was more strongly related to fluid than crystallized intelligence in brain-injured groups. It is possible that the lack of relationship in control participants could be related to the fact that the average of crystallized intelligence scores correlate better with $g$ than individual scores and, to
observe a significant correlation, this study would have needed more than one test of crystallized intelligence. Alternatively, this finding may reflect the significant difference in IQ between the two groups. The discrepant WTAR performance between the groups may have confounded the results with the brain-injured group being expected to obtain lower scores on executive function tasks based on their lower IQ rather than their lack of a brain injury. In other words, any differences observed on the 12ET, may actually be a reflection of \( g \). If this were the case, one would expect the brain-injured participants’ WTAR scores to be significantly related to all of the executive function tests. The WTAR was related to several measures in the brain injured group. The most notable difference between the control and brain-injured participants’ intercorrelations for the neuropsychological measures was the relative lack of correlation between the executive function tests and the WTAR for the control group (the WTAR was only significantly correlated with Trailmaking B, and Trailmaking A – Trailmaking B for the control group) as compared to the brain-injured group (the WTAR was significantly correlated with Trailmaking B, Trailmaking A – Trailmaking B, TOL Total Move Score and WCST Categories). This suggests that a common factor, \( g \), was not responsible for the variance in the executive function performance of the control group, but it may have had a larger impact for the brain-injured group. The relationship between the 12ET and intelligence will be revisited below.

Working Memory and Memory

Despite the significant difference in IQ estimates between the two groups, the groups did not differ on a measure of working memory. This similarity between the groups allowed for the assumption that any differences in test scores on the executive
function tasks was not due to deficient working memory in the brain-injured group. In contrast, the matched control group performed significantly better on the verbal memory task than the brain-injured group in terms of free recall, and number of repetitions. Therefore, differences between the groups on the executive function measures could have been mediated by deficits in verbal memory and self monitoring in the brain-injured group.

**Part 1. Evaluating the Validity of the 12ET**

Contrary to the hypothesis that brain-injured participants would perform more poorly on the 12ET variables than the control participants, there were no significant differences between the two groups on the two tested variables of the 12ET despite a trend towards significance for the Number of Tasks Attempted. Both groups performed the same with respect to whether or not they broke the rule (it was very rare for both groups). Given these findings, it might be argued that non parametric tests, such as the Mann-Whitney test, would have been more appropriate than the t test for the analysis of some of these data. Non parametric tests are less powerful than parametric tests when the assumptions of the parametric test are met. However, if the assumptions of the parametric test are not met (such as non normally distributed data with heterogeneity of variance) as with the current data set, the Type I error rate would not be 5% as it would not be based on a normal distribution. Non parametric tests only have an increased chance of a Type II error if the data is normally distributed (Field, 2005). However, as noted above, the t test is robust to violations of its assumptions. Therefore, using the modified t test, with equal variance not assumed, allowed for retention of as much power as possible while
maintaining the integrity of the analyses. The lack of significant results was likely not
due to the choice of statistic, but it may have been due to a lack of power.

Because the Number of Tasks Attempted t test had a small to moderate effect size
(Cohen, 1992), there may not have been enough power to demonstrate a significant
difference between the two groups. A slightly larger sample of participants may have
been adequate to detect a statistically significant difference between the two groups.
However, because the ABI group did not demonstrate significant executive deficits, the
similarities in 12ET scores between the groups may simply reflect good planning and
organisational skills in both groups. It may be useful at this point to look carefully at the
differences between the Six Element Test, and 12ET. Both the SET and 12ET are
complex tasks, and success or failure of these tasks can result for a number of reasons.
An examinee may fail due to damage to non-executive (peripheral) systems such as poor
verbal comprehension (Burgess, 1997). Alternatively, an examinee may fail to form a
plan, or fail to follow their plan. There is evidence of separate executive processes
systems underlying the ability to follow a plan (i.e. prospective memory), separate from
(but linked to) other behaviours (Cockburn, 1995; Einstein & McDaniel, 1990).

Chan and Manly (2002) administered a translated version of the SET to
neurologically healthy controls and brain-injured participants in China. They found the
control group’s profile score on the SET ($M = 3.58$, $SD=0.56$, range 2-4) was very similar
to that reported by Wilson et al, (1996; $M = 3.52$, $SD = 0.80$). The authors also reported
the SET Number of Tasks Attempted was at ceiling levels for the healthy control group
($M = 5.68$, $SD = 0.68$) but significantly lower for the brain-injured group ($M = 4.47$ $SD =
1.48$, $t= 4.28$, $p <.001$). As with the 12ET, participants in both groups did not tend to
break the rule of the task, and the groups were not significantly different on this variable. The authors concluded that the basis for the lower scores on the SET by the brain-injured participants was that they got caught up in one task to the detriment of the overall goal. The results of the current study are consistent with this hypothesis and will be further discussed below.

The increase in difficulty on the 12ET as compared to the SET did not have the intended effect of more clearly separating the brain-injured and control groups. In fact, the mean scores, and range of scores on the 12ET test variables were very similar for both groups and the means overlapped when the standard deviations were taken into account (similar to the overlapping means for the SET reported earlier). That said, the increase in range reduced the number of participants who performed at ceiling levels overall, though both groups contained individuals who performed at ceiling levels on both testing occasions.

The fact that many control participants performed quite poorly on the 12ET merits comment. When we consider why people fail tests such as the SET and 12ET, we tend to search for an organic deficit related explanation when a brain-injured person fails a neuropsychological test. However, this explanation clearly is inadequate when we try to understand the poor performance of neurologically intact individuals. It is not uncommon for healthy individuals to fail executive tests and distributions of scores often include a few control participants who perform as poorly as the brain-injured participants (Burgess, 27/04/03, personal communication). It is possible that the reason for the large range in test scores in the control group lies in the above discussion of g, and Duncan’s (1995) finding that healthy controls at the low end of the distribution of intelligence often
Personality and Neuropsychological Test Performance

The hypothesis that the Conscientiousness factor of the NEO-FFI (which refers to a preference for structure and self discipline with respect to goal directed behaviour) would be related to the Number of Tasks Completed on the 12ET was not supported. However, in the control participants, Conscientiousness was related to 12ET Number of Time Requests, with control participants who scored higher on Conscientiousness tending to ask about the time remaining less frequently. Perhaps these participants were confident in their ability to plan and organize their behaviour, given their preference for structure and self discipline. Alternatively it may have been more important for these participants to focus on obeying the rules and doing as many of the twelve tasks as possible. Their increased confidence, or increased focus on the task and rule may have been reflected in less need, or less willingness, to request the time remaining.

As with the pilot study, control participants tended to complete tasks on the 12ET, despite being expressly instructed not to do so. In Part 2 of the current study, we saw that intelligent and quick control participants were more likely to complete tasks. Were these participants overconfident in their ability to be able to complete several tasks and still attempt all 12? Clearly, the Conscientiousness factor of the NEO-FFI did not clarify the reason for participants choosing to complete tasks. However, it is possible that while Conscientiousness was not related, another personality factor may well have been. Could the elusive variable be self confidence? Stankov and Crawford (1997) reported that self confidence did not predict performance on neuropsychological tests and feedback about
performance did not change confidence ratings. They noted that, in general, people tend to be overconfident when predicting their own performance. Others have also noted that people systematically overestimate their own abilities, and that this misperception may lead them to make poor decisions (Larricka, Buronb & Solla, 2007). Moreover, those who believe they are better than average, as the participants in this study with a high IQ would likely believe, are less likely to listen to the advice of others (Gino & Moore, 2007). Could this overconfidence and over-evaluation of abilities be applied to these data? Overconfidence may have been related to the unique aspects of the 12ET in that it is a relatively unstructured task with no feedback provided. The instructions for the 12ET explicitly say “There is NO WAY you can finish all of the tasks in the time provided”. Rather than a sensible warning to organize one’s time, could this instruction be taken as a challenge to a bright individual who is overconfident about her abilities or, as the research above suggests, simply ignored? More importantly, could this overestimation of ability help explain why participants do not improve their 12ET scores by a significant margin on the second testing occasion?

