2007

# Developing a spatial analogue of the Reliable Digit Span: A within-test measure of performance validity in neuropsychological examination after suspected traumatic brain injury

Shelley Ylioja
*University of Windsor*

# DEVELOPING A SPATIAL ANALOGUE OF THE RELIABLE DIGIT SPAN: A WITHIN-TEST MEASURE OF PERFORMANCE VALIDITY IN NEUROPSYCHOLOGICAL EXAMINATION AFTER SUSPECTED TRAUMATIC BRAIN INJURY

by

Shelley Ylioja

A Thesis
Submitted to the Faculty of Graduate Studies
Through Psychology
in Partial Fulfillment of the Requirements for
the Degree of Master of Arts at the
University of Windsor

Windsor, Ontario, Canada

2007

*Your file Votre référence*
*ISBN: 978-0-494-34980-9*
*Our file Notre référence*
*ISBN: 978-0-494-34980-9*

Abstract

It is recommended that validity of performance be assessed in all neuropsychological cases involving external incentive. The present study sought to develop a within-test performance validity measure based on the spatial span task. The Reliable Spatial Span (RSS) calculation had specificity, sensitivity, and predictive power values within the range of other within-test measures, which suggests RSS is able to distinguish between mild TBI cases demonstrating valid and invalid performance. Reliable Digit Span (RDS) classification accuracy within the present sample was lower than that of previous research, as well as of RSS. The possibility that spatial span may be a better indicator of invalid performance than digit span is discussed. Finally, involvement of working memory system components in the spatial and digit span tasks was explored, with some support for the span tasks being less closely analogous than typically assumed.

Table of Contents

## Introduction

It has traditionally been assumed that poor performance on neuropsychological tests is indicative of cognitive impairment, with poorer scores being observed in cases of more severe impairment. However, neuropsychologists have realized that this is not always the case, and that incentive to perform well or to perform poorly has a large influence on a test profile. Thus, a lot of attention has been focused recently on the development of measures to include in the neuropsychological test battery that will allow clinicians to make informed conclusions with regards to the validity of client performance. The main purpose of the present study was to develop a performance validity marker from a commonly-used neuropsychological test, without compromising the role of the measure in assessing an aspect of cognitive function.

### *Effort in Neuropsychological Testing*

Several studies highlight the influence of client effort on test findings. Green (2006) has shown that results on a wide range of neuropsychological tests are more related to effort than to the severity of brain injury/disease in a variety of neurological disorders, with a progressive decrease in test scores corresponding to decreasing level of effort. Another study found that effort explained 53% of the test score variance, while variables such as education and age only explained 11% and 4%, respectively (Green, Rohling, Lees-Haley, & Allen, 2001). This study also showed that sub-optimal effort suppressed performance on neuropsychological tests over four times more than did moderate/severe brain injury, with a steady decrease in test scores occurring with decreasing effort scores. In fact, no difference in test performance was observed between mild injury severity and moderate/severe injury severity groups until client effort was considered. Once those clients who had displayed inadequate effort had been removed from the analysis, the expected negative relationship between head injury severity and performance on cognitive tests was evident. Others have found similar results, with results from clients who demonstrate adequate effort also showing this expected negative relationship between injury severity (classified into six groups) and test performance, while no relationship was evident in the data from clients demonstrating inadequate effort (Moss, Jones, Fokias, & Quinn, 2003). In addition, in a sample comprised entirely of mild traumatic brain injury (TBI) it was demonstrated that 47% of the variance of the General Neuropsychological Deficit Score on the Halstead-Reitan battery was explained by effort scores (Constantinou, Bauer, Ashendorf, Fisher, & McCaffrey, 2005). Thus, it is obvious that the validity of neuropsychological testing depends to a large extent on the level of effort put forth by the client to perform to the best of his or her ability.

Failure to exert optimal effort compromises test results and gives an inaccurate picture of true cognitive abilities. Important decisions and recommendations with respect to presence of brain injury, a client's daily life, rehabilitation, return to work, and educational pursuits are made based on test results. In addition, test data inform theories about brain-behaviour and neurocognitive relationships, as well as expected performance patterns in particular brain injuries/disorders.

A recent study underscored the importance of this by demonstrating how effort was a confounding factor that had resulted in a false belief held by neuropsychologists regarding the level of cognitive impairment in a particular client group. Based on neuropsychological test performance patterns it was previously believed that patients with psychogenic nonepileptic seizures (PNES) had equal or greater cognitive impairment than patients with epilepsy, even though magnetic resonance imaging and electroencephalography failed to show evidence of brain disease in the former group. A study comparing these groups found that roughly 50% of the PNES group failed a widely-used effort test, compared to only 8% of the epilepsy group (Drane et al., 2006). Once those displaying poor effort had been removed from the analysis, the PNES group showed significantly less impairment than the epilepsy group, a pattern that would be expected given the objective evidence of brain injury in the epilepsy group compared to the PNES group. This example illustrates how a failure to take degree of effort into consideration can lead to inaccurate conclusions and expectancies with respect to a particular disorder.

It is, therefore, important to include measures of effort in a neuropsychological test battery. This has recently been made clear in a National Academy of Neuropsychology (NAN) position paper on validity assessment, which states that such an assessment "…is an essential part of a neuropsychological evaluation. The clinician should be prepared to justify a decision not to assess symptom validity…" (Bush et al., 2005, p. 421).

*Impact of Financial Compensation*

The inclusion of validity measures is important in all neuropsychological test settings, but it becomes especially crucial in cases involving possible financial compensation. Examples of such cases would include clients involved in an automobile insurance claim and/or civil litigation after a motor vehicle accident (MVA), a worker's compensation claim, or a medical disability insurance claim. In such a situation there is a greater possibility of malingering, or intentionally performing poorly for monetary gain.

It has been estimated that anywhere from 20 to 60% of clients in litigation or other types of compensation cases involving mild head trauma are malingering (Binder, 1993; Constantinou et al., 2005; Greiffenstein, Baker, & Gola, 1994; Langeluddecke & Lucas, 2003; Meyers & Volbrecht, 1998). This factor must be considered seriously, given that a number of studies have found compensation-seeking status to be a significant variable in test performance.

For example, Binder and Willis (1991) found poor performance on the Portland Digit Recognition Test (PDRT), a measure designed to assess motivation, to be more related to financial incentive than to severity of injury. Clients with well-documented brain dysfunction (moderate to severe traumatic brain injury, cerebrovascular disease, central nervous system infection) who were not seeking financial compensation scored significantly better on the PDRT than clients who were seeking monetary gain through worker's compensation, a personal injury claim, or a lawsuit who had an injury of similar severity or a mild head injury. In fact, 26% of the mild head injury group performed below the lowest score attained on the PDRT by anyone in the group with documented brain injury not seeking compensation.

A study using the Word Memory Test (WMT), designed to measure biased responding, found that those with mild TBI performed worse on the WMT than those with moderate to severe TBI (Green, Iverson, & Allen, 1999). All patients included in the analysis were involved in worker's compensation, medical disability insurance, or a lawsuit. However, the group of clients with a moderate or severe TBI had their insurance/disability claims already guaranteed while those in the mild TBI group did not (as is often the case), thus giving the latter group more incentive to feign impairment.

Also, financial incentive has been associated with nearly a 4-fold increase in invalid performance, as measured by the Test of Memory Malingering (TOMM) and forced-choice recognition of the California Verbal Learning Test –II (CVLT-II). Injury severity (mild versus moderate to severe TBI), on the other hand, was not found to be a significant predictor of performance validity (Moore & Donders, 2004).

In another study, a comparison of two groups of clients with equal injury severity revealed that those with a financial incentive demonstrated greater impairment on a within-test measure of effort (Reliable Digit Span; RDS) as well as on measures of cognitive functioning from the Wechsler Adult Intelligence Scale-Revised (Meyers & Volbrecht, 1998). While 48.9% of the group with financial incentive failed RDS, only 4.1% of the group without financial incentive did so. All clients had suffered a mild TBI, defined as having experienced loss of consciousness of less than one hour following a MVA or blow to the head.

While the studies just outlined give support for financial compensation as a significant factor in test performance, not all studies support this. A recent multiple regression analysis found presence of financial incentive to be predictive of test scores in only three of thirteen cognitive domains, with the rate of invalid performance no higher in the group having financial incentive (74.8% of sample) compared to the group without such incentive (Ross, Putman, & Adams, 2006). Performance validity, however, was measured with the Recognition Memory Test which was primarily developed as a test of memory (Warrington, 1984) and secondly as a performance validity indicator (Millis, 1992). Thus, because this test has weaker classification power than tests designed specifically to detect invalid performance, the results should not be weighted as heavily in assessing effort as the studies mentioned above that used the WMT, TOMM, or PDRT. However, the majority of studies demonstrate that the presence of monetary incentive alone is related to test results, which speaks to the necessity of including measures of performance validity in such cases.

*Mild Traumatic Brain Injury and Neuropsychological Assessment*

The present study involves a sample evaluated for suspected mild TBI within the context of litigation. Performance validity may be an especially important issue in cases of mild TBI, where there is often no objective medical evidence that a brain injury has been sustained (Echemendia, Putukian, Mackin, Julian, & Shoss, 2001). While the majority of mild TBI cases resolve naturally within the first days to weeks following injury, there are a minority of cases who do not follow the normal recovery pattern and have permanent cognitive impairments (Alexander, 1995; Reitan & Wolfson, 2000; Vanderploeg, Curtiss, & Belanger, 2005). Domains of impairment following a mild TBI can vary widely across individuals, with the most frequent involving processing speed, working memory, attention, and executive function (Alexander, 1995; Echemendia et al., 2001; Immediate Post-concussion Assessment and Cognitive Testing (ImPACT), 2006; McHugh et al., 2006; Psychological Corporation, 1997; Reitan & Wolfson, 2000). Considering the large influence that effort has on test scores, it is necessary to control for it in such cases where cognitive impairments may be subtle in nature. If there are individuals claiming to have sustained a mild TBI and who fake or exaggerate impairment during testing, they will appear much more impaired than those who have real deficits as a result of mild TBI and are exerting their best effort during testing. If performance validity is not controlled such a situation could cause inaccurate decisions regarding degree of impairment, resulting in the former being entitled to assistance ranging from monetary to cognitive remediation to vocational or educational resources, while the latter may not.

*Terminology: Effort, Malingering, and Validity of Performance*

The topics of effort and malingering are typically conflated in research literature. Malingering has been defined as involving a conscious attempt to feign or exaggerate impairments as a result of motivation by external incentives (Slick, Sherman, & Iverson, 1999). However, it should be clarified that insufficient effort can be present without malingering. A client may intend to perform to the best of her or his ability, but effort may be compromised by a variety of factors such as poor attention, fatigue, psychological disturbance, irritation with the testing process, or failure to become fully engaged in the task (Frederick, Crosby, & Wynkoop, 2000). Or there may be internal incentives involved, rather than or in addition to external ones, such as managing psychological stress, playing a sick role, or escaping from informal duty (Slick et al., 1999). In analyzing an invalid profile of results, it is important to make the distinction between a person who had the intention of doing poorly and one who had the intention of performing well but whose effort was compromised for some other reason. The distinction between external and/or internal incentives also is important.

While much of the research surrounding the topic broadly refers to it as malingering, such terminology implies that researchers have been able to isolate the group of individuals who are intentionally performing poorly and who are motivated by external incentive. To avoid such an implication the present paper will refer to the measurement of a “valid” or “invalid” performance profile, whether the invalid profile be a result of intentional poor performance or not. Likewise, the term "effort" will not be used because a client consciously performing poorly is likely extending effort to ensure his profile looks this way (Frederick et al., 2000). It is acknowledged that the results of the present study will not allow the clinician to distinguish among sub-types of invalid responses.

*Tests of Valid Performance*

Previously, assessment of valid performance was appraised solely through clinical judgement. Research suggests, however, that reliance on such judgement alone is far from accurate (Faust & Ackley, 1998). Attention has been recently directed toward the development of tests to measure performance validity in a more objective manner.