Part 2. External Validity

Based on the above discussion of the expected intercorrelational test differences for brain-injured and healthy control participants, it is appropriate to discuss the results of this dissertation in terms of each group of participants.

Convergent Validity for the Brain-injured Participants

There is clearly a relationship between executive function tests and measures of general intellectual functioning, though the exact nature of the relationship remains under
debate. The interpretation of the correlations for the brain-injured participants between the WTAR and the 12ET test variables would appear to be consistent with past research for other executive tests. The 12ET variables Number of Tasks Attempted, Time Deviation, Number of Time Requests, and whether or not participants spent >7 minutes on one task were all significantly correlated with the WTAR. For the brain-injured participants, general intellectual functioning as measured by recognition reading (crystallized intelligence) was related to performance on the 12ET. In the brain-injured group, several other executive tests were also related to the WTAR, such as the WCST Categories Completed, TOL Total Move Score, and Trailmaking A - Trailmaking B. This suggests the 12ET and other executive tests all shared variance that was likely accounted for by g. The 12ET Rule Break variable was not correlated with the WTAR standard score, but this may be due to the lack of data points on this variable.

The 12ET and Memory

The correlations between CVLT II Immediate memory score and the 12ET variables for the brain-injured participants were non significant, suggesting poor memory in the brain-injured participants was not a contributor to 12ET scores. This would be expected, as any memory deficits were supported by providing written instructions in full view for the entirety of the 15-minute test.

The 12ET and Tests of Cognitive Flexibility

Trailmaking B was significantly related to one 12ET variable in the brain-injured group (spending >7 minutes on one task). The relationship between 12ET Number of Tasks Attempted and Trailmaking B was significant at the .05 level, but because this was
a planned comparison, it was no longer significant following the Bonferroni correction.

Trailmaking B is considered to be a measure of visual scanning, divided attention and cognitive flexibility (Wood & Liossi, 2007), however the difference score between Trailmaking A and Trailmaking B may be a more relevant measure of switching, as it controls for visual scanning and processing speed (Salthouse et al., 1998). The difference score was significantly related to three 12ET variables: 12ET Number of Tasks Attempted, Time Deviation, and spending more than 7 minutes on one task. There was a trend towards significance for the correlation between the Trailmaking difference score and 12ET Rule Breaks. The 12ET scores that had significant correlations, were highly intercorrelated, therefore only the relationship between the Trailmaking difference score and 12ET Number of Tasks Attempted will be discussed in more detail. Both 12ET Number of Tasks Attempted and the Trailmaking difference score were significantly related to the WTAR in the brain-injured group. As a result, it is difficult to determine what portion of the variance was accounted for by intelligence, and what portion was attributable to executive function. For the moment, let us assume that at least some of the variance shared between these test scores was uniquely executive. In addition to the above mentioned domains, Trailmaking Part B is also a measure of working memory, because it requires the participant to keep in mind where they are in each of the number and letter sequences to avoid errors. The SET has been described to be a test that requires goal directed planning, flexible strategy generation and self monitoring (Chan et al., 2006) and by extension the 12ET may also rely on these processes. That said there are conceptual similarities between the underlying skills needed for success on both the Trailmaking B Task as well as the 12ET. The 12ET however, does not appear to rely on
working memory as heavily as Trailmaking B. Digit Span scores were not significantly correlated with any of the 12ET variables in the brain-injured group whereas Trailmaking B and the Trailmaking difference score were both significantly correlated with Digit Span scores.

Further evidence of convergent validity regarding cognitive flexibility came from correlations between the 12ET variables and the WCST Categories Completed. Again, this relationship is complicated by the significant correlations between the WCST Categories Completed and the WTAR, and general intellectual function may well underlie this relationship. The WCST Perseverative Errors score did not correlate with the 12ET test variables, nor did perseverative errors significantly vary with the WTAR, though these errors were, of course, significantly correlated with the WCST Categories Completed. The WCST Perseverative Errors appear to measure unique variance that may be executive. However, the correlations with the WTAR for the 12ET, Trailmaking and the WCST Categories Completed may weaken the argument that the shared variance between the 12ET and tests of cognitive flexibility is due to executive function rather than $g$.

*The 12ET and a Test of Planning*

The 12ET was related to the TOL, a measure of planning, for both groups, though in differing ways. The brain-injured participants’ 12ET Time Requests variable was strongly correlated with the TOL Total Move Score (with fewer time requests made by individuals who made fewer moves on the TOL). 12ET Time Requests might be considered a reflection of self monitoring but also a reflection of overall ability to plan and organize one’s behaviour. Someone who plans their approach to the 12ET at the
outset, may have less need to request the time, as they are aware that by following their plan they will get to as many tasks as possible within the time limit. In addition, 12ET Time Deviations were significantly related to TOL Rule Violations with individuals who made a rule violation on the TOL being more likely to deviate from the optimal amount of time per task on the 12ET. Rule Violations on the TOL consist of taking more than one bead off the pegs at a time (the other Rule Violation of placing more beads on a peg than the peg can hold never occurred). Behaviourally, this action was associated with individuals who appeared to get caught up in the task, and “forget themselves” in that when reminded of the rule they invariably acknowledged they remembered it but said they “forgot it momentarily”. It was also noted that individuals who broke the rule of the TOL task typically expressed frustration with the Tower Task, and would comment that the particular item “couldn’t be done”. One wonders if these people also tended to get “caught up” in doing a task on the 12ET, to the detriment of the other tasks, which would lead to large Time Deviation scores. Because the TOL Rule Violation was not correlated with the WTAR in the brain-injured group, these significant relationships can be more confidently attributed to uniquely shared executive function variance than some of the correlations described above.

The 12ET and Questionnaire Data

The DEX was completed by 22 brain-injured participant and informant pairs. The difference between the DEX Informant Report and Self Report was considered to be the variable of interest. This difference score was significantly correlated with whether or not participants spent longer than 7 minutes on one task. This significant correlation may
reflect poor self monitoring (poor insight into the extent of difficulties on the DEX and poor awareness of the passage of time and the overall goal of the task on the 12ET).

Divergent Validity for the Brain-injured Participants

*The 12ET and Visual Scanning Speed*

From these data, the ability to quickly scan through a complex visual array was important for success on the 12ET. This makes sense, given the size of the space needed to arrange the 12ET materials, and the fact that it is a complex visual array (see Figure 1). In retrospect, Trailmaking A was not a good choice of measure of divergent validity for the 12ET as it also requires quick scanning through a complex visual array. Moreover, despite the significant correlation between Trailmaking A and the 12ET variables, the scatterplot showed a tremendous amount of scatter. Therefore while the 12ET may indeed require efficient visual scanning for task completion (as there is a great deal of visual information presented on the table) the scatter plot suggests the possibility that the significant effect may be a result of error, noise in the data, or chance.

Given the complexity of the 12ET, it is difficult to predict what test would show clear divergent validity. From the data, the Stroop Interference score had the lowest correlations with the 12ET variables for the brain-injured participants. Inhibition of a prepotent response might not be expected to correlate strongly with the 12ET. Another possible contender for divergent validity testing is the RFF Unique Designs Score. Even though figural fluency is considered a measure of executive function, strategic generation of figures was not related to performance on the 12ET particularly for the brain-injured participants. Similarly, WCST Perseverative Errors were unrelated to performance on the 12ET, but this divergence was quite unexpected. One might expect a perseverative
tendency to be related to a tendency to spend too long on one task, but in these data this was not the case. Trailmaking A was chosen, in part, because it was a convenient measure to use as it was already part of the battery. However, a more specific measure of visuospatial function, such as the short form of the Hooper Visual Organization Test (Merten, 2004) may have provided better a measure of divergent validity.

12ET Intercorrelations in the Brain-Injured Group

Several of the 12ET variables were highly correlated with each other, and while some of these correlations were expected (e.g. the correlation between Time Deviation and spending > 7 minutes on one task) some of them were quite interesting and might help explain participants’ performance on the test. It is interesting to note that those brain-injured participants who requested how much time was remaining (at least once) tended to attempt more of the 12 tasks than those who made no, or only one, time remaining request. This suggests that those participants who remembered to check the time, were able to use that information to help refocus their attention on the task at hand. Qualitatively, on hearing how much time was remaining, participants would often say something like “Oh dear, I’d better hurry up” and then soon after they would often stop what they were doing and switch to another task. Asking for the time remaining, could be considered a task that must be remembered prospectively. Applying Shallice and Burgess’s (1991) marker hypothesis, the participants’ executive system must set up a marker so that when they reach that point in the task, they will complete the proposed activity (e.g. to check the time remaining, and/or switch tasks). The brain-injured participants’ behaviour on the 12ET is reminiscent of that observed by Manly, Hawkins, Evans, and Robertson (2002) when they improved participant scores on the Hotel Task (a
more ecologically valid multitasking version of the SET) by having an alarm sound periodically on six occasions (there were six different tasks on the Hotel Task and participants were given 15 minutes). Participants had been instructed, that on hearing the alarm, they should consider their current behaviour and overall goal. These authors suggested that providing support to one aspect of executive functioning (in this case goal neglect) could help brain-injured individuals make better use of their intact executive skills. Of course, there were important procedural differences in the Hotel Task, as compared to the 12ET. The first difference was that for the Hotel Task, the examiner sat out of the participant’s view. The second difference was that the maximum number of tasks attempted on the Hotel Task was five, though there was a sixth task of opening garage doors at predetermined times. The third difference was that a clock was placed on the table in view of the participants (and it was covered with a towel so the examiner could note when participants checked the time). It is unknown what effect having the examiner sit outside the participants’ view would have, but it is likely that keeping the clock on the table in the participants’ field of view cued participants to check the time more frequently than having them ask the examiner for the time remaining. With these distinctions in mind, it appears from the current data, that participants who were able to prospectively remember to check the time remaining on the 12ET, were successful in keeping their overall goal in mind. It would have been very interesting to have included a condition where a periodic alert sounded for half of the brain-injured and control participants, to see if an alert improved the participants’ performance. However, the focus of this study was psychometric evaluation of the 12ET as an extension of the SET, rather than maximizing performance on the task with external manipulations such as an alert.
As above, 12ET Time Requests were significantly related to planning ability, as measured by the TOL Total Move score. Planning is not a unitary activity that occurs once at the beginning of an activity. It is a dynamic process that is activated each time an individual completes one task, meets a situation for which their plan is no longer working, or meets an unexpected event (Shallice & Burgess, 1991). Therefore the significant correlations between the TOL, Time Requests, and number of Tasks Attempted support the hypothesis that the 12ET is a test of planning, and that Time Requests interrupted activity and helped participants modify and maximize their planned behaviour.

Inter correlations between the Executive Function Tests in the Brain-Injured Group

It was interesting to see that the WCST Perseverative Errors were significantly correlated with several executive variables including the RFF Total Unique Designs Score, the Trailmaking difference score, and CVLT II repetitions. Moreover, WCST Perseverative Errors were also significantly related to Digit Span, CVLT II Total Immediate Memory and Trailmaking A. the brain injured participants’ scores on these variables were significantly lower than the matched controls, suggesting the WCST may be sensitive to the effects of brain injury. Unfortunately, as above, the WCST Perseverative Errors did not correlate with the 12ET.

Convergent Validity for the Control Participants

While the t test on the difference between the groups on the 12ET Number of Tasks Attempted did not quite reach significance, the control participants exhibited quite a different pattern of results on the 12ET than the brain-injured group. Not only did the
control participants tend to complete more tasks on the 12ET (as reported in Part 1), but this tendency was significantly related to the WTAR. There were also significant relationships between Trailmaking A, RFF Unique Designs and 12ET Tasks Completed for the control group. Individuals with fast visual scanning tended to complete more tasks on the 12ET, as well as create more unique designs on the RFF. While working as quickly as one can was a reasonable tactic for test taking, completing tasks on the 12ET was not, so having quick processing and visual scanning speed may actually have worked against participants. It appears that intelligent people were quick and efficient at completing tasks, however on the 12ET, this quality may have led to difficulty attempting all 12 tasks (and thus interfered with obtaining optimal scores). These correlations may help explain why control participants did not score at ceiling levels on the 12ET Number of Tasks Attempted.

The 12ET and Memory

Unlike the brain-injured group who had non significant correlations between 12ET test variables and the CVLT II, the correlation between CVLT II Immediate memory score and 12ET Time Requests was significant for the control group. That is, individuals with better memories, tended to request the time remaining more often. Perhaps those individuals with strong memory skills frequently remembered to check the time remaining, whereas other individuals forgot to do so. However, CVLT II and 12ET Time Request scores were not related to the Number of 12ET Tasks Attempted. Therefore these time remaining requests did not appear to have an effect on the overall performance on the 12ET for the control participants as they did for the brain-injured participants.
The 12ET and Tests of Cognitive Flexibility

Neither Trailmaking B, nor the Trailmaking A – B difference score was significantly related to any 12ET variables in the control group. The WCST was also nonsignificantly correlated with the 12ET in the control group, but because the groups did not significantly differ on the WCST variables, this result was not surprising. As would be expected, there was less range in the control participants’ scores on these tasks than in the brain-injured group, and this may have reduced the possibility for significant correlations.

The 12ET and a Test of Planning in the Control Group

The 12ET Rule Break variable was significantly negatively correlated with the TOL Rule Violation score. Moreover, the control participants made significantly fewer TOL Rule Violations than the brain-injured group, but a similar number of 12ET Rule Breaks. That is, individuals who broke the rule of the 12ET, did not necessarily make a rule violation on the TOL suggesting the two rules are conceptually distinct in the control group, but may tap similar processes in the brain-injured group. The TOL Total Move score was also significantly negatively related to the 12ET Rule Breaks. Again, doing well on the TOL did appear to be related to the ability to follow the rule on the 12ET. The pattern of results also suggests that the 12ET rule (of not attempting two parts of the same subtask consecutively) may be more difficult for participants to follow than the TOL Rule (or removing only one bead off the pegs at a time). Finally, the control group showed the same pattern of relationship as compared to the brain-injured group with respect to time requests. TOL Total Move score was significantly negatively correlated with 12ET Number of Time Requests. Those individuals who made few moves on the
TOL (which suggested intact planning skills), tended not to request the time remaining on the 12ET.

Divergent Validity for the Control Participants

*The 12ET and Visual Scanning Speed*

For the control group, Trailmaking A was significantly correlated with 12ET Number of Tasks Attempted, and Time Deviation. The discussion above regarding participants’ ability to manage a complex visual array and general processing speed is applicable.

*12ET Intercorrelations for the Control Group*

Unlike the brain-injured group, the 12ET number of Tasks Attempted were not significantly related to the Number of Time Requests for the control participants. It is possible that brain-injured participants used Time Requests as a compensatory strategy to help them keep track of their behaviour. Time Requests could be incorporated into a “self talk” compensatory strategy that would allow brain-injured participants to improve their overall score on the 12ET Tasks Attempted. Control participants would not be expected to require such a compensatory strategy for success on the 12ET and may therefore have less need to request the time remaining.

Intercorrelations between Published Neuropsychological Tests.

Few of the executive function variables were significantly correlated in the control group, and the correlations between the executive function tests and the WTAR IQ estimate were low except for the Trailmaking Test.
Differences in Performance on Published Executive Function Tests.

It was expected that the two groups would differ on all of the executive function measures, with the matched control participants performing better than the brain-injured participants. In fact, this was observed for some, but not all tests. The matched control group did perform significantly better on the Trailmaking B task, which is a measure of cognitive flexibility, in addition to visual scanning speed. This is in line with past research as Trailmaking B is a well recognized measure of executive dysfunction and fMRI studies have confirmed its sensitivity to left frontal lobe lesions (Zakzanis, Mraz, & Graham, 2005). In addition, the matched control group performed significantly better on the Ruff Figural Fluency task in terms of figure generation than the brain-injured group, providing support for the sensitivity of the RFF to brain injury. The brain-injured group performed equally well on the Tower of London Total Move Score, but made significantly more TOL Rule Violations than the control group. Rule Violations on the TOL Drexel version have been found to be the most sensitive variable on other studies as well (Riccio, et al., 2004).