Strategies for detection of invalid performance were identified by Rogers, Harrell, and Liff (1993) and include floor effects, symptom validity testing, atypical presentation, performance curves, magnitude of error, and psychological sequelae. The floor effect phenomenon is the failure to perform at a level expected of even clients with severe impairment, or the failure to answer very basic questions or perform simple tasks. Symptom validity tests use a two-alternative forced choice format in which validity is identified as suspicious if scores are significantly below chance level or below some pre-determined cut-off score based on minimum performance of people with moderate or severe brain injury. The atypical performance method refers to examination of consistency of performance within or across test administrations. Performance curve analysis examines performance across item difficulty to determine whether the expected inverse relationship between accuracy and difficulty level is present. The magnitude of error approach involves quantitative and qualitative analysis of error. An example is analysis of "near miss" responses in which one looks at how far off from the correct answer the response was. Lastly, the method of psychological sequelae is based on the observation that clients who are malingering often also present with feigned psychological/psychiatric symptoms which are odd or inaccurate.

The developers of performance validity measures have taken two routes. The first involves designing tests specifically to assess such. These tests are usually designed so that they have a low difficulty level but appear to have a higher one. For example, a large number of stimuli have been used to give a test an appearance of difficulty, when in actuality clients with severe head injuries of various aetiologies are capable of near-perfect performance on the measure. The purpose of these tests is usually to trick malingerers into performing poorly, due to their assumption that someone experiencing cognitive difficulties would be unable to perform well (Inman & Berry, 2002). These tests generally utilize floor effects or symptom validity measures. Common examples include the Test of Memory Malingering (Tombaugh, 1996), Rey 15-Item Memory test (Rey, 1958, cited in Frederick et al., 2000), Computerized Assessment of Response Bias (Allen, Conder, Green, & Cox, 1997), Victoria Symptom Validity Test (Slick, Hopp, & Strauss, 1997), Portland Digit Recognition Test (Binder & Willis, 1991; Binder, 1993), and the Word Memory Test (Green, Allen, & Astner, 1996). The Validity Indicator Profile is one that uses performance curve analysis (Frederick, 1997, cited in Frederick, 2002).

The second route involves the development of cut-off scores for tests commonly used within the neuropsychological battery. Cut-off points are generally determined either by minimum performance of severe head injury groups or by the point that best classifies individuals suspected of invalid performance from those who are not. Suspected invalid performance groups are usually based on one or more of the following: a failing score on a widely accepted and validated performance validity test(s), discrepancy between test performance and/or self-reported symptoms and known brain functioning patterns, contradiction between test performance and observed behaviour, and divergence of self-report and collateral report (Slick et al., 1999). Examples of this type of validity measure include the California Verbal Learning Test forced-choice recognition (Millis, Putnam, Adams, & Ricker, 1995), Reliable Digit Span (Greiffenstein et al., 1994; Greiffenstein, Gola, & Baker, 1995), Recognition Memory Test cut-off score (Millis, 1992), Rey Auditory Verbal Learning Test-Recognition (Meyers, Morrison, & Miller, 2001), Memory Assessment Scale cut-off scores (O'Bryant, Duff, Fisher, & McCaffrey, 2004), as well as unique responses on the Wisconsin Card Sorting Test (Greve, Bianchini, Mathias, Houston, & Crouch, 2002).

There are several possible benefits of using performance validity measures that are built into commonly used neuropsychological tests. Firstly, it is much more efficient if a test can play a dual role in assessing cognitive function as well as performance validity. That is, with built-in validity checks less time in an already-full day of testing needs to be allotted to specific validity tests which are often lengthy to administer (Meyers & Volbrecht, 2003; Langeluddecke & Lucas, 2003). Also, measures within the battery itself work as validity checks of the integrity of the results throughout the entire assessment (Meyers & Volbrecht, 1998). In other words, validity is able to be measured throughout the session, while a single validity test administered at a certain point during the session does not necessarily indicate that performance was such throughout the entire battery. In addition, it has been noted that a client will generally attempt to feign symptoms in a specific area of cognitive functioning rather than feign a global impairment (Greiffenstein et al., 1995). Thus, clients may respond in an invalid way in one domain but not in another, and a single test designed specifically to detect malingering may not be viewed by the client as involving the cognitive area in which he or she is feigning impairment (Iverson & Binder, 2000). In such an instance, malingered performance would be completely missed by the clinician.

Lastly, built-in validity measures may also be less susceptible to intentional poor effort made by clients who have been made aware by their lawyers of tests designed to detect malingering (Mathias, Greve, Bianchini, Houston, & Crouch, 2002) or who have found such information on the internet (Bauer & McCaffrey, 2006). This is an area of concern for neuropsychologists given the evidence that some attorneys do coach their clients regarding tests used to detect malingering and test-taking strategies to avoid detection (Essig, Mittenberg, Petersen, Strauman, & Cooper, 2001; Wetter & Corrigan, 1995; Youngjohn, 1995). Information is also available on the internet about some of the common performance validity tests used by neuropsychologists, explaining the test format as well as providing the suggested cut-off scores (Bauer & McCaffrey, 2006). If only one or two validity tests are used by clinicians, clients may recognize these tests when they encounter them in the test situation and may be able to escape detection by altering responding during these measures. It is thought that they especially may recognize the forced-choice test (Suhr & Gunstad, 2000), and it has been noted that methods used by neuropsychologists are often very transparent (Faust & Ackley, 1998). In fact, Rogers and colleagues (1993) explained that people attempting to feign impairment are better able to escape detection if they have received information on how validity of performance is calculated on standardized tests as opposed to information about the disorder they are feigning. This finding has been replicated by Gorny and Merten (2005). This, along with the evidence that attorney coaching does occur and that clients may easily recognize commonly-used validity measures, speaks for the importance of incorporating validity scales into the neuropsychological tests already in use.

While objective validity measures are necessary, they should not be the foundation for determination of whether feigned/exaggerated symptoms or poor effort are present. There is, of course, still room for clinical judgment. "In the final analysis, behavioural (subjective) and empirical (objective) evidence must be integrated by a seasoned clinician who evaluates this evidence together with premorbid, comorbid, and postmorbid history and then renders his or her best judgment" (Ruff, Wylie, & Tennant, 1993, p. 70).

*Spatial Span Analogue of the Reliable Digit Span*

Reliable Digit Span (RDS) is a performance validity measure that has been developed using a test commonly administered in a neuropsychological battery. It involves a calculation from the Wechsler Adult Intelligence Scale-III (WAIS-III) or Wechsler Memory Scale (WMS-III) digit span of the sum of the highest number of digits forward on which both trials are correct plus the highest number of digits backwards on which both trials are correct (Greiffenstein et al., 1994). RDS has been suggested to be one of the best-validated within-test clinical measures of "effort" (Heinly, Greve, Bianchini, Love, & Brennan, 2005), and has been mainly tested in clinical cases of suspected and definite traumatic brain injury (Greiffenstein et al., 1994, 1995; Heinly et al., 2005; Mathias et al., 2002). Other RDS validation samples have included simulated TBI (Inman & Berry, 2002; Strauss et al., 2002), a general clinical sample of medical and/or psychiatric disorders (Babikian, Boone, Lu, & Arnold, 2006), patients reporting chronic pain (Etherton, Bianchini, Greve, & Heinly, 2005) as well as a forensic sample without diagnosed neurologic impairment undergoing pretrial/pre-sentence assessment (Duncan & Ausborn, 2002). A cut-off score of seven or lower has been shown to differentiate suspected valid and invalid responders with acceptable sensitivity, specificity, and predictive power in all of these studies. However, with respect to clinical samples, classification accuracy is lower in cases of documented moderate to severe TBI compared to cases of mild TBI, with a higher rate of misclassifying valid responders as invalid based on RDS scores (Greiffenstein et al., 1994, 1995). This suggests RDS is better suited to differentiating valid and invalid performance in cases of mild TBI.

The spatial span test, often administered in its original Corsi Block Test form, is used frequently in clinical neuropsychological testing (Vandierendonck, Kemps, Fastame, & Szmalec, 2004; Mammarella & Cornoldi, 2005). Spatial span has traditionally been viewed as a non-verbal or visuospatial analogue of digit span (Mammarella & Cornoldi, 2005), a simple test of verbal or auditory attention and working memory. Research looking at the underlying cognitive processes of these two span tests has documented that the spatial and digit span tasks seem to utilize partially overlapping but separable neuronal networks (the visuospatial sketchpad and phonological loop, respectively, of Baddeley's model of working memory) in a similar fashion and rely upon executive functioning resources with spans of longer length

(Vandierendonck, De Vooght, & Van der Goten, 1998; Vandierendonck et al., 2004). In addition, similarities between the span tasks were found by Smyth and Scholey (1996) with respect to serial order and position effects. However, as will be discussed shortly, there is also evidence that the span tasks may not be direct analogues of each other as has been traditionally assumed.

Given the positive findings regarding the utility of digit span's RDS in detecting invalid performance, it was hypothesized that its non-verbal "counterpart" would be sensitive to invalid performance and could also act as a simple within-test measure of performance validity. Specifically, it was expected that a Reliable Spatial Span (RSS) score, calculated in an identical manner to RDS, or some other calculation involving spatial span scores, would accurately differentiate an individual's performance as valid or invalid within a sample of clients examined for suspected mild TBI in the context of litigation. Among the various tests of performance validity already developed, there appears to be a paucity of measures that rely on visuospatial ability. As previously mentioned, clients who malinger generally choose different areas in which to feign difficulties (Greiffenstein et al., 1995; Iverson & Binder, 2000). There is evidence that a variety of cognitive domains may be compromised across individuals who have sustained a TBI, with visuospatial memory being one of them (Chuah, Maybery, & Fox, 2004; Reitan & Wolfson, 2000). It would, therefore, be reasonable to assume that some clients feigning or exaggerating symptoms would do so in this area of functioning. Also, effort is required even by clients who have no visuospatial deficits in order to do well on the spatial span task, making it likely to distinguish those exerting adequate effort from those who are not. In addition, effort is likely to fluctuate to some degree over the course of the testing session, and thus a validity measure based on spatial span would add another within-test check across the administration session.

The present study also sought to replicate past research regarding classification accuracy of RDS. It was hypothesized that RDS in the current study would be associated with sensitivity and specificity values similar to those found previously for a cut-off score of 7 or lower. Other combinations of digit span scores were also explored for utility in classification of performance validity, with no hypotheses extended as to whether use of these alternatives will provide more or less classification accuracy than RDS.

*Underlying Cognitive Constructs Measured by the Span Tests*

The argument that RSS also might be a useful validity indicator rests in part on similarity between the spatial and digit span tasks in terms of underlying constructs. The literature on the underlying abilities involved in the span tasks is mixed. Evidence exists to suggest that the span tasks may not be verbal and visuospatial analogues of each other, as has been previously assumed (Giggey, Spencer, Rice, Katzel, & Waldstein, 2006; Hester, Kinsella, & Ong, 2004; Mammarella & Cornoldi, 2005; Mertens, Gagnon, Coulombe, & Messier, 2006; Smythe & Scholey, 1992; Szmalec, Vandierendonck, & Kemps, 2005; Vandierendonck et al., 2004; Wilde, Strauss, & Tulsky, 2004; Wilde & Strauss, 2002). Moreover, there is conflicting evidence as to whether the forward and backward trials of each involve somewhat different cognitive processes (Mammarella & Cornoldi, 2005; Szmalec et al., 2005; Vandierendonck et al., 2004). Analogousness of the digit and spatial span tasks refers to more than outward similarities of tasks, but also to reliance upon the same combination and weighting of underlying cognitive constructs as well as measurement of the same abilities within the verbal and visuospatial domains, respectively. A second goal of the present study was to try to delineate what cognitive constructs underlie the digit and spatial span tasks overall as well as their forward and backward trials.