There were also some unexpected findings where the brain-injured group performed as well as the control group. Both groups performed similarly on the WCST, a test that reportedly measures novel problem solving and cognitive flexibility. In addition, the brain-injured group did not demonstrate more sensitivity to interference on the Stroop test than did the matched control group. Finally, the brain-injured group’s ability to estimate the passage of time as measured by the Time Estimation subtest of the BADS was not significantly worse than the control group’s scores. The similar scores between the two groups on the Time Estimation task may be misleading. Rather than both groups
scoring high on this task, both groups scored very poorly on this task in keeping with other reported results (Gillepsie et al., 2002; Ponsford, et al., 2005). This may reflect the fact that the test was developed and normed in Britain, or may also reflect low validity of this particular subtest of the BADS as the normative British sample also scored relatively low. As with other reports, Time Estimation was not a valid measure of Temporal Judgment in this sample.

Overall, while the two groups scores typically differed in the expected direction on the executive function tests, the scores were not always significantly different. These similarities between the groups might be expected to lead to a similarity in test scores on other tests of planning, including the 12ET.

The brain-injured group was made up of participants with heterogeneous injuries, including stroke, aneurysm, and traumatic brain injury. By including participants with acquired brain injuries (ABI), rather than restricting the sample to only those participants with traumatic brain injuries (TBI), it was possible that the brain-injured group may have discrepant cognitive deficits, and there might have been several participants who did not exhibit executive deficits. When the brain-injured group was broken into groups based on their injury the only test they differed on was the 12ET, which suggests the ABI participants and TBI participants were not significantly different in their cognitive deficits. Interestingly, individuals with TBI tended to perform better on the 12ET in terms of Number of Tasks Attempted, despite being more likely to break the rule of the task. Therefore, while the TBI group did appear to exhibit a deficit that impaired their ability to follow the 12ET rule (perhaps a memory deficit, self monitoring deficit, or tendency towards impulsivity) this impairment did not reduce their overall score on the 12ET
Number of Tasks Attempted. However, as noted in the results, on average, the ABI group did not demonstrate executive impairment. Therefore, the small sample of ABI participants in this study, for one reason or another, was not representative of the brain injured population who demonstrate executive impairment, and were therefore a poor comparison group for the evaluation study.

Part 3. Normative Data and Reliability of the 12ET

Normative Data

The normative data for the 12ET was presented as population derived frequencies, rather than mean scores. This is because the distributions for both the brain-injured and control groups were not normal, and traditional statistically driven norms (i.e. –3 SD from the mean is abnormal) are not appropriate. Neuropsychological test data are often non normally distributed. They have a low ceiling, as they are designed to detect impairment. As such, the data might be expected to be negatively skewed, particularly for the matched control group. This is indeed the case for the 12ET Number of Tasks Attempted (See Figure 2). While neither distribution was normal, the brain-injured group’s scores were closer to a normal distribution than the negatively skewed control group’s data. However, the control participants’ scores were not as negatively skewed as expected. As noted above, many control participants did not perform the 12ET at ceiling levels. Because the two group means did not significantly differ, the normative data shows the expected overlap in test scores. Unfortunately, this overlap suggests the 12ET is not a sensitive measure of brain injury.

Practice Effects and Reliable Change on the 12ET
Because of the factors listed in the introduction, the reliability of established executive function tests is notoriously poor. As a result, the reliabilities of the 12ET were expected to be similarly low. When one considers executive function tests, it is usually preferable to choose a measure with a higher level of validity even if it is slightly less reliable (Strauss, Sherman & Spreen, 2006). However, as the reader can see, some of the 12ET variables have low reliability coefficients, whereas some are almost equal to the minimum of 0.80 recommended by Anastasi and Urbina (1997). In fact, the 12ET Number of Tasks Attempted, and 12ET Time Deviations were two of the most reliable variables of the test battery employed in this study for both groups. These strong test-retest correlations, however, were tempered by reasonably large standard deviations. Therefore the 12ET may suffer the same fate as many executive function tests; a very large increase in scores may be necessary to conclude that an observed change is reliable.

From Table 17, it is clear that large changes on the 12ET Number of Tasks Attempted would be required to conclude the change had been reliable. This is a result of the large Standard Error of Prediction for this variable in both groups. The 12ET variables that showed the most promise as sensitive measures of reliable change are the Number of Tasks Completed, Time Requests and Time Deviation variable for the control group, and the Number of Tasks Completed and Time Deviation for the brain-injured group.

One might look at these reliable change scores and consider the 12ET to be an unsuitable tool to assess change over time. When compared to an intelligence test (such as the WTAR with an r >.90), that is a fair characterization. However, when compared to other executive function tests, the 12ET has similar reliability coefficients (TOL Total
Move score $r = .75$; the information to calculate SEP is not available in the TOL manual. Given the relatively low reliability scores and large changes needed for reliable change, an important question to consider is whether or not reliable testing of complex executive functioning is possible. It may be that the quest for an executive function test that has good test-retest reliability and small reliable change indices is impossible given the nature of the tests. Test-retest reliability as defined by classical test theory, is used to assess the construct validity of a test, by ensuring scores are the same from one testing occasion to the next, or that everyone in the sample increases or decreases their score by a similar amount. Because the construct validity of an executive function test depends on its ability to measure information processing during novel problem solving, repeated testing will necessarily reduce the construct validity of the test. Therefore, traditional psychometrics derived from classical test theory may be inappropriate for executive function tests that are inherently unstable. The authors of the WCST support this view, and have reported generalizability coefficients rather than reliability coefficients in the test manual (Heaton et al., 1993). This author chose to report the normative data of the 12ET in terms of frequencies rather than mean scores in agreement with this view. Perhaps then, rather than attempting to develop a version of the SET that has greater range and more difficulty (to allow for repeated testing) it may be more appropriate to try to develop a completely different test of the same construct (de Zubicary, Smith, Chalk & Semple, 1998) that could be administered on the second testing occasion. Obviously this venture is immensely complicated, but it may be where we should put our efforts if repeat testing of executive function is a desirable goal. The best approach may be to accept that we must assess executive functioning as the poorly defined heterogeneous groups of cognitive
functions it is (Bennet, Ong & Ponsford, 2005). That is, measure it similarly to intelligence, with a battery of subtests each requiring a discrete executive skill. The Behavioural Assessment of the Dysexecutive Syndrome (BADS; Wilson et al., 1997), and Delis Kaplan Executive Function System (DKEFS; Delis, Kaplan & Kramer, 2001) are examples such batteries, and batteries such as these may represent the way we will assess executive function from hereon.

Limitations

The sample size in this study was very small, when compared to other reliability and validity studies (e.g. Salthouse et al., 2003). In addition, while the current sample was not particularly small when compared to some normative data sets (e.g. the Trailmaking Test, Strauss et al, 2006) it was quite small when compared to others (e.g. WAIS III; Wechsler, 1997). As a result, the characteristics and performance of the brain-injured and matched control groups may not be representative of the general population.

With respect to the brain-injured participants, most of the participants in this study reported PTA consistent with a severe brain injury. In the general population, the majority of acquired brain injuries are mild (Sherer et al., 2003). Therefore this group of brain-injured participants likely performed more poorly than a larger, more heterogeneous group of brain-injured participants. Of course, this weighting of the sample as having more severe brain injuries might have maximized the differences in performance on the 12ET, so one could consider this study an evaluation of the 12ET under conditions that would be most likely to detect a meaningful difference between the two groups. In addition, ABI severity was not confirmed through chart records for the brain-injured participants. Obtaining ethical permission to access patient charts is very
difficult, and significantly invades participant privacy. Therefore, the severity of brain injury was estimated from the brain-injured participants’ self report (and confirmed by a significant other if they were present). All participants, except one mildly brain-injured participant, had completed a rehabilitation programs. Therefore the likelihood that these participants had suffered a genuine brain injury was high. Having said that, it is clear that despite their brain injuries, these participants did not, as a group, exhibit executive impairment. This was unexpected, and unfortunately made it difficult to accurately assess the 12ET’s sensitivity to executive impairment.

A limitation of the study was the choice of subtasks for the 12ET. The SET is made up of a dictation task, a picture naming task, and an arithmetic task. These three tasks were preserved in the 12ET to maximize the similarities between the SET and the 12ET. In addition to these three tasks, the 12ET also included a card sorting task, a money sorting task, and a restaurant menu task, where participants had to write down what they would order off restaurant menus. The restaurant menu task appeared to be problematic, as it required considerably more writing than the other tasks. As a result, participants tended to avoid it. In addition, restaurant task was placed in the upper right corner of the table. While it was in full view, and it was included in the list of tasks in the instructions, some participants claimed they forgot about it, and stated they didn’t see it while they were working. Perhaps their attention was drawn to the tasks that were directly in front of them and, in their haste, they did not scan the entire table. The other two new tasks, money sorting and card sorting, were also different from the original SET tasks because they potentially took up a large amount of space on the table while the participant was working on them. Some participants sorted the cards and money in their
hand (by “fanning” them out) but most participants sorted via laying individual cards or
bills out on the table and sorting onto each individual pile (thereby having approximately
10-15 piles). Because this method of sorting interfered with the completion of other tasks
(because the cards or money were physically in the way), it appeared that some
participants felt compelled to complete the task in order to be able to clean up the cards
or money and move them aside. There were also participants who simply pushed the
unfinished piles of money or cards to one side and continued, and others who worked on
top of the piles of money or cards, though the cards and money left on the table could
have still served as a distraction. An additional concern regarding the sorting tasks was
that they took quite a long time to complete (approximately 5 minutes on average). Many
participants reported that they preferred card sorting over the other 12ET tasks. One
wonders if the addition of a long, but relatively enjoyable task that participants tended to
complete, skewed the data in favour of participants having a higher score on the Number
of Tasks Completed variable than would have been observed without the card sorting
task. In future, it may be useful to consider the affective component of different tasks
when choosing subtasks for tests such as the 12ET.

It may have beneficial to administer a short inventory, such as the Mayo Portland
Adaptability Inventory (MPAI-IV; Malec, Moessner, Kragness, & Lezak, 2000) to
evaluate functional impairment. This 29 item questionnaire only takes 20 minutes to
complete, and would fit nicely into a research or clinical protocol.

Another potential limitation of this study is the relationship of the significant
others to the brain-injured participants. Some significant others were spouses, some were
parents, some were children, and some were professional care providers of the brain-
injured participants. This variability of raters may have had an impact on the comparability of the ratings of the DEX.

Conclusion

It has been proposed by Ardila, Ostrosky-Solis and Bernal (2006) that psychological and neuropsychological batteries fulfill four criteria in order to be considered reliable and valid measures: they have large enough normative databases; the effects on performance of brain damage is known; the areas of the brain that are activated under testing conditions are known; and how the test correlates with other measures is known. To this end, tests must be administered to diverse populations, be clinically useful, and have demonstrated reliability and validity in different populations. The 12ET, along with many neuropsychological measures, does not meet these criteria. The 12ET was developed to be an improvement over the SET in terms of increased sensitivity to mild executive impairment, and increased range of scores to make it more amenable to serial assessments. It is difficult to determine whether the 12ET is indeed an improvement, as the ABI sample used in this study did not demonstrate executive impairments on published measures. The brain injured participants in this study had remarkably good executive functioning (as measured by published tests). The lack of diversity between the groups precluded the ability to draw firm conclusions regarding the 12ET’s ability to measure executive function. The 12ET was considerably more difficult than the SET, and this difficulty was intended to increase the range of test scores to allow for increased sensitivity and to better allow practice effects to be observed without participants obtaining ceiling level scores. While the goal of increased range in scores was certainly met, the increased difficulty of the 12ET was so great, that the test became
too difficult for some brain-injured participants and control participants. For some, it became a test of ability, rather than a test of disability. Neuropsychological tests are designed to be sensitive to impairment, rather than reflect ability. While the 12ET may reflect impairment in those who have genuine executive impairment, it was difficult to observe because many brain injured participants performed relatively well, and many control participants performed relatively poorly, even on the second testing occasion, which reduced the difference between the groups. If there were brain-injured participants with planning and organizational deficits who performed poorly on the 12ET, it was difficulty to see them in the results because of the small sample size. In addition those who had difficulty on the 12ET due to brain injury were mixed in with participants who may not have functionally impaired planning and organization, but who nevertheless performed poorly on the 12ET.

One difference on the 12ET that could be considered as a potential change to the SET was that the math items did not become more difficult as the participants got closer to the end of the booklet, as they do in the SET. This alteration was made in order to minimize the effects of a potential external cue, which may have encouraged participants to stop doing the math items. This change was considered an advantage of the 12ET as no participants were observed to complain that the math items were getting difficult (and subsequently shift to a new task) as the principal investigator had observed on several occasions previously when administering the SET in a clinical context.

Future Directions

The data gathered in this study can be analysed in more ways than presented in this dissertation. The purpose of this dissertation was to establish specific psychometric
parameters of the 12ET, and therefore not all possible analyses have been included. It will be interesting to further investigate the manner in which the WTAR and executive function tests are related. A factor analysis of these data would show what variance of the executive function tests is accounted for by crystallized intelligence, and what variance (if any) is unique to the executive function tests, though the sample may be too small for this analysis.

Another interesting line of research involves the DEX. The Informant version of the DEX has been factor analysed and has three distinct factors (Burgess et al., 1998). The factors are Inhibition, Intentionality, and Executive Memory. Burgess et al. (1998) found that only the Intentionality factor was related to performance on the SET in a UK sample of brain-injured participants. Chan and Manly (2002) found a similar result in a healthy Chinese sample. Intentionality includes the items pertaining to planning problems, poor decision making, lack of insight, distractibility and the knowing-doing dissociation. It will be interesting to investigate whether the DEX data gathered from the informants in the current study fall into the same three factors, and if so, which factors correlate with the 12ET and the other executive function measures.

The above discussion of the relationship between executive function tests and fluid intelligence suggests the 12ET would correlate with fluid intelligence. It would be interesting to include a fluid reasoning measure, such as Matrix Reasoning from the WAIS III, in future studies employing the 12ET. It would also be informative to evaluate personality more fully and correlate further personality variables with performance on the 12ET.
Another direction future studies on the 12ET could take would be to include an auditory alert that sounds periodically throughout the test administration time, similar to that employed in the Hotel Task (Manly et al., 2002). It would be very interesting to evaluate whether a periodic alert improved participant performance on the 12ET, and whether or not healthy controls and brain-injured participants obtain the same benefit from the alert.

In future research, the 12ET could be improved in several ways based on the lessons learned here. When considering how to improve participant performance on the 12ET, an obvious place to look is the administration instructions. The instructions for the 12ET were read aloud to the participants, and they were asked to follow along on the instruction sheet that was on the table in front of them. At the conclusion of the instructions, participants were asked to tell the examiner what they must do. If participants were not able to clearly state the intention of the task (including the rule) in their own words, they were cued but were not asked to state the instructions again. It may have been beneficial to have the participants learn the 12ET instructions under conditions of errorless learning. In this way, as each portion of the instructions was learned (i.e. the list of subtask requirements, the overall goal, and the rule) the participants would have been able to recite the instructions from memory. Errorless learning techniques, along with the provision of the task instruction sheet may increase examiner confidence that the reason for task failure is not misunderstanding of the instructions.

Further refinements to the 12ET would include reducing the amount of money in the money sorting subtasks, to make them shorter tasks, and removing the card sorting task from the test entirely so as not to have two tasks that are so similar. Instead of the
card sorting task, a document editing task such as employed in the Hotel Task (Manly et al., 2002) would be useful, as it is effortful and does not have an obvious end as the card sorting task did. In addition, it does not require a great deal of writing, which may be a deterrent to some participants. On that note, the restaurant task could be modified to reduce the writing load. Rather than writing down meal combinations the participants could simply read out loud from the menu (and their responses would be recorded on the tape recorder). Another modification taken from the Hotel Task would be to keep a covered clock on table. The examiner kept a stopwatch on her clipboard, and the participants were instructed they could ask how much time was remaining whenever they wanted. However, because they had to ask the examiner, they may have been reluctant to ask (perhaps they thought they would be scored lower if they asked frequently, or they were embarrassed). If the clock were available to them to check when they wanted, one wonders if they would check more frequently, and potentially improve their scores on the Number of Tasks Attempted as Time Requests appeared to cue task switches.