While some talk about the span tasks as measures of immediate attention (e.g., Langeluddecke & Lucas, 2003; Larrabee & Curtiss, 1995), most refer to these tasks chiefly as measures of working memory (e.g., Handley, Capon, Copp, & Harper, 2002; Mertens et al., 2006), and both tasks are included under this category in the WAIS-III and WMS-III. The terms attention, short-term memory, and working memory are often used interchangeably in the literature to describe similar processes. This area of cognitive processing is currently under exploration, and thus our understanding is not yet complete. The model most widely used within clinical literature and much of cognitive literature to explain this construct is Baddeley's model of working memory, which was initially proposed in 1976 and continues to be updated. The current study uses this theory in reference to aspects of working memory that may underlie the spatial and digit span tasks.

Baddeley's theory of working memory involves a multi-component system for the temporary maintenance and manipulation of information (Baddeley, 1996, 2001). He proposes a system comprised of two neuroanatomically distinct "slave systems", the phonological loop and visuospatial sketchpad (VSSP), as well as a central executive and episodic buffer. The phonological loop is thought to have two components, a phonological store and an articulatory rehearsal system, which allow verbal information to be held in mind for a brief period of time through active rehearsal. The VSSP is thought to have two components as well, one for visual and the other for spatial information, with integration occurring between the two. This slave system holds visuospatial information in mind for a short time. The central executive is seen as an attentional system directing active processing and manipulation of information within the slave systems (Baddeley, 1996, 2001). The episodic buffer, the most recent component of the model, is seen as responsible for combining information between the slave systems and long-term memory, although not much research has been conducted yet on this component of the model (Baddeley, 2001). The slave systems and central executive are the aspects of this theory that are of relevance to the present discussion.

Working memory as defined by Baddeley's model is not a unitary construct, and thus although digit and spatial span tasks may both be measures of working memory, they may tap different aspects of the working memory system. This, then, could account for the literature suggesting that the tasks are not as analogous as traditionally assumed. The second goal of the present study is to try to delineate involvement of specific aspects of Baddeley's theory of working memory within the digit and spatial span tasks, as well as forward and backward components of each task.

*Evidence of the span tasks involving similar aspects of working memory.* Evidence from dual-task studies show involvement of the slave system and central executive components of working memory in forward and backward trials of both span tasks, and thus in digit and spatial span tasks overall. One study examined both a visuospatial span task (computerized Corsi test) and a verbal span task (string of consonants presented visually on a computer) (Szmalec et al., 2005). Both tasks were performed alone as well as concurrently with an articulatory suppression task (thought to utilize the phonological loop), a matrix tapping task (thought to utilize the VSSP), a choice response task (pressing a particular key in response to a high versus low tone, thought to utilize the central executive), and a simple response task (pressing a key in response to a tone, thought to minimally utilize the central executive and control for the impact of motor movement in the choice response task). When a span task is performed concurrently with an interference task, the assumption is that if span and/or interference task performance significantly decreases in comparison to performance of the tasks separately the two tasks are utilizing some of the same resources (i.e., phonological loop, VSSP, or central executive). The authors of this study concluded that VSSP resources are required in both the forward and backward tasks, although to a significantly greater degree in the spatial span forward task. Central executive resources were concluded to be involved in both forward and backward performance to a similar degree. It was also determined that forward and backward verbal span tasks require the phonological loop in a comparable manner, and while central executive resources are involved in both, the backward verbal span requires such resources to a significantly greater degree. The results of this study imply that forward and backward trials of both span tasks, and thus the digit and spatial span tasks overall, involve both slave system and central executive resources of the working memory system.

Vandierendonck et al. (2004) employed a similar dual-task paradigm looking at interference on a visuospatial span task (computerized Corsi) by articulatory suppression, matrix tapping, random interval (central executive), and fixed interval (motor control task for the random interval production) tasks. This study broke down forward and backward spatial span further into short (3-4), intermediate (5-6), and long (7-8) span lengths. The VSSP was concluded to be involved at all span lengths in both forward and backward variations, with central executive involvement in forward and backward tasks at intermediate and long span lengths. The findings are in agreement with those of Szmalec and colleagues (2005) in that the central executive appears to be involved in both forward and backward spatial span. In an earlier study by the same group (Vandierendonck et al., 1998) participants performed the same interference tasks during a verbal span task in which consonants were presented visually on a computer screen. Results suggested that the phonological loop was involved in forward and backward span tasks of all lengths, with central executive involvement in forward span trials of long length (10) and in all backward spans. Taken together, these two studies (Vandierendonck et al., 1998, 2004) suggest that central executive resources are involved in visuospatial and verbal span tasks at spans of longer length for both forward and backward recall, while backward verbal span may involve central executive resources at shorter lengths as well. In addition, recruitment of resources of their respective working memory slave systems is required in forward and backward trials of both span tasks.

*Evidence of the span tasks involving different aspects of working memory.* Several studies show differences in working memory system involvement in the span tasks and their forward and backward components. A recent factor analysis found spatial span and digit span tasks to load onto different factors, with the suggestion that while digit span measures working memory, spatial span measures visual scanning and is fairly independent of the type of immediate memory required for verbal span tasks (Giggey et al., 2006). However, since digit and word span comprised the working memory factor, with a verbal memory task and a visual memory task creating their own factors, the working memory factor could possibly reflect verbal ability of some sort. In that case, the spatial and digit span would be loading onto similar factors according to modality, revealing nothing about reliance on working memory processes. Also, the forward and backward trials of each task loaded together, adding evidence of similarity between the processes involved in the recall variations within each modality. However, it should be noted that this study was conducted in an older population (average age = 65.9 yrs), and it is possible that results could be different in a younger adult population. For instance, visuospatial ability has been shown by some research to decrease more with increasing age than verbal ability (Hester et al., 2004; Tubi & Calev, 1989) which could result in a weaker correlation between the span tasks in an older population than in a younger population.

Another study revealed that forward and backward digit span loaded onto a factor of auditory/visual working memory and complex attention, while forward and backward spatial span failed to load onto either of the two factors extracted by the analysis ("memory and information tracking" being the second factor; Mertens et al., 2006). The authors hypothesized that different mechanisms are at work in the spatial span task versus tasks requiring one to attend to and hold information in immediate memory. While these two studies do not address involvement of specific aspects of working memory, they suggest that while the digit span task appears to involve working memory processes, the spatial span task may not. Congruence of processes involved in the forward and backward trials within each modality is also indicated.

Further discrepancy between the spatial and digit span tasks has been documented. For example, while it is well known that digit span forward generally has a longer recall span than digit span backwards (e.g., Hester et al., 2004), this appears to not be replicated in spatial span. Backward spatial span lengths have been found to be equal or even superior to forward lengths (Mammarella & Cornoldi, 2005; Szmalec et al., 2005; Vandierendonck et al., 2004; Wilde & Strauss, 2002). While a study involving the WMS-III standardization sample found no evidence of this overall pattern (Hester et al., 2004), examination of the data at an individual level revealed a substantially greater percentage of people with a backward spatial span greater than or equal to their forward span (34.5%), as compared to equal or superior performance of backward digit span scores (7.1%) (Wilde et al., 2004). It has been hypothesized that such a pattern implies similar central executive involvement in spatial span forward and backwards, and limited or at least less central executive involvement in digit span forward compared to backward (Mammarella & Cornoldi, 2005). However, it is also possible that this could be due to differences in storage capacity requirements as opposed to, or in addition to, central executive requirements. Overall, such results suggest that the spatial and digit span tasks may involve different aspects of the working memory system in general, and it cannot be determined whether these discrepancies lie primarily in storage or executive resources.

However, it has been pointed out that observed discrepancy between the span tasks may have nothing to do with underlying processes at all, but merely result from differences in stimulus presentation. For example, this finding could be due to differences in the forward and backward sequences of digit span on the Wechsler tests, while the sequences are identical for spatial span (Wilde & Strauss, 2002). Additional exposure could facilitate performance on spatial span backward compared to digit span backward.

Another potential reason for this discrepancy has been outlined by Farrand and Jones (1996). They postulated that one must recall both "item" and "order" in digit span, while one must only recall "order" in spatial span. That is, the blocks are present during the recall phase and thus eliminate item recall, but the numbers are not presented again during digit recall. They matched verbal and visuospatial span tasks on item and order recall requirements, showing that when only order had to be recalled there was no difference between forward and backward recall. When both item and order had to be recalled though, forward recall surpassed backward recall on both span tasks. Based on these results the authors suggested that differences between the span tasks in the number of items recalled may stem from retrieval requirements rather than stimulus modality. Overall, the research seems to suggest superior recall for digit span forward compared to digit span backward, while equal or lower performance is true of the forward compared to the backward spatial span. This discrepancy provides evidence that the customary digit and spatial span tasks may not be as analogous as they are generally assumed to be, whether as a result of underlying cognitive processes or of stimulus presentation disparities.

*Summary of working memory involvement in the span tasks.* It is clear that there is discrepancy in the literature regarding the involvement of working memory processes in the digit and spatial span tasks. The dual-task research studies reviewed indicate that digit and spatial span tasks both involve working memory processing, with storage and central executive resources required in forward and backward trials of both span tasks. In contrast, results of two factor analytic studies suggested that working memory processing may be involved in the digit span task but not in the spatial span task. Studies that show a different pattern in average forward versus backward recall for the spatial span and digit span tasks imply that the span tasks may depend on a different combination and weighting of storage and executive components of working memory.

The goal of this portion of the present study was to explore involvement of specific aspects of Baddeley's theory of working memory within the digit and spatial span tasks, as well as forward and backward components of each task. The degree of correlation between performance on the span tasks and performance on marker tasks of the working memory slave system and central executive components was examined. In terms of Baddeley's model, maintenance and rehearsal of information involving one of the slave systems is referred to as working memory without central executive involvement (WMCE-), while active manipulation and processing of information with involvement of central executive resources is referred to as working memory with central executive involvement (WMCE+). Only those cases with a valid performance profile, as determined by a passing score on both the WMT and TOMM, were included in the analysis to eliminate the potential confound of invalid performance in exploring relationships between tests.

*Hypotheses of the Present Study*

The hypotheses of the main portion of the study were that spatial span scores would efficiently differentiate an individual's performance as valid or invalid among a sample of clients seen for suspected mild head injury in the context of civil litigation and could thus be used as a within-test measure of performance validity. It was also expected that utility of RDS as a performance validity classification technique would be replicated in the present sample.

No specific hypotheses were extended for the second portion of the study regarding the relation between span tasks and WMCE- and WMCE+ marker tasks within the subset of valid responders in the present sample since past research fails to indicate conclusively the involvement of slave system and central executive processes in the forward and backward trials within each span task and the digit and spatial span tasks overall. It was, however, expected that performance on the spatial span backward task would be equal to or greater than that of spatial span forward, and that digit span forward performance would exceed that of backward, since results of prior research seem to demonstrate this pattern consistently.

## Method

### *Participants*

Participants were drawn from an archival database of clients aged 18 to 55 years evaluated between January 2002 and February 2007 for suspected mild head injury by a neuropsychologist at an outpatient clinic in a large urban health care system in Michigan based on the following criteria: loss of consciousness (LOC) no longer than one hour, post-traumatic amnesia (PTA) extending no more than one day, and an emergency room Glasgow Coma Scale (GCS) of higher than 13 (American Congress of Rehabilitation, 1993). Since the GCS was not available for the majority of clients, injury severity was generally based upon LOC and PTA criteria. Participants were involved in civil litigation and were seen for independent evaluation ($n = 78$) or had been referred clinically (i.e., referred by primary care doctor or neurologist; $n = 4$). Exclusion criteria included history of neurological intervention (e.g., craniotomy), documented seizure disorder, brain cancer, encephalitis, stroke, myocardial infarction, substance abuse, and psychiatric history of bipolar disorder or schizophrenia. Those with a documented history of moderate to severe TBI and those who speak English as a second language were also excluded

Based on scores from well-validated forced-choice effort tests, the Word Memory Test (WMT; Green et al., 1996) and Test of Memory Malingering (TOMM; Tombaugh, 1996), participants were classified as having an Invalid (failed both), Valid (passed both), or Suspect (passed one, failed one) performance profile. A score of ≤ 82.5% on one or more of three WMT scores (Immediate Recognition, Delayed Recognition, and Consistency) constituted a failure, as did a score of < 45 on trial two or the retention trial of the TOMM.