Finally, the choice of relevant variables of the 12ET may need to be changed. While the Number of Tasks Attempted seems to make the most intuitive sense as the variable that is most relevant, perhaps the gross nature of this variable, and limited range (even on the 12ET) limits its usefulness. The Time Deviation variable seems as though it may be a more useful criterion variable. Not only did it correlate with more executive tasks than the Number of Tasks Attempted, but it seem to be a more promising variable in terms of evaluating reliable change. This is in keeping with advanced normative data on the SET (P.W. Burgess, 15/08/07, personal communication) where the longest time spent on any one task is considered a criterion variable along with Number of Tasks
Attempted and Rule Breaks. In Burgess’s unpublished data, the longest time spent on any one task was highly correlated with the number of tasks attempted, similar to the strong relationship between Time Deviation and Number of Tasks Attempted in these data.

Test development is a remarkably difficult process. In particular, obtaining an adequate sample that is large enough, and which demonstrates the required characteristics, is a challenging endeavour that takes a great deal of time and resources. Having made an attempt to improve on a published test, this author has newfound respect for those who persevere in this process and eventually publish tests for clinical and research purposes. While it is important to use measurement tools that have strong psychometric properties, the development of such tools is time consuming and expensive, particularly when complex cognitive abilities are the object of the assessment. While more psychometrically sound and ecologically valid measures will continue to be developed, current measurement tools can be useful, as long as one remains cognizant of the strengths and limitations of those tools within various populations and for different purposes.
REFERENCES


analysis of the Mayo-Portland Adaptability Inventory. *Journal of Head Trauma Rehabilitation*, 15(1), 670-682.


APPENDIX A

INSTRUCTIONS FOR THE TWELVE ELEMENTS TASK

Examiner copy

You get 15 minutes for this task, and in this task you will be doing 6 different subtasks. The first involves describing events. The second involves solving some simple arithmetic problems on these cards here, and writing the answers on the paper I’ve given you. The third involves writing down the names of some pictures shown on these cards on the paper I’ve given you. The fourth subtask involves sorting playing cards into sequential order. The fifth involves putting play money onto order from smallest bills to largest. The sixth subtask involves choosing menu items to make three course meals.

Refer to their sheet

Each of these tasks is divided into two parts. Part A and Part B. There are two sets of dictation exercises, two sets of arithmetic problems, two sets of pictures, two sets of playing cards, two piles of play money, and two restaurant menus.

During the next 15 minutes, I would like you to try to complete at least some of the 12 individual parts. There is NO WAY that you will be able to complete everything in just 15 minutes. So, the most important thing is not to try to complete any one task, but to make sure you try at least some of all 12 parts.

However, there is one rule that you must follow: You cannot move on to the second part of a task immediately after you have attempted the first part of the same task, and of course, you can’t do the first part of the task immediately after the second part of the same task. Instead, you must choose one of the other tasks to do. For e.g., if you do the first part of the arithmetic, you cannot immediately then go on to try the second part: you must instead try one of the other subtasks.

In summary, you have 15 minutes for this entire task. You must organise your work so that you do at least something from each of the 12 parts, and you must not attempt one part of a particular subtask immediately after you have tried the other part. I will set this stopwatch to count down from 15 minutes, and you can ask me how much time you have left whenever you want. Do you have any questions?

Now tell me what you must do.
Appendix A Instructions for The Twelve Elements Task – Participant copy

The subtasks:

1. Describing a memorable personal vacation (Part A) and a memorable personal event (Part B) into the tape recorder
2. Solving as many math problems as you can. There are two sets of simple math problems (Part A and Part B). Record your answers on the sheet provided.
3. Recording the names of as many pictured objects as you can on the sheet provided (Part A and Part B).
4. Sorting as many cards as you can into sequential order (Aces, Twos, Threes, up to Kings). There are two piles of cards (Part A and Part B).
5. Putting as much play money as you can from low bills to high bills (Part A and Part B).
6. Reading two different menus (Part A and Part B) and choosing as many different meal combinations (each) consisting of an appetizer, entree and dessert item as you can.

Summary instructions:

1. Attempt as many of the 12 tasks as you can
2. Do as much as you can within 15 minutes
3. You can do the tasks in any order, except that you cannot do the second part of any of the 6 tasks immediately after doing the first part. For example, if you do Part A of the math task, you cannot do Part B immediately afterwards.
APPENDIX B CONSENT FORMS

Psychometric evaluation of the Twelve Elements Test, and other commonly used measures of executive function.

You are being invited to participate in a study called “Development and evaluation of the Twelve Elements Test, and evaluation of commonly used measures of executive function”. The researcher is Claire Sira who is a graduate student in the department of Psychology at the University of Victoria. Her number is 721-7538. Dr. Adele Hern is Claire Sira's supervisor at the Gorge Road Hospital and Victoria General Hospital and Dr. Catherine Mateer is the supervisor at the University of Victoria (721-7012).

The purpose of this study is to investigate a new test of thinking abilities. Research of this type is important because we need to know if a neuropsychological test is useful in the assessment of thinking skills following brain injury. To better understand the test, we need to know how people without a brain injury perform on it. This knowledge allows clinicians to use tests with confidence in working with their clients.

People who have trouble with planning and problem solving sometimes have more trouble managing at home after they are discharged from hospital. This research will help develop a neuropsychological test that assesses difficulties in planning and organization.

You have been selected for this study because you do not have a brain injury or neurological condition and have not completed a neuropsychological assessment in the past 12 months. You have also been selected because you know one of the brain-injured participants well (and see him/her every day, or almost every day), and are the same age and gender as the brain-injured participant.

If you decide to participate in this research, this is what your participation will include:

- Completing a brief interview and assessment of your thinking on 2 occasions, 8 weeks apart. During the 8 weeks, you don’t need to do anything different; all you have to do it wait for me to contact you again.

The neuropsychological assessment and interview takes about 2 hours and it includes:
- A new test that I am developing. It has 6 different tasks, each with two parts, and you will be asked to complete some of each task within 15 minutes.
- A few other neuropsychological tests that take about 1 hour. Some of these tasks are easy, and some are more difficult. You will not be expected to have perfect performance. I just want you to try your best.
- An interview and questionnaire that ask you questions about your brain-injured family member or friend.

Your participation in this research must be completely voluntary, and you can withdraw at any time. Claire Sira does not have any impact on any future care you may have in
hospital. If you choose to decline to participate, or withdraw from the study, it will have no effect on any future treatment you may have at the hospital. If you choose to withdraw during the study, there will be no penalty, and you will not have to return the money you have been paid. If you choose to continue, but are still feeling distressed, then we can discuss your concerns after the session, and can make an appointment with your doctor for you to discuss your concerns with him/her if you want. If you withdraw from the study, your information will not be kept without your consent. I will ask you if you agree to my using your information if you withdraw from the study. If you say no, the information will be destroyed.

**RISKS AND BENEFITS**

This study may cause you to feel tired. We will take breaks if you get tired. During the assessment, you may find that you feel upset if you don’t do as well as you had hoped on the thinking tasks. The tasks can be difficult and you are not expected to achieve perfect performance. You can stop the assessment at any time if you feel too uncomfortable.

Participation in this study may also be beneficial to you. The assessment measures used in this study (other than the Twelve Elements Test) are commonly used when assessing thinking after brain injury. Those participants who are interested in getting feedback on their results will be provided with feedback after the second assessment.

**COMPENSATION**

To compensate you for any costs and/or inconvenience related to your participation, you will be paid for your participation:
$15 for each of the 2 assessment sessions for a total of $30. If you don’t think you would choose to participate if the money was not offered, then you should decline.