*Measures*

Clients completed a comprehensive neuropsychological battery including the Wechsler Memory Scale-III (WMS-III; Wechsler, 1997), Wechsler Adult Intelligence Scale-III (WAIS-III; The Psychological Corporation, 1997) or Wechsler Abbreviated Scale of Intelligence (WASI; The Psychological Corporation, 1999), Wide Range Achievement Test-3 (WRAT3; Wilkinson, 1993), California Verbal Learning Test-II (CVLT-II; Delis, Kramer, Kaplan, & Ober, 2000), Judgment of Line Orientation (JOLO; Benton, Hamsher, Varney, & Spreen, 1983), Ruff Figural Fluency (Ruff, 1988), Ruff 2 & 7 Selective Attention Test (Ruff & Allen, 1996), Finger Tapping Test (FTT; Reitan & Wolfson, 1985), Grooved Pegboard (Matthews & Klove, 1964), Grip Strength test (Reitan & Wolfson, 1985), Wisconsin Card Sorting Test (WCST; Heaton, Chelune, Talley, Kay, & Curtis, 1993), Trail Making test, Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Greene, Brown, & Kovan, 1998) or Personality Assessment

Inventory (PAI; Morey, 1996), and the Test of Memory Malingering (Tombaugh, 1996) and/or the Word Memory Test (Green et al., 1996). The core measures used in the present study included the TOMM, WMT, spatial span and digit span sub-tests of the WMS-III, CVLT-II, and Arithmetic sub-test from the WAIS-III.

*Test of Memory Malingering.* The TOMM is a forced-choice test that was designed to measure effort and malingering. The client views 50 simple pictures one at a time, after which the examinee is asked to choose which of two pictures was presented before. The second trial is identical to the first, except the pictures are presented in a varied order, and new foil pictures are presented during the testing phase. A delayed retention trial involves only the test phase, in which the original pictures are shown with different foils than in the previous trial. The examiner gives feedback regarding responses during all three test phases. A score of less than 45 correct on trial two or the retention trial has been shown to be suggestive of malingering (Tombaugh, 1996).

*Word Memory Test.* The WMT was designed to measure malingering as well as to assess cognitive functioning. It is a computerized test in which a person learns a list of 20 semantically-related pairs of common words (e.g., dog/cat) that is presented twice at a rate of 6 seconds per word pair. This is followed by an immediate recognition subtest where the person selects each of the 40 original words from 40 new pairs, in which the original words are now paired with a foil word (e.g., dog/rabbit, cat/mouse). A delayed recognition subtest is administered 30 minutes later in which the person selects each of the 40 original words from 40 new pairs, with the original words again paired with new foil words (e.g., dog/rat). A consistency score between immediate and delayed recognition tests is also computed. A score at or below 82.5 percent correct on any of these three conditions is indicative of invalid performance (Green et al., 1996).

Since it has been stated that a test of performance validity developed on a certain clinical population is not necessarily relevant to another population (Faust & Ackley, 1998), it should be noted that both the TOMM and WMT used TBI groups in their normative samples. This is of importance given that the clients in the present study are suspected to have had a mild TBI.

*Spatial span.* The spatial span task is a measure of visual attention and working memory in which the examiner touches a series of blocks, after which the client must touch the same blocks in the same order. The test includes a forward and backward trial with the series increasing in length throughout (Wechsler, 1997).

*Digit span.* The digit span task is a measure of verbal attention and working memory in which the client is required to repeat sequences of digits that increase in length throughout the test. Like spatial span, this task includes a forward and backward trial (Wechsler, 1997).

*CVLT-II (WMCE- marker task).* The CVLT-II is a verbal learning and memory list task. Trial one of the CVLT-II was used in the current study and involves the examiner reading aloud a list of 16 words (which are comprised of 4 semantic categories), after which the client repeats back as many of the words as possible in any order (Delis et al., 2000). Trial one of the CVLT-II was used as a cognitive marker task of WMCE-, and was conceptualized as involving the phonological loop, one of the two slave systems outlined in Baddeley's model of working memory (Baddeley, 1996, 2001). This construct is defined as the maintenance of information by the slave system for a brief time period, with no involvement of the central executive. From the battery given to the present sample, no task appeared to be a good measure of the VSSP. It is assumed that although the spatial span task would be more modestly associated with the WMCE- marker task than would digit span, inferences about the contribution of storage capacity to spatial span might be made on the basis of the relationship between spatial span and trial one of the CVLT-II.

*Arithmetic (WMCE+ marker task).* The Arithmetic subtest of the WAIS-III is a test of working memory in which math problems of increasing difficulty are read aloud to the client, who must mentally calculate the answer and report it aloud within a time limit (The Psychological Corporation, 1997). The Arithmetic subtest of the WAIS-III is widely accepted as having a large manipulation component in addition to requiring the examinee to hold information in mind (DeStefano & LeFevre, 2004; Imbo, Vandierendonck, & De Rammalaere, 2007), and was thus used as a measure of WMCE+, conceptualized as using one of the two slave systems with additional involvement of the central executive. Similar to the WMCE- measure, the WMCE+ measure is largely in the verbal modality and thus likely would require storage by the phonological loop. However, it could be argued that spatial ability is required in the Arithmetic sub-test as well, and therefore that the visuospatial sketchpad may be involved to some degree. Again, by looking at relative correlation patterns, it was hoped that further information could be obtained with regards to WMCE+ involvement in these span tests.

*Spatial Span alternatives.* Five spatial span alternatives were calculated including a reliable spatial span (RSS; the sum of the longest string forward and backward correctly repeated over two trials, identical to the RDS), longest string forward correctly repeated over two trials [SSf(2)], longest string backward correctly repeated over two trials [SSb(2)], longest string forward [SSf(raw)], and longest string backward [SSb(raw)].

*Digit Span Alternatives.* Five digit span alternatives were also calculated including reliable digit span (RDS; the sum of the longest string forward and backward correctly repeated over two trials), longest string forward correctly repeated over two trials [DSf(2)], longest string backward correctly repeated over two trials [DSb(2)], longest string forward [DSf(raw)], and longest string backward [DSb(raw)].

# Results

## *Demographic Characteristics*

Once exclusion criteria were applied to the 201 cases, the total number of participants was 83 (31 male, 52 female). Average age was 38.9 years (*SD* = 10.3, range 18 to 55), with an average education level of 12.6 years (*SD* = 2.1, range 9 to 20). Race included 49 Caucasian, 31 African American, and 3 participants of some other background. Seventy-six participants were right-handed, 5 left-handed, and 2 ambidextrous. Time since the injury ranged from 43 to 2835 days (*M* = 806, *SD* = 611). Testing was conducted more than six months post-injury in 94% (*n* = 78) of the cases, and 81% (*n* = 67) were tested more than one year post-injury. Date of injury was missing for 2 cases. A total of 79 were involved in litigation and 4 were clinically referred. The majority (*n* = 80) were involved in a motor vehicle accident (MVA; 76 as a driver or passenger, 4 as a pedestrian) and 3 had received a blow to the head by other means (car hood fell on head – clinical referral, sign fell on head, hit on head by a falling box while shopping). Information regarding site of the injury, however, was not available for any of the cases.

Groups (Valid [$n = 29$], Invalid [$n = 33$], Suspect [$n = 21$]) were analyzed using a one-way Analysis of Variance (ANOVA) or Chi-Square analysis on gender, age, race, education, and time since injury. No significant differences were observed among the three groups (see Appendix A). Litigation status, injury type, handedness, length of LOC, PTA, and GCS could not be submitted to Chi-Square analysis due to low expected cell frequencies, but visual inspection suggests no group differences (also see Appendix A).

*Analysis of Variance Assumptions*

The Shapiro-Wilk test of normality was performed for each spatial and digit span alternative [RSS, SSf(2), SSb(2), SSf(raw), SSb(raw), RDS, DSf(2), DSb(2), DSf(raw), DSb(raw)] by group (Valid, Invalid, Suspect). All but 9 of 30 variables [RSS and RDS in the Valid group, RSS, SSb(raw), and DSf(raw) in the Invalid group, and RSS, SSf(2), DSf(raw), and DSb(raw) in the Suspect] were significant, meaning that data in the majority of the groups did not follow a normal distribution. Scores on all of the alternatives of spatial and digit span were converted to z-scores, and ANOVAs were run with and without outliers greater than 3.29 ($p < .001$) (Field, 2005; Tabachnick & Fidell, 2001). Because the results did not change in any meaningful way and because there was no reason to believe that these cases did not actually belong to their group population, it was decided that the outliers would not be removed. Skewness and kurtosis values were also obtained and transformed to z-scores. DSf(raw) in the Valid group, RDS, and DSf(2) in the Invalid group, and RDS and DSb(2) in the Suspect group had skewness and/or kurtosis z-scores greater than 2.58, a level that has been suggested as indicative of a significant violation of normality (Field, 2005). Since such violations only serve to make a test more conservative due to reduced power (kurtosis) or have no impact at all (skewness), and the ANOVA is robust to violations of the normality assumption (Kirk, 1995), it was decided that the use of a different statistical test was unwarranted. Homogeneity of variance was tested for each alternative of spatial and digit span by group using Levene's test. All tests were non-significant, meaning that the assumption of homogeneity of variance was met. Independence of observations, the last ANOVA criterion, was assumed. Therefore, it was judged that the data could be assessed using an ANOVA.

*Spatial Span Alternatives*

One-way ANOVA ratios were significant for four of the five spatial span alternatives (see Appendix B). Tukey post-hoc analyses were conducted on the significant ANOVA results. The Suspect group did not differ significantly from the Invalid or the Valid group for any alternative of spatial span (see Appendix C). Because sensitivity and specificity calculations require a comparison of two groups, it was decided that the Suspect group would be excluded from these calculations. Had the Suspect group differed significantly from the Valid group but not the Invalid group, the Suspect and Invalid groups would have been combined into one for these calculations. The inverse strategy could be used if the Suspect group differed significantly from the Invalid group but not the Valid one. Results from Tukey post-hoc analyses also revealed that the Invalid and Valid groups differed significantly on all four spatial span alternatives for which the overall ANOVA was significant (see Appendix C). Because the significance values of these four spatial span alternatives were close in size, all four were submitted to sensitivity and specificity value calculations to determine which has the best utility in detecting invalid performers.

*Sensitivity, specificity, and positive predictive power values.* Classification accuracy was determined based on the methodology used by Heinly and colleagues (2005), which is outlined here using the RSS alternative as an example. First, the cumulative percentage of individuals in each group (Invalid and Valid) obtaining particular scores on RSS was identified (see Appendix D), providing sensitivity and specificity values for each score. Sensitivity is the true positive (hit) rate and is the cumulative percentage of people correctly classified as showing invalid performance by scoring at or below a particular RSS score. The cumulative percentage of valid performers scoring at or below this RSS score is the false positive error rate and is subtracted from 1 to give the specificity. Thus, specificity is the percentage of valid cases correctly classified at a particular RSS score.

Sensitivity and specificity values were calculated for a range of scores on the four significant spatial span alternatives, RSS, SSf(2), SSb(2), and SSf(raw). These values are presented in Appendix E. For each spatial span alternative, sensitivity and specificity values were examined simultaneously across the range of scores in order to recommend a cut-off score that correctly classifies the largest number of invalid performers (high sensitivity) while resulting in very few misclassifications of valid responders (high specificity). In other words, in determining a cut-off score, a compromise must be made between test sensitivity and specificity. A cut-off score with low sensitivity includes a large number of false negatives, or participants with an invalid performance profile who go undetected. Poor specificity, on the other hand, means the cut-off score results in a large number of false positives, or valid responders misclassified as invalid. It has thus been suggested that when determining a cut-off score it is preferable to misclassify invalid profiles as valid rather than the opposite misclassification (Type II error), and more weight should therefore be put on obtaining a high specificity value (Greve & Bianchini, 2004). Of the four spatial span alternatives, RSS provides the best compromise between specificity and sensitivity resulting in relatively high values of both. A cut-off of six or less on RSS correctly classifies approximately 55% of invalid performers and misclassifies 14% of valid performers. A cut-off of seven or less correctly classifies approximately 70% of invalid performers with a slight rise in misclassifications to 20%. Because specificity and sensitivity values are based on a sub-set of the population, they have associated estimation errors. Error associated with a 95% confidence interval was calculated for each RSS score using a computer program called DAG_Stat (Mackinnon, 2000). This range of percentages more accurately represents the true sensitivity and specificity of the particular RSS score (see Appendix F).

Sensitivity and specificity are independent of base rates (Greve & Bianchini, 2004), which are especially important to consider given the wide variation in malingering rates reported among clinics. Base rate reports range anywhere from 20 to 60% (Binder, 1993; Greiffenstein et al., 1994; Meyers & Volbrecht, 1998). This rate depends to a large degree on what percentage of clinic cases are litigation as opposed to clinically-referred in nature. The base rate of the present sample was 53%, calculated with only the Valid and Invalid groups as these were the two used in sensitivity and specificity calculations.

Base rates are accounted for by calculating predictive power, which tells the clinician how confident he or she can be that an individual's performance validity test result is accurate (Greve & Bianchini, 2004). In the present study, positive predictive power (PPP) was calculated for a range of RSS scores to enable the clinician to determine the probability that a client's profile is invalid, given the individual's RSS score and the base rate in that particular clinic. The base rate of invalid performance from the present sample (approximately 50%), as well as base rates between the range that others have suggested (20 to 60%) was used (see Appendix F). For example, in a clinic with a 50% base rate a client with a RSS value of 6 can be assumed to be performing in an invalid manner with a probability of 80%. In reality, the predictive power is also associated with error in estimation of the sensitivity and specificity values and is better expressed as a range of probabilities. These probability ranges are also presented in Appendix F. PPP was calculated using the following formulae (Heinly et al., 2005): Likelihood ratio = sensitivity / (1 - specificity); pretest odds = base rate / (1 - base rate); posttest odds = likelihood ratio x pretest odds; PPP = posttest odds / (1 + posttest odds). As can be seen, PPP is dependent on the accuracy of the test (specificity and sensitivity) and on the base rate of the condition in the population in question (Greve & Bianchini, 2004). Thus, predictive power cannot be high if the base rate is low, even if sensitivity and specificity are good (Slick et al., 1997).

*RDS Replication*

One-way ANOVA tests revealed that none of the five digit span alternatives had a significant result, although results from all five show a trend toward higher scores by the Valid group (see Appendix G). Although it was non-significant, sensitivity and specificity were calculated using the Valid and Invalid groups for RDS to allow direct comparison of values to previous research.

*Sensitivity, specificity, and positive predictive power values.* Sensitivity and specificity values were calculated for RDS (see Appendix H). RDS has a specificity of

79% at a value of 6 with classification of invalid performers (sensitivity) at 27%. Estimation error was calculated at a 95% confidence interval for values of RDS. Positive predictive power was also calculated for a range of RDS values which showed the best sensitivity and specificity values. These values are presented in Appendix H.

*Correlations of Span Tasks with WMCE- and WMCE+*

Data on two cognitive marker tasks (CVLT-II trial one [WMCE- task] and Arithmetic [WMCE+ task]), as well as six span task variations (longest spatial span forward [SSf], longest spatial span backward [SSb], total of the longest forward and backward spatial span [SSt], longest digit span forward [DSf], longest digit span backward [DSb], and total of the longest forward and backward digit span [DSt]) from the 29 Valid subjects were explored using Spearman's rho correlations. One case was missing CVLT-II trial one data ($n = 28$) and 5 cases were missing Arithmetic data ($n = 24$). Four of the 6 variables followed a non-normal distribution as determined by the Shapiro-Wilk's test of normality (DSt and SSt were the exceptions). Neither removal of outliers greater than a z-score of 3.29 nor replacement of the outlier's score with that of the next closest case plus one (Tabachnick &Fidell, 2001) affected normality values, and removal of outliers resulted in noticeable changes in Pearson correlation coefficients. It was therefore decided that ranking the data for use with Spearman's rho correlation test was more appropriate than use of Pearson's correlation test.

Spearman's rho correlation coefficients are presented in Appendix I. Because these were exploratory correlations without specific hypotheses, no correction was done for significance levels of p-values. Significant correlations ($p < .05$) were observed in all cases except DSf with CVLT-II trial 1.

The mean scores of all variables involved in the correlation analyses are presented in Appendix J. The average SSf and DSf scores are higher than the SSb and DSb scores, respectively. SSf was higher than SSb in 9 cases (31%), lower than SSb in 2 cases (7%), and equal in 18 cases (62%). DSf was higher than DSb in 24 of 29 cases (83%), lower than DSb in 2 cases (7%), and equal in 3 cases (10%).

## Discussion

The main goal of the present study was to explore the utility of the spatial span task as a within-test measure of performance validity for use with clients presenting with a suspected mild TBI. It was hypothesized that spatial span scores would have the ability to classify clients as either valid or invalid responders. RSS scores showed a good balance of sensitivity and specificity, thus supporting this hypothesis.

Digit span was also examined in the present sample in an attempt to replicate previous research which has documented high classification accuracy of RDS. It was hypothesized that RDS classification accuracy would be replicated in the present sample and provide further support for its use in a clinical setting as a within-test measure of performance validity. Other variations of digit span scores were also examined. These hypotheses were not supported since classification accuracy of RDS was found to be substantially lower than that of prior research. Better classification accuracy by RSS than RDS suggests that spatial span may be more susceptible to invalid performance than digit span is, at least within the present sample.

The second goal of the study was to explore the involvement of aspects of working memory in the spatial and digit span tasks. Examination of correlation coefficients suggested that slave system and central executive resources are involved in the digit and spatial span tasks overall, the forward and backward trials of the spatial span task, and the backward trial of the digit span task. Only central executive resources were found to be involved in the forward trial of digit span. Some support was provided for the increasing evidence in the literature that the digit and spatial span tasks are not as analogous as traditionally assumed.

*Spatial Span as a Performance Validity Indicator*

Spatial span, in particular the RSS calculation, appears to have promise as a classification index of performance validity in neuropsychological testing of suspected mild TBI. A cut-off score of 6 or less is recommended, correctly detecting 55% (true estimate of 36 to 72%) of invalid performers with a low misclassification rate of valid performers at 14% (true estimate of 4 to 32%). Alternately, a score of 7 or less correctly detects a substantially greater proportion of invalid performers (70% [true estimate of 51 to 84%]) while still maintaining a relatively low percentage of false positives (20% [true estimate of 8 to 38%]). These are both acceptable rates, given values reported by other within-test measures of performance validity that have been developed (Greiffenstein et al., 1995; Greve et al., 2003; Heinly, 2005; O'Bryant et al., 2003). However, considering the upper limit of the true estimate of false positives with a score of 7 or less nearing 40%, a cut-off 6 or lower is perhaps a better alternative. Previous measures of performance validity have not reported the error of estimation associated with sensitivity and specificity values (with the exception of Heinly et al.'s 2005 replication of RDS), and they are therefore likely subject to the same high upper limits of false positive rates. Other alternative estimates of performance validity using spatial span examined in the present study failed to show as high classification accuracy as RSS. Therefore only the

RSS alternative is recommended for use as a within-test measures of performance validity.

Positive predictive power was calculated for a range of cut-off scores with varying base rates of invalid performance. There is generally no recommendation in the literature regarding what constitutes acceptable predictive power. This is a decision that must be made by each clinician who, by looking up a client's RSS score in the table provided and knowing the approximate base rate in the clinic in question, can decide what level of predictive power he or she is comfortable with using. True estimates of the predictive power should also be considered in making this decision. For an RSS score of 6 or less, positive predictive power is 50% or higher from even the lowest base rate of 20%, or at a base rate of 30% and higher when considering the true PPP range. As cut-off values rise, positive predictive power decreases.

With a sensitivity value of 55 to 70%, associated with RSS cut-off scores of 6 and 7 respectively, 30 to 45% of clients in the Invalid performance group remain undetected. That is, a number of clients who failed both the WMT and TOMM, and therefore were classed as invalid responders, attained a score higher than the RSS cut-off recommended to be indicative of invalid responding. The question that arises is why these clients were not classified as invalid by RSS. First, it is possible that those clients who remain undetected by RSS actually did have valid performance throughout their testing session. This, however, is unlikely given that they failed two well-validated performance validity tests. Second, it is possible that while they did perform in an invalid manner on the WMT and the TOMM, they performed in a valid manner on the spatial span task. A single test can only measure validity of responding at a particular point in time, and it is very possible that performance could vary throughout the test battery. Malingerers generally choose specific areas of functioning to fake or exaggerate as opposed to manifesting a general cognitive deficit (Greiffenstein et al., 1995; Iverson & Binder, 2000). Perhaps these clients were true malingerers who perceived the spatial span task as not involving processes in which they were trying to portray deficits. Or these clients may have had varying effort across the test battery and for some reason may have been more awake and engaged during the spatial span task than during the WMT or TOMM.

In addition, specificity rates of 80 to 86% mean that 14 to 20% of the valid group, who passed both the TOMM and WMT, are classified as invalid responders with RSS cut-off scores of 7 and 6. Similar to the opposite misclassification, it is possible that these clients were fatigued or lacked engagement during the spatial span task, thereby causing their performance to be invalid on this task although it was valid on the WMT and TOMM. Test anxiety could also have caused low performance on the spatial span task while it would not have resulted in failure on the WMT and TOMM. Anxiety has been shown to affect working memory (e.g., Darke, 1988), which is measured by the spatial span task. Another possibility is that this percentage of the Valid group actually has cognitive deficits due to brain injury or some other neurological disorder that impacted performance on the spatial span task.

The spatial span task was designed to measure an aspect of cognitive function while the TOMM and WMT were specifically designed to have a low difficulty level so that failure would occur rarely even in individuals with well-documented moderate to severe cognitive deficits (e.g., neurological patients with brain tumour, stroke, severe TBI, as well as others have been shown to reliably pass; Green et al., 1996; Tombaugh, 1996). Therefore, it is plausible that a few individuals with mild TBI might pass malingering tests but score below the RSS cut-off as a result of marked deficits in visuospatial working memory due to mild TBI. Although the majority of individuals sustaining a mild TBI spontaneously recover to pre-injury levels in a short period of time and suffer no permanent deficits, there is a minority of individuals who have persisting deficits in areas of cognitive function (Alexander, 1995; McHugh, 2006; Reitan & Wolfson, 2000). While deficits following a mild TBI can occur in a range of cognitive domains across individuals, the majority of studies indicate that persisting deficits occur in areas of attention, working memory, processing speed, and executive function (Alexander, 1995; Echemendia et al., 2001; ImPACT, 2006; McHugh et al., 2006; Psychological Corporation, 1997). Statements regarding the possibility of cognitive deficits following a mild TBI have especially strong support from studies including athletes who have experienced a concussion and are required to be back to pre-injury level of cognitive functioning in order to return to play (e.g., Iverson, Brooks, Collins, & Lovell, 2006). Subjects in the return to play studies can be assumed to have motivation to perform as well as possible, giving strong support to the evidence that cognitive deficits do occur in individuals who have sustained even a very mild TBI. Therefore, it is very possible that a subset of the Valid group had a deficit in the area of working memory which could have impacted spatial span performance. Naturally, this is a possibility with the Invalid group as well, but since they responded in an invalid manner on two well-validated effort tests, no conclusions about their true cognitive abilities can be made.

Of course, it must be reiterated that the RSS does not have sufficient classification power to act as a stand-alone determinant of invalid performance. If a client scores below the recommended cut-off, this score must be considered along with other measures of performance validity as well as clinical judgment. Ideally, to act alone as a classification technique, the false positive rate would be quite close to 0% and the detection rate would be as close to 100% as possible.