**CONFIDENTIALITY**

Your identity will be protected by coding your information. All information gathered through the study will be kept in a locked filing cabinet at the University of Victoria. None of the data will identify individual participants, and only Claire Sira (me) and my supervisors will have access to your test and rehabilitation results.

The results will be included as part of a university dissertation and may be written into a report that will be submitted for publication in a scientific journal. The results may also be presented as a poster or paper at conferences or workshops.

The study is expected to be completed by April 2005. At the end of the study, I will discuss your assessment results with you. A letter outlining the general findings of the study will be sent to those participants who request it. Information gathered from this study will be shredded five years after the completion of the study (April 2010).
In order to assure myself that you are continuing to give your consent to participate in this research, I will ask you over the next few months if you are still willing to participate.

In addition to being able to contact the researcher and the supervisors at the above phone numbers, you can check the ethical approval of this study, or raise any concerns you might have, by contacting the Associate Vice-President, Research at the University of Victoria (250-472-4362). If you would like to discuss your rights as a research participant with the South Island Medical Director and Medical Administrative Consultant for the Vancouver Island Health Authority, Dr. Ernie Higgs, his contact number is 250-727-4110.

Your signature below indicates that you understand the above conditions of participation in this study and that you have had the opportunity to have your questions answered.

Name of Participant ____________________ Signature ____________________ Date ________________

A copy of this consent will be left with you, and a copy will be taken by the researcher
Psychometric evaluation of the Twelve Elements Test, and other commonly used measures of executive function.

You are being invited to participate in a study called “Development and evaluation of the Twelve Elements Test, and evaluation of commonly used measures of executive function”. The researcher is Claire Sira who is a graduate student in the department of Psychology at the University of Victoria. Her number is 721-7538. Dr. Adele Hern is Claire Sira's supervisor at the Gorge Road Hospital and Victoria General Hospital and Dr. Catherine Mateer is the supervisor at the University of Victoria (721-7012).

The purpose of this study is to investigate a new test of thinking abilities. Research of this type is important because we need to know if a neuropsychological test is useful in the assessment of thinking skills following brain injury. We need to understand how people with a brain injury perform on this test. This knowledge allows clinicians to use tests with confidence in working with their clients.

People who have trouble with planning and problem solving sometimes have more trouble managing at home after they are discharged from hospital. This research will help develop a neuropsychological test that assesses difficulties in planning and organization.

You have been selected for this study because you have had a brain injury, and you are not planning on taking part in any rehabilitation of your thinking in the next 2 months.

If you decide to participate in this research, this is what your participation will include:

 Completing a brief interview and assessment of your thinking (about 2 hours in length) on 2 occasions, 8 weeks apart. A person who knows you well will also be asked to complete an interview and assessment on these 2 occasions. During the 8 weeks, you don’t need to do anything different; all you have to do it wait for me to call you again.

 For the duration of this study, it would be best if you did not take part in any formal rehabilitation program (for your thinking) at the same time. If you decide to do a formal rehabilitation program, you need to tell me, so I can note that you did so.

 The neuropsychological assessment and interview takes about 2 hours and it includes:
 A short interview about how some of your thinking skills affect your daily life.
 A new test that I am developing. It has 6 different tasks, each with two parts, and you will be asked to complete some of each part within 15 minutes.
 A few other neuropsychological tests that take about 1 hour. Some of these tasks are easy, and some are more difficult. You will not be expected to have perfect performance. I just want you to try your best.
 A questionnaire for you to fill out about everyday problems people who have had a brain injury may experience. The questionnaire also has a form to be completed by someone
who knows you well. This person is someone such as a carer, family member or friend that you see every day, or almost every day.

Your participation in this research must be completely voluntary, and you can withdraw at any time. The assessments will take place at the University of Victoria, and no one from the hospital will know that you took part. Claire Sira does not have any impact on any future care you may have at the hospital. However, you may have an ongoing professional relationship with Dr. Hern. If you choose to decline to participate, or withdraw from the study, it will have no effect on any future treatment you may have at the Neuro-rehabilitation Unit, and Dr. Hern will not refer to this study when she meets with you. If you do not participate, or choose to withdraw during the study, there will be no penalty, and you will not have to return the money you have been paid. If you choose to continue, but are still feeling distressed, then we can discuss your concerns after the session, and can make an appointment with your doctor for you to discuss your concerns with him/her if you want. If you withdraw from the study, your information will not be kept without your consent. I will ask you if you agree to my using your information if you withdraw from the study. If you say no, the information will be destroyed.

**RISKS AND BENEFITS**

This study may cause you to feel tired. We will take breaks if you get tired. During the assessment, you may find that you feel upset if you don’t do as well as you had hoped on the thinking tasks. The tasks can be difficult and you are not expected to achieve perfect performance. You can stop the assessment at any time if you feel too uncomfortable.

Participation in this study may also be beneficial to you. The assessment measures used in this study (other than the Twelve Elements Test) are commonly used when assessing thinking after brain injury. You will be provided with feedback after the second assessment, including recommendations based on your specific strengths and weakness.

**COMPENSATION**

To compensate you for any costs and/or inconvenience related to your participation, you will be paid for your participation:
$15 for each of the 2 assessment sessions for a total of $30. If you don’t think you would choose to participate if the money was not offered, then you should decline.

**CONFIDENTIALITY**

Your identity will be protected by coding your information. All information gathered through the study will be kept in a locked filing cabinet at the University of Victoria. None of the data will identify individual participants, and only Claire Sira (me) and my supervisors will have access to your test and rehabilitation results.
The results will be included as part of a university dissertation and may be written into a report that will be submitted for publication in a scientific journal. The results may also be presented as a poster or paper at conferences or workshops.

The study is expected to be completed by April 2005. At the end of the study, I will discuss your assessment results with you. A letter outlining the general findings of the study will be sent to those participants who request it. Information gathered from this study will be shredded five years after the completion of the study (April 2010).

In order to assure myself that you are continuing to give your consent to participate in this research, I will ask you over the next few months if you are still willing to participate.

In addition to being able to contact the researcher and the supervisors at the above phone numbers, you can check the ethical approval of this study, or raise any concerns you might have, by contacting the Associate Vice-President, Research at the University of Victoria (250-472-4362). If you would like to discuss your rights as a research participant with the South Island Medical Director and Medical Administrative Consultant for the Vancouver Island Health Authority, Dr. Ernie Higgs, his contact number is 250-727-4110.

Your signature below indicates that you understand the above conditions of participation in this study and that you have had the opportunity to have your questions answered.

_________________________  _______________________  ____________
Name of Participant            Signature               Date

A copy of this consent will be left with you, and the researcher will take a copy.
Psychometric evaluation of the Twelve Elements Test, and other commonly used measures of executive function.

You are being invited to participate in a study called “Development and evaluation of the Twelve Elements Test, and evaluation of commonly used measures of executive function”. The researcher is Claire Sira who is a graduate student in the department of Psychology at the University of Victoria and who is currently working at Foothills Hospital (403-277-8998). Dr. Catherine Mateer is Claire Sira’s supervisor at the University of Victoria (250-721-7012).

The purpose of this study is to investigate a new test of thinking abilities. Research of this type is important because we need to know if a neuropsychological test is useful in the assessment of thinking skills following brain injury. We need to understand how people with a brain injury perform on this test. This knowledge allows clinicians to use tests with confidence in working with their clients.

People who have trouble with planning and problem solving sometimes have more trouble managing at home after they are discharged from hospital. This research will help develop a neuropsychological test that assesses difficulties in planning and organization.

You have been selected for this study because you have had a brain injury, and you are not planning on taking part in any formal rehabilitation of your thinking in the next 2 months.

If you decide to participate in this research, your participation will include:

Completing a brief interview and assessment of your thinking on 2 occasions, 8 weeks apart. A person who knows you well will also be asked to complete an interview and assessment on these 2 occasions. During the 8 weeks, you don’t need to do anything different; all you have to do is wait for me to call you again.

The neuropsychological assessment and interview takes about 2 hours and it includes: A short interview about how some of your thinking skills affect your daily life. A new test that I am developing. It has 6 different tasks, each with two parts, and you will be asked to complete some of each part within 15 minutes. A few other neuropsychological tests that take about 1 hour. Some of these tasks are easy, and some are more difficult. You will not be expected to have perfect performance. I just want you to try your best.