*Replication of Reliable Digit Span*

The use of RDS as a method of performance validity classification has been supported by several studies, both for use in clinical TBI cases (Babikian et al., 2006; Greiffenstein et al., 1994, 1995; Heinly et al., 2005; Mathias et al., 2002) as well as for other populations (Babikian et al., 2006; Duncan & Ausborn, 2002; Etherton et al., 2005; Inman & Berry, 2002; Strauss et al., 2002). However, the present results failed to support this, with RDS having substantially lower classification accuracy than that of the majority of the previous research studies. The best balance achieved between specificity and sensitivity for values of RDS in the current sample was at a cut-off of 6. Specificity was somewhat low although still adequate at 79% (true estimate of 60 to 92%), with a sensitivity of only 27% (true estimate of 13 to 46%). Past research has recommended a RDS cut-off score of 7 or lower, which corresponds to a specificity of 59% (true estimate of 39 to 77%) and a sensitivity of 49% (true estimate of 31 to 67%) in the current sample. This results in an unacceptably high false positive rate of 41%. Previous RDS studies using clinical TBI samples, on the other hand, have found specificity and sensitivity values, respectively, of 89% and 68% (Greiffenstein et al., 1994), 93% and 67% (Mathias et al., 2002), 83% and 71% (Heinly et al., 2005), 77% and 44% (Babikian et al., 2006), and 68% and 89% (Greiffenstein et al., 1995). While the last two studies document a relatively low specificity, the other three studies show good to very high classification accuracy with specificity and sensitivity levels much better than those obtained in the present study. However, when the upper level of estimation error associated with the sensitivity and specificity values in the present study is considered, since the confidence interval is a truer estimate of the actual values, our findings perhaps do not show as much discrepancy from previous research as the single specificity and sensitivity levels imply.

A number of reasons may account for the failure of the present study to strongly replicate RDS classification accuracy of prior clinical TBI research, all of which have to do with methodological differences between the studies. The first, and likely the largest factor, is the set of criteria used to classify subjects as invalid or valid responders. Since specificity and sensitivity values are based on the cumulative percentages of each group at or below a particular RDS score, they are directly dependent on how clients were classified into groups. The previous RDS studies used a combination of objective (test scores) and subjective (clinical judgement) methods to classify subjects (see Appendix K), while the present study relied solely upon objective methods. The archival nature of the present study made access to subjective data non-uniform across individuals. Generally, neuropsychological test data and the final report were all that was available at the time of data coding. Thus, while information to make decisions on subjective data would have been present in some cases, it would not have been available for all. As such, group classification based on subjective data, in addition to objective data, is more suited to a prospective study design. Therefore, to provide continuity across all subjects in the current study, use of two objective performance validity measures was chosen. This, however, may have resulted in substantial differences across the studies with respect to the type of clients making up the invalid and valid groups.

In addition to discrepancy with respect to the involvement of subjective criteria, variation in the objective measures that were used across the studies could have resulted in sizeable differences in the initial group classification. Both studies by Grieffenstein and colleagues (1994, 1995) used an impairment rating of severe on two or more neuropsychological tests throughout the assessment battery to meet the objective criteria. Mathias and colleagues' (2002) objective criterion was a failure on either the TOMM or the PDRT. Babikian et al. (2006) used two within-test measures, and the study by Heinly et al (2005) used failure on either the TOMM or PDRT or else failure on two or more within-test measures. Within-test measures of performance validity do not have as high classification accuracy as tests developed specifically to measure effort, especially when used alone or in combination with only one other within-test measure. Thus, the RDS studies involving within-test measures as objective grouping criteria likely have a substantial difference from the present study which employed two tests that have been designed specifically to measure performance validity.

In addition, tests designed specifically for the purpose of testing validity of performance also display disagreement in classification of individuals. For instance, discrepancy has been found between classifications made by the TOMM and WMT with 23% of a sample of 1046 passing the TOMM but failing the WMT while less than 1% showed the opposite pattern (Green, 2005). Given that two of the prior RDS studies used the TOMM as grouping criteria, this could result in a large degree of discrepancy in the initial classification of invalid responders compared to the present study. Therefore, it is possible that marked grouping differences exist between the present study and prior studies as a result of dual use of the WMT and TOMM in making classifications in this study.

Further, classification of invalid performance in the two studies conducted by Greiffenstein et al (1994, 1995) may not have even involved a piece of objective criterion because a person could be placed in this group by meeting two of the subjective criteria alone. In addition, it is possible that the studies using subjective criteria could have based discrepancy between test results and known patterns of brain functioning/documented history/behavioural observations in part on digit span scores. In other words, it appears that in most studies of within-test performance validity indicators which primarily assess cognitive abilities, the score in question was available to the clinician who determined whether an atypical pattern of test findings was present overall. This could cause lower RDS scores within the pre-classified invalid group, and artificially increase classification accuracy of RDS. Overall, it appears that criteria used for initial group classification may account to a large extent for failure of the present study to replicate RDS classification accuracy of prior research.

Second, severity of TBI is not constant across studies. The study of Mathias and colleagues (2002) included mild, moderate, and severe TBI. Babikian and colleagues (2006) failed to specify the injury severity of their TBI sample. The other three studies had a moderate to severe TBI comparison group in addition to their mild TBI group, but classification accuracy values were calculated separately for the two groups and only the mild TBI group values were compared to the present study.

Also, mild TBI definitions vary somewhat across studies (see Appendix K) although it appears that differences in definitions did not likely result in a large discrepancy of the injury severity across studies. Two of these studies used a sample comprised of individuals with a history of very mild TBI (Greiffenstein et al., 1994, 1995), but while criteria are different from the present study they may not necessarily result in large actual differences between the groups. The largest difference between the present study and that conducted by Heinly and colleagues (2005) was their slightly shorter LOC length of 30 minutes. However, although the LOC time length in the present study was set at one hour a priori, due to other exclusion criteria no subjects actually had an LOC longer than 30 minutes.

Other methodological differences, although not as obvious as those just discussed, also could have also influenced the failure of RDS replication. The number of subjects included in the studies could be a potential influence, although the sample size of the present study was not markedly different from that of Greiffenstein and colleagues (1994) and Mathias and colleagues (2002). The former had 30 valid and 43 invalid, and the latter 30 valid and 24 invalid, in comparison to the current sample of 29 valid and 33 invalid. Numbers were even lower in the TBI portion of Babikian and colleagues' (2006) sample, with 13 valid and 28 invalid. The others had higher numbers of 53 valid and 68 invalid (Greiffenstein et al., 1995) and 77 valid and 48 invalid (Heinly et al., 2005). It has been pointed out that a small valid group makes cut-off scores less stable across studies (Greve et al., 2004). Also, although three of the five prior RDS studies had valid and invalid groups that were comprised mainly of those with financial incentive, similar to the present study, two studies (Babikian et al., 2006; Mathias et al., 2002) had a valid group without financial incentive. Given the large influence this has been shown to have, such a difference warrants caution in making direct comparison of these two studies to the others. A fifth potential influence could be age of the subjects, as digit span scores are known to decrease with age (Hester et al., 2004; Wilde, Strauss, & Tulsky, 2004). However, with the exception of one sample that had an average age of 43 to 48 (Babikian et al., 2006), the mean age in other studies was close to that of the current sample (38 years) with average ages ranging from 34 to 41 years.

Overall, a number of methodological differences exist between the present study and prior RDS studies that have used clinical TBI samples. The studies seeming to have the largest methodological differences from the present study are those conducted by Mathias and colleagues (2002) and by Babikian and colleagues (2006), suggesting direct comparisons between classification accuracy may be inappropriate. This leaves three studies (Greiffenstein et al., 1994, 1995; Heinly et al., 2005) more relevant to the present one. This results in a range of 68 to 89% for both sensitivity and specificity values. The values of the present study still fall short, even when considering their associated confidence intervals. Discrepancy of classification accuracy values between the current study and these other three RDS studies may be largely due to differences in criteria used to group clients into valid and invalid groups. This is obviously an important factor to be considered by clinicians who use these suggested cut-offs, as they are in turn implicitly applying the same classification criteria to their own clients.

*Classification of Performance Validity: Spatial Span versus Digit Span*

The results of the present study suggest that spatial span may have better classification of performance validity than does digit span. RSS specificity (80 to 86%) and sensitivity (55 to 70%) values at scores of 6 and 7 are within the range of values reported for RDS in previous clinical TBI samples of 68 to 93% specificity and 43 to

89% sensitivity. Of course, such comparisons are subject to the methodology differences just discussed for across-study comparisons of RDS. The present examination of RSS and RDS within the same sample, which has the benefit of identical group criteria and other methodological factors, shows that RSS has considerably higher specificity and sensitivity than RDS. Such a finding is not entirely unexpected.

A previous study looking at the utility of the entire WMS-III for validity classification in mild head injury litigants found that although both digit and spatial span total scores had low sensitivity, the spatial span score detected more invalid performers (24%) than did the digit span score (16%) at an identical specificity level for a total span score of 8 or lower (Langeluddecke & Lucas, 2003). Additionally, a study employing simulated malingering using the WMS-Revised found the difference between simulated malingerers and controls was greater for spatial span total score (termed "Visual Memory Span" in the WMS-R) than it was for digit span total score (Bernard, 1990). Although the latter study does not use a real-world sample, it still demonstrates that the spatial span task may be more susceptible to exaggeration or faking than the digit span task. This also makes sense when one considers that overall spatial span scores tend to be lower than digit span scores (Hester et al., 2004). The greater difficulty level of spatial span may make it more susceptible to incomplete effort and also lend itself to being chosen as a task that a client who is malingering would deem as one on which someone with cognitive deficits would be unable to perform well.

Given the popularity of RDS as a within-test measure of performance validity, the present study suggests RSS may have similar or even greater utility. This pattern of higher classification by RSS may, of course, be merely a phenomenon restricted to the present sample. To determine relative accuracy of classification of the two methods, it would be necessary to examine these two calculations together in further samples to see whether the pattern found between RSS and RDS in the present study is replicated. It may be especially enlightening to compare the two in a study using alternate methods of determining group membership: two pieces of objective evidence as used in the present study versus one piece each of objective and subjective evidence as used in previous RDS research.

*Generalizability*

The findings of the present study indicate that calculation of RSS from clients' spatial span results has promise as a classification index of performance validity in neuropsychological testing of suspected mild TBI. Scores lower than 6 or 7, depending on what levels of specificity and sensitivity each clinician is comfortable using, are recommended for use in detecting invalid performance in combination with other previously validated measures. It is important to note that this method has some limitations. These results may not be applicable to clients outside an age range of 18 to 55, or clients with a history of substance abuse, prior neurological intervention, prior moderate to severe TBI, and diagnosis of schizophrenia, nor is it applicable to individuals whose first language is other than English. For the purposes of an initial exploratory study, it was felt that stringent selection criteria for inclusion were necessary to minimize potential confounds of other factors that could impact test data in addition to validity of performance. It would be of interest for future research to determine whether the present results can be extended to older adults and to those with a history of substance abuse, given the prevalence of these individuals in clinical settings, in order to increase the utility of RSS as a within-test measure of performance validity.

Another limitation is that a number of participants ($n = 32$) were excluded from the analysis because they had been administered only a single performance validity measure. When planning the inclusion criteria of the present study, it was not known that this would be the case. Had inclusion criteria differed and these participants could have been included, it is possible that the results of the analyses may have been altered to some extent due to the impact this number of participants can have in cases of relatively small sample sizes. Because of the retrospective nature of the study it is unknown why these participants were only administered a single performance validity measure, but it is possible that obvious poor effort could have been one reason. This, then, may have excluded a large number of invalid performers from the analysis.

Because a real-world sample of suspected mild TBI cases was used, the present results likely have good ecological validity. While it would, of course, be preferable to have an experimental group of invalid responders compared to a control group of valid responders, such a comparison is impossible since it is never known who actually is performing in an invalid manner (Faust & Ackley, 1998). Some researchers have attempted to get around this issue by using a group of simulated malingerers (e.g., Bernard, 1990; Inman & Berry, 2002; Strauss et al., 2002). For the present study, use of such a group was rejected due to research revealing that the general population usually does not have an accurate perception of what sequelae follow a traumatic brain injury (e.g., Hux, Schram, & Goeken, 2006). In addition, the incentive to perform poorly is not comparable in clinical patients who have the possibility of large monetary gain and experimental simulators, and the former are likely more skilled than simulators (Faust & Ackley, 1998). Because of this, high classification rates of simulation studies often plummet when used on a clinical group suspected of invalid performance, making generalization of such findings to the real world questionable (Rogers et al., 1993).

The 53% base rate of invalid performance in the present study's sample (which was calculated using the Valid and Invalid groups) is close to the 49% rate reported by Meyers and Volbrecht (1998), and lends support to the suggestion made by Greiffenstien and colleagues (1994) that base rates may be higher than has generally been assumed within a population of litigating clients seen for independent evaluation with a suspected mild TBI. While a little lower than the 60% rate reported by Greiffenstein, it is higher than that of other reports (Binder, 1993; Constantinou et al., 2005; Langeluddecke & Lucas, 2003).

The finding that spatial span is able to classify invalid and valid performance with a high level of accuracy adds another within-test performance validity measure to the neuropsychological test battery. The move toward the development of within-test measures has been supported by the need to decrease time and money spent on neuropsychological examination (Meyers & Volbrecht, 2003; Langeluddecke & Lucas, 2003), to insert validity checks throughout the entire assessment (Meyers & Volbrecht, 1998), and to reduce the susceptibility of validity tests to lawyer coaching (Mathias et al., 2002). In addition, it has been noted that a client will generally attempt to feign symptoms in a specific area of cognitive functioning rather than feign a global impairment (Greiffenstein et al., 1995). Since performance validity measures currently in use are mainly in the verbal modality, a client exaggerating or faking non-verbal deficits could avoid detection. Therefore, the results of the present study not only add a within-test measure of performance validity, but also one in the non-verbal domain, and thereby make the findings especially useful.

*Aspects of the Working Memory System Underlying the Span Tasks*

Results seem to indicate that the total spatial span task, total digit span task, spatial span forward and backward, and digit span backward scores reflect contributions by the working memory slave system and central executive. The significant correlation coefficients are all around the 0.3 to 0.5 range, indicative of moderate correlation sizes. The present findings are in contrast to the two factor analytic studies reviewed which suggested no involvement of working memory in the spatial span task (Giggey et al., 2006; Mertens et al., 2006). The results are somewhat in line with dual-task studies which have indicated both storage capacity and central executive involvement in the forward and backward trials of the spatial span task, as well as digit span (Szmalec et al., 2005; Vandierendonck et al., 1998, 2004).

The non-significant correlation between digit span forward recall and the WMCE-marker task was unexpected and is contrary to these dual-task studies which have suggested involvement of the phonological loop in addition to the central executive in digit span forward recall (Szmalec et al., 2005; Vandierendonck et al., 1998). However, it is noted that the present results should be interpreted somewhat cautiously because the cognitive marker tasks employed are not pure measures of working memory system components. For example, while trial one of the CVLT-II was deemed to be the best cognitive marker of the phonological loop within the battery of tests available in the database, it is not a pure measure of the holding capacity of the phonological loop without any involvement of active processing and manipulation. It could be argued that organizing the word list into semantic categories involves manipulation of information to a certain extent, and that recall of the CVLT-II word list may depend upon long-term memory resources more than phonological loop resources due to the supraspan list length and to the semantic properties of the stimuli.

The average backward score on spatial span was less than the forward score, a pattern similar to that observed in digit span. This appears to be in contrast to past research findings which have shown equal or better performance on spatial span backward compared to forward (Mammarella & Cornoldi, 2005; Szmalec et al., 2005; Vandierendonck et al., 2004; Wilde & Strauss, 2002). However, inspection of scores at the individual level revealed that a substantially larger number of cases had equal or higher performance on backward compared to forward trials for spatial span (69%) than for digit span (17%), which follows the pattern reported by Wilde and colleagues (2004).

The results may provide support for the increasingly numerous research findings suggesting that the spatial and digit span tasks are not as closely analogous as generally assumed (Giggey et al., 2006; Hester et al., 2004; Mammarella & Cornoldi, 2005; Mertens et al., 2006; Smyth & Scholey, 1992; Wilde et al., 2004). This may be due to reliance upon different combinations and weightings of working memory system components or to differing presentation and recall requirements that have been noted to exist between the two tasks (Farrand & Jones, 1996; Wilde & Strauss, 2002).

*Future Research*

A number of further studies are recommended to expand the utility of RSS as a within-test measure of performance validity for use in clinical neuropsychological practice. It would be beneficial for future research to examine RSS in combination with other within-test measures of performance validity to determine how certain a clinician can be that a client is performing in an invalid or a valid manner based on the simultaneous consideration of RSS and several other scores. A within-test measure is never used alone to determine how valid a performance profile is, and therefore such research could expand the utility of the RSS for use in such classification. It would also be of importance to determine if RSS classification accuracy is able to be replicated using initial grouping criteria that is based on a combination of subjective and objective grouping criteria, since it is generally recommended that both be used in making decisions about the validity of a client's profile (Ruff et al., 1993). Further, examination of RSS in a larger sample would indicate how stable the recommended cut-off scores and their associated sensitivity and specificity values are. Another analysis that would lend power to RSS as a classification technique would be to show that a very low percentage of a sample of documented moderate to severe TBI score below the recommended RSS cut-off score. It could also be beneficial to extend the findings using less stringent inclusion criteria such as a wider range of ages, substance abuse history, and other neurological groups so RSS would have greater utility in general clinical settings. Lastly, the relative accuracy of classification of RSS and RDS should be examined simultaneously in additional samples to see whether the pattern found between RSS and RDS in the present study is replicated.

Appendix A: Means/SD/F or Frequency/Percentage/ $\chi^2$ for Demographic Variables by Group

| | Valid | Suspect | Invalid | F / $\chi^2$ |
|---|---|---|---|---|
| Age | 39.41(SD 11.83) | 38.80(SD 11.24) | 38.57 (SD 8.46) | 0.051 |
| Education | 12.37 y (SD 2.39) | 12.71 y (SD 1.90) | 12.78 y (SD 2.02) | 0.306 |
| Time since injury | 803.82 d (SD 671.85) | 779.60 d (SD 678.03) | 821.33 d (SD 528.13) | 0.028 |
| Sex | | | | 1.059[a] |
| Male | 11 (37.9%) | 6 (28.6%) | 14 (42.4%) | |
| Female | 18 (62.1%) | 15 (71.4%) | 19 (57.6%) | |
| Race | | | | 2.081[ab] |
| Caucasian | 19 (65.5%) | 14 (66.7%) | 16 (48.5%) | |
| A. American | 10 (34.5%) | 6 (28.6%) | 15 (45.5%) | |
| Other | 0 | 1 (4.8%) | 2 (6.0%) | |
| Handedness | | | | |
| Right | 27 (93.1%) | 19 (90.5%) | 30 (90.9%) | |
| Left | 1 (3.4%) | 1 (4.8%) | 3 (9.1%) | |
| Ambidextrous | 1 (3.4%) | 1 (4.8%) | 0 | |
| Injury type | | | | |
| MVA-in car | 26 (89.7%) | 18 (85.7%) | 32 (97.0%) | |
| MVA-pedestrian | 2 (6.9%) | 2 (9.5%) | 0 | |

| | | | |
|---|---|---|---|
| Other | 1 (3.4%) | 1 (4.8%) | 1 (3.0%) |
| Referral | | | |
| Litigation | 29 (100%) | 17 (81.0%) | 33 (100%) |
| Clinical | 0 | 4 (19.0%) | 0 |
| LOC | | | |
| None | 14 (48.3%) | 10 (47.6%) | 16 (48.5%) |
| < 5 min | 7 (24.1%) | 4 (19.0%) | 8 (24.2%) |
| < 30 min | 2 (6.9%) | 3 (14.3%) | 2 (6.1%) |
| In & out | 0 | 0 | 1 (3.0%) |
| Not in chart | 6(20.7%) | 4 (19.0%) | 6 (18.2%) |
| PTA | | | |
| None | 17 (58.6%) | 11 (52.4%) | 15 (45.5%) |
| Brief | 5(17.2%) | 5 (23.8%) | 7 (21.2%) |
| Patchy | 5(17.2%) | 2 (9.5%) | 1 (3.0%) |
| Not in chart | 2 (6.9%) | 3 (10.3%) | 7 (21.2%) |
| GCS | | | |
| 14-15 | 2 (6.9%) | 1 (4.8%) | 4 (12.1%) |
| Not in chart | 27 (93.1%) | 20 (95.2%) | 29 (87.9%) |

Note: SD = standard deviation, F = F-ratio from ANOVA, $\chi^2$ = Chi-Square, A. American = African American, y = year, d = day, MVA = motor vehicle accident, LOC = loss of consciousness, PTA = post-traumatic amnesia, GCS = Glasgow Coma Scale, [a] Chi-Square analysis, [b] "other" group removed to allow Chi-Square analysis

Appendix B: Spatial Span Alternatives - Mean (SD), F-ratio, Effect Size

| Spatial Span Alternative | Valid | Suspect | Invalid | F | Effect size $omega^2$ |
|---|---|---|---|---|---|
| RSS | 8.75 (2.33) | 7.19 (1.96) | 6.54 (2.35) | 7.633*** | .137 |
| SSf(2) | 4.51 (1.12) | 3.95 (1.11) | 3.36 (1.11) | 8.246*** | .148 |
| SSb(2) | 4.24 (1.40) | 3.23 (1.13) | 3.18 (1.53) | 5.219** | .092 |
| SSf(raw) | 5.24 (1.21) | 4.90 (0.88) | 4.27 (1.00) | 6.685** | .120 |
| SSb(raw) | 4.86 (1.30) | 4.38 (1.02) | 4.33 (1.29) | 1.623 | - |

** $p < .01$, *** $p < .001$

Appendix C: Spatial Span Alternatives – Tukey Post-hoc Analysis Results

| | Mean1-Mean2 | Significance |
|---|---|---|
| RSS | | |
| Valid vs. Invalid | 2.21 | .001** |
| Valid vs. Suspect | 1.56 | .056 |
| Suspect vs. Invalid | 0.64 | .565 |
| SSf(2) | | |
| Valid vs. Invalid | 1.15 | .000*** |
| Valid vs. Suspect | 0.56 | .188 |
| Suspect vs. Invalid | 0.58 | .149 |
| SSb(2) | | |
| Valid vs. Invalid | 1.05 | .011* |
| Valid vs. Suspect | 1.00 | .054 |
| Suspect vs. Invalid | 0.05 | .989 |
| SSf(raw) | | |
| Valid vs. Invalid | 0.96 | .002** |
| Valid vs. Suspect | 0.33 | .511 |
| Suspect vs. Invalid | 0.63 | .089 |

*p < .05, **p <.01, ***p < .001; RSS = reliable spatial span, SSf(2) = spatial span forward both trials correct, SSb(2) = spatial span backward both trials correct, SSf(raw) = spatial span forward one trial correct

Appendix D: Frequency, Cumulative Frequency, and Cumulative Percentage of Subjects Scoring at a RSS Score for the Valid and Invalid Groups

| RSS cut-off score | Frequency (Valid) | Cumulative Frequency (Valid) | Cumulative % | Frequency (Invalid) | Cumulative Frequency (Invalid) | Cumulative % |
|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 1 | 1 | 3.0 |
| 1 | 0 | 0 | 0 | 0 | 1 | 3.0 |
| 2 | 1 | 1 | 3.4 | 0 | 1 | 3.0 |
| 3 | 0 | 1 | 3.4 | 1 | 2 | 6.1 |
| 4 | 0 | 1 | 3.4 | 4 | 6 | 18.2 |
| 5 | 2 | 3 | 10.3 | 2 | 8 | 24.2 |
| 6 | 1 | 4 | 13.8 | 10 | 18 | 54.5 |
| 7 | 2 | 6 | 20.7 | 5 | 23 | 69.7 |
| 8 | 6 | 12 | 41.4 | 4 | 27 | 81.8 |
| 9 | 7 | 19 | 65.5 | 3 | 30 | 90.9 |
| 10 | 3 | 22 | 75.9 | 1 | 31 | 93.9 |
| 11 | 3 | 25 | 86.2 | 1 | 32 | 97.0 |
| 12 | 4 | 29 | 100 | 1 | 33 | 100 |