A questionnaire for you to fill out about everyday problems people who have had a brain injury may experience. The questionnaire also has a form to be completed by someone who knows you well. This person will be a family member or friend that you see often.

COMPENSATION

To compensate you for any costs and/or inconvenience related to your participation, you will be paid for your participation:
$20 for each of the 2 assessment sessions for a total of $40.
If you would not choose to participate if the money were not offered, then you should decline.

Your participation in this research must be completely voluntary, and you can withdraw at any time. The assessments will take place at Columbia Health, but no one other than Claire Sira will know that you took part. Claire Sira does not have any impact on any future care you may have at Columbia Health, or involvement you may have at the University of Victoria.
If you do not participate, or choose to withdraw during the study, there will be no penalty, and you will not have to return the money you have been paid. If you choose to continue, but are still feeling distressed, then we can discuss your concerns after the session, and can make an appointment with your doctor for you to discuss your concerns with him/her if you want. If you withdraw from the study, your information will not be kept without your consent. I will ask you if you agree to my using your information if you withdraw from the study. If you say no, the information will be destroyed.

RISKS AND BENEFITS

This study may cause you to feel tired. We will take breaks if you get tired. During the assessment, you may find that you feel upset if you don’t do as well as you had hoped on the thinking tasks. The tasks can be difficult and you are not expected to achieve perfect performance. You can stop the assessment at any time if you feel too uncomfortable.

For the duration of this study, it would be best if you did not take part in any formal rehabilitation program (for your thinking). If you decide to do a formal rehabilitation program, you need to tell me, so I can note that you did so.

Participation in this study may also be beneficial to you. The assessment measures used in this study (other than the Twelve Elements Test) are commonly used when assessing thinking after brain injury. If you choose, you will be provided with feedback after the second assessment, including recommendations based on your specific strengths and weakness.

CONFIDENTIALITY

Your identity will be protected by coding your information. All information gathered through the study will be kept in a locked filing cabinet and stored at the University of Victoria. None of the data will identify individual participants, and only Claire Sira and her university supervisor will have access to your test results.

The results will be included as part of a university dissertation and may be written into a report that will be submitted for publication in a scientific journal. The results may also be presented as a poster or paper at conferences or workshops. No identifying information will be included in any of these works.
The study is expected to be completed by July 2006. At the end of your participation, your assessment results will be shared with you. A letter outlining the general findings of the study will be sent to those participants who request it. Information gathered from this study will be shredded five years after the completion of the study (July 2011).

In order to assure myself that you are continuing to give your consent to participate in this research, I will ask you over the next few months if you are still willing to participate.

In addition to being able to contact the researcher and the supervisor at the above phone numbers, you can check the ethical approval of this study, or raise any concerns you might have, by contacting the Associate Vice-President, Research at the University of Victoria (250-472-4362).

Your signature below indicates that you understand the above conditions of participation in this study and that you have had the opportunity to have your questions answered.

_________________________  ________________________  ________________
Name of Participant          Signature                  Date

A copy of this consent will be left with you, and the researcher will take a copy.
Psychometric evaluation of the Twelve Elements Test, and other commonly used measures of executive function.

You are being invited to participate in a study called “Development and evaluation of the Twelve Elements Test, and evaluation of commonly used measures of executive function”. The researcher is Claire Sira who is a graduate student in the department of Psychology at the University of Victoria and who is currently completing her internship at Foothills Hospital (403-277-8998). Dr. Catherine Mateer is Claire Sira’s supervisor at the University of Victoria (250-721-7012).

The purpose of this study is to investigate a new test of thinking abilities. Research of this type is important because we need to know if a neuropsychological test is useful in the assessment of thinking skills following brain injury. To better understand the test, we need to know how people without a brain injury perform on it. This knowledge allows clinicians to use tests with confidence in working with their clients.

People who have trouble with planning and problem solving sometimes have more trouble managing at home after they are discharged from hospital. This research will help develop a neuropsychological test that assesses difficulties in planning and organization.

You have been selected for this study because you do not have a brain injury or neurological condition and have not completed a neuropsychological assessment in the past 12 months.

If you decide to participate in this research, this is what your participation will include:

Completing a brief interview and assessment of your thinking on 2 occasions, 8 weeks apart. During the 8 weeks, you don’t need to do anything different; all you have to do is wait for me to contact you again.

The neuropsychological assessment and interview takes about 2 hours and it includes:
A new test that I am developing. It has 6 different tasks, each with two parts, and you will be asked to complete some of each task within 15 minutes.
A few other neuropsychological tests that take about 1 hour. Some of these tasks are easy, and some are more difficult. You will not be expected to have perfect performance. I just want you to try your best.

COMPENSATION

To compensate you for any costs and/or inconvenience related to your participation, you will be paid for your participation:
$20 for each of the 2 assessment sessions for a total of $40.
If you would not choose to participate if the money were not offered, then you should decline.
Your participation in this research must be completely voluntary, and you can withdraw at any time. The assessments will take place at Columbia Health, but no one other than Claire Sira will know that you took part. Claire Sira does not have any impact on any future care you may have at Columbia Health, or involvement you may have at the University of Victoria.

If you choose to withdraw during the study, there will be no penalty, and you will not have to return the money you have been paid. If you choose to continue, but are still feeling distressed, then we can discuss your concerns after the session, and can make an appointment with your doctor for you to discuss your concerns with him/her if you want. If you withdraw from the study, your information will not be kept without your consent. I will ask you if you agree to my using your information if you withdraw from the study. If you say no, the information will be destroyed.

**RISKS AND BENEFITS**

This study may cause you to feel tired. We will take breaks if you get tired. During the assessment, you may find that you feel upset if you don’t do as well as you had hoped on the thinking tasks. The tasks can be difficult and you are not expected to achieve perfect performance. You can stop the assessment at any time if you feel too uncomfortable.

Participation in this study may also be beneficial to you. The assessment measures used in this study (other than the Twelve Elements Test) are commonly used when assessing thinking after brain injury. Those participants who are interested in getting feedback on their results will be provided with feedback after the second assessment.

**CONFIDENTIALITY**

Your identity will be protected by coding your information. All information gathered through the study will be stored in a locked filing cabinet at the University of Victoria. None of the data will identify individual participants, and only Claire Sira and her supervisor will have access to your test results.

The results will be included as part of a university dissertation and may be written into a report that will be submitted for publication in a scientific journal. The results may also be presented as a poster or paper at conferences or workshops.

The study is expected to be completed by Dec 2007. At the end of your participation, your assessment results will be shared with you. A letter outlining the general findings of the study will be sent to those participants who request it. Information gathered from this study will be shredded five years after the completion of the study (Dec 2012).

In order to assure myself that you are continuing to give your consent to participate in this research, I will ask you over the next few months if you are still willing to participate.
In addition to being able to contact the researcher and the supervisor at the above phone numbers, you can check the ethical approval of this study, or raise any concerns you might have, by contacting the Associate Vice-President, Research at the University of Victoria (250-472-4362).

Your signature below indicates that you understand the above conditions of participation in this study and that you have had the opportunity to have your questions answered.

____________________________  ______________________________  __________________
Name of Participant                Signature                           Date

_A copy of this consent will be left with you, and a copy will be taken by the researcher_
APPENDIX C

SEVERITY OF DEPENDENCE SCALE

Please answer each question by circling one response only
These questions are about how you felt about your [named drug] use in the last year.

1. Did you ever think your [named drug] use was out of control?
Never or almost never 0
Sometimes 1
Often 2
Always 3

2. Did the prospect of missing a [fix or shot or drink] make you very anxious or worried?
Never or almost never 0
Sometimes 1
Often 2
Always 3

3. Did you worry about your [named drug] use?
Not at all 0
A little 1
Often 2
Always or nearly always 3

4. Did you wish you could stop?
Never or almost never 0
Sometimes 1
Often 2
Always 3

5. How difficult would you find it to stop or go without?
Not difficult at all 0
Quite difficult 1
Very difficult 2
Impossible 3