Appendix E: Specificity & Sensitivity for the Significant Spatial Span Alternatives

| Cut-off Score | Specificity | Sensitivity |
|---|---|---|
| RSS | | |
| 4 | 96.6 | 18.2 |
| 5 | 89.7 | 24.2 |
| 6 | 86.2 | 54.5 |
| 7 | 79.3 | 69.7 |
| 8 | 58.6 | 81.8 |
| SSf(2) | | |
| 2 | 96.6 | 15.2 |
| 3 | 79.3 | 57.6 |
| 4 | 55.2 | 87.9 |
| SSb(2) | | |
| 2 | 89.7 | 30.3 |
| 3 | 82.8 | 54.5 |
| 4 | 34.5 | 87.9 |
| SSf(raw) | | |
| 2 | 100 | 3.0 |
| 3 | 89.7 | 21.2 |
| 4 | 72.4 | 60.6 |
| 5 | 51.7 | 87.9 |

RSS = reliable spatial span, SSf(2) = spatial span forward both trials correct, SSb(2) = spatial span backward both trials correct, SSf(raw) = spatial span forward one trial correct

Appendix F: Sensitivity, Specificity, False Positives, Positive Predictive Power and 95% Confidence Intervals by Base Rate for RSS

| Cut-off | Sensitivity | Specificity | False Positives | PPP BR = .2 | BR=.3 | BR=.4 | BR=.5* | BR=.6 |
|---|---|---|---|---|---|---|---|---|
| 6 | 54.5 | 86.2 | 13.8 | .50 | .63 | .72 | .80 | .86 |
| | 36.4-71.9 | 68.3-96.1 | 3.9-31.7 | .36-.70 | .49-.80 | .60-.86 | .69-.90 | .77-.93 |
| 7 | 69.7 | 79.3 | 20.7 | .46 | .59 | .69 | .77 | .83 |
| | 51.3-84.4 | 60.3-92.0 | 8.0-37.7 | .36-.62 | .49-.73 | .60-.81 | .69-.87 | .77-.91 |
| 8 | 81.8 | 58.6 | 41.4 | .33 | .46 | .57 | .66 | .75 |
| | 64.5-93.0 | 38.9-76.5 | 23.5-61.1 | .28-.41 | .39-.54 | .50-.65 | .60-.73 | .70-.80 |

*Base rate from current sample, range is standard error at a 95% confidence interval, PPP = positive predictive power, BR = base rate

Appendix G: Digit Span Alternatives - Mean (SD), F-ratio

| | Valid | Suspect | Invalid | F |
|---|---|---|---|---|
| RDS | 8.41 (2.62) | 7.48 (1.56) | 7.54 (2.04) | 1.950 |
| DSf(2) | 4.90 (1.56) | 4.52 (0.98) | 4.30 (1.33) | 1.515 |
| DSb(2) | 3.52 (1.40) | 2.76 (0.99) | 3.24 (1.09) | 2.470 |
| DSf(raw) | 5.89 (1.75) | 5.71 (1.34) | 5.27 (1.35) | 1.401 |
| DSb(raw) | 4.41 (1.45) | 3.71 (1.10) | 3.69 (1.23) | 2.878 |

Appendix H: Sensitivity, Specificity, False Positives, Positive Predictive Power and 95% Confidence Intervals by Base Rate for RDS

| Cut-off | Sensitivity | Specificity | False Positives | PPP BR = .2 | BR=.3 | BR=.4 | BR=.5* | BR=.6 |
|---|---|---|---|---|---|---|---|---|
| 6 | 27.3 | 79.3 | 20.7 | 0.25 | 0.36 | 0.47 | 0.57 | 0.66 |
| | 13.3-45.5 | 60.3-92.0 | 8.0-39.7 | .22-.29 | .33-.42 | .43-.53 | .53-.62 | .63-.71 |
| 7 | 48.5 | 58.6 | 41.4 | 0.23 | 0.33 | 0.44 | 0.54 | 0.64 |
| | 30.8-66.5 | 38.9-76.5 | 23.5-61.1 | .21-.25 | .32-.36 | .42-.47 | .52-.57 | .62-.66 |
| 8 | 51.5 | 37.9 | 62.1 | 0.17 | 0.26 | 0.36 | 0.45 | 0.55 |
| | 33.5-69.2 | 20.7-57.7 | 42.3-79.3 | .17-.18 | .25-.26 | .35-.37 | .44-.47 | .54-.57 |

*Base rate from current sample, range is standard error at a 95% confidence interval, PPP = positive predictive power, BR = base rate

Appendix I: Correlation Coefficients for Valid Group

| | CVLT trial1 (n=28) | | Arithmetic (n=24) | |
|---|---|---|---|---|
| | r | p | r | p |
| SSf | .474 | .005 | .364 | .040 |
| SSb | .477 | .005 | .482 | .009 |
| DSf | .141 | .237 | .398 | .027 |
| DSb | .542 | .001 | .406 | .024 |
| SSt | .471 | .006 | .453 | .013 |
| DSt | .389 | .020 | .441 | .016 |

SSf = spatial span forward one correct, SSb = spatial span backward one correct, DSf = digit span forward one correct, DSb = digit span backward one correct, SSt = spatial span total score, DSt = digit span total score

Appendix J: Mean Scores and Standard Deviations for Subjects Included in Correlational Analyses

| Task | Mean (SD) |
| --- | --- |
| CVLT-II trial 1 | 6.17 (2.14) |
| Arithmetic | 7.83 (3.01) |
| SSf | 5.24 (1.21) |
| SSb | 4.86 (1.30) |
| SSt | 10.10 (2.36) |
| DSf | 5.89 (1.75) |
| DSb | 4.41 (1.45) |
| DSt | 10.31 (2.82) |

SSf = spatial span forward one correct, SSb = spatial span backward one correct, DSf = digit span forward one correct, DSb = digit span backward one correct, SSt = spatial span total score, DSt = digit span total score

Appendix K: Group Criteria for Past RDS Studies with Clinical TBI Samples

| Study | Invalid group | Valid group | Mild TBI criteria |
|---|---|---|---|
| Greiffenstein et al. (1994) | PPCS and meet 2+ of following:<br>- 2+ impairment ratings of severe on NP tests<br>- Improbable symptom history contradicted by records or surveillance film<br>- Total disability in work or a major social role after 1 yr<br>- Claims of remote memory loss | PPCS and not meeting criteria for probable malingering | - PPCS 1 yr post-injury<br>- PTA≤20 mins<br>- ER GCS=15<br>- Hospital stay ≤48 hrs<br>- Normal CT & neurological exam results |
| Greiffenstein et al. (1995) | Same as Greiffenstein et al. (1994) | PPCS at least 1 yr post-TBI, not meeting probable malingering criteria | - PPCS 1yr post-injury<br>- PTA ≤20mins |
| Mathias et al. (2002) | Slick criteria (probable MND = A plus 2 B or 1 B & 1+ C)<br>A. presence of substantial external incentive | No external incentive, passed PDRT & TOMM, not | Irrelevant because mild and moderate-severe TBI groups |

| | | | |
|---|---|---|---|
| | B. evidence from NP testing (failed effort test, discrepancy between test data & documented history, known patterns of brain function, behavioural observations, or info from collaterals)<br>C. evidence from self-report (Discrepancy between self-report & documented history, known patterns of brain function, behavioural observations, or info from collaterals)<br>D. B&C not accounted for by psychiatric, neurologic, or developmental factors | meeting Slick criteria | examined together |
| Heinly (2005) | Slick et al. (1999) criteria used<br>- Effort test criteria: fail TOMM or PDRT, or 2+ of 3 within-test measures | Not meeting Slick criteria for malingering | - PTA≤24hrs<br>- GCS 13-15 after 30mins<br>- LOC≤30mins |

| | | | |
|---|---|---|---|
| | (CVLT Millis formula, WCST Suhr formula, WCST unique responses) | | - No +ve neuroradiological or focal neurological signs |
| Babikian et al. (2006) | Criteria of suspect effort:<br>- Failure on 2 within-test measures of performance validity (Rey Dot Counting Test, Rey Word Recognition, Harbour-UCLA b Test, Warrington Recognition Memory Test – Words, RAVLT, Rey-15*)<br>- Met at least one of Slick's behavioural criteria | Not meeting suspect effort criteria | Irrelevant because no mention of TBI severity |

PPCS = persistent post-concussive syndrome, NP = neuropsychological, PTA = post-traumatic amnesia, GCS = Glasgow Coma Scale, ER = emergency room, LOC = loss of consciousness, CT = computerized tomography, MND = malingered neurocognitive dysfunction, 1+ = one or more, *not a within-test measure

## References

Alexander, M.P. (1995). Mild traumatic brain injury: Pathophysiology, natural history, and clinical management. *Neurology, 45,* 1253-1260.

Allen, L., Conder, R.L., Green, P., & Cox, D.R. (1997). *CARB' 97 Manual for the Computerized Assessment of Response Bias.* Durham, NC: CogniSyst.

American Congress of Rehabilitation Medicine Head Injury Interdisciplinary Special Interest Group. (1993). Definition of mild traumatic brain injury. *Journal of Head Trauma Rehabilitation, 8*, 86–87.

Babikian, T., Boone, K.B., Lu, P., & Arnold, G. (2006). Sensitivity and specificity of various digit span scores in the detection of suspect effort. *The Clinical Neuropsychologist, 20,* 145-159.

Baddeley, A.D. (1996). The fractionation of working memory. *Proceedings of the National Academy of Sciences of the United States of America, 93,* 13468-13472.

Baddeley, A.D. (2001). Is working memory still working? *American Psychologist, 56,* 851-864.

Bate, A.J., Mathias, J.L., & Crawford, J.R. (2001). Performance on the Test of Everyday Attention and standard tests of attention following severe traumatic brain injury. *The Clinical Neuropsychologist, 15*, 405-422.

Bauer, L., & McCaffrey, R.J. (2006). Coverage of the Test of Memory Malingering, Victoria Symptom Validity Test, and Word Memory Test on the internet: Is test security threatened? *Archives of Clinical Neuropsychology, 21,* 121-126.

Benton, A.L., Sivan, A.B., Hamsher, K.deS., Varney, N.R., & Spreen, O. (1983). *Contributions to neuropsychological assessment.* Orlando, FL: Psychological Assessment Resources.

Bernard, L.C. (1990). Prospects for faking believable memory deficits on neuropsychological tests and the use of incentives in simulation research. *Journal of Clinical and Experimental Neuropsychology, 12,* 715-728.

Binder, L.M. (1993). Assessment of malingering after mild head trauma with the Portland Digit Recognition Test. *Journal of Clinical and Experimental Neuropsychology, 15*, 170-182.

Binder, L.M. & Willis, S.C. (1991). Assessment of motivation after financially compensable minor head trauma. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3,* 175-181.

Bush, S.S., Ruff, R.M., Troster, A.I., Barth, J.T., Koffler, S.P., Pliskin, N.H., Reynolds, C.R., & Silver, C.H. (2005). Symptom validity assessment: Practice issues and medical necessity. NAN policy & planning committee. *Archives of Clinical Neuropsychology, 20*:419-426.

Chuah, Y.M.L., Maybery, M.T., & Fox, A.M. (2004). The long-term effects of mild head injury on short-term memory for visual form, spatial location, and their conjunction in well-functioning university students. *Brain and Cognition, 56*, 304-312.

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

Constantinou, M., Bauer, L., Ashendorf, L., Fisher, J.M., & McCaffrey, R.J. (2005). Is poor performance on recognition memory effort measures indicative of generalized poor performance on neuropsychological tests? *Archives of Clinical Neuropsychology, 20*, 191-198.

Darke, S. (1988). Anxiety and working memory capacity. *Cognition and Emotion, 2*, 145-154.

Delis, D.C., Kramer, J.H., Kaplan, E., & Ober, B.A. (2000). *California Verbal Learning Test – Second edition.* San Antonio, TX: The Psychological Corporation.

DeStefano, D., & LeFevre, J. (2004). The role of working memory in mental arithmetic. *European Journal of Cognitive Psychology, 16*, 353-386.

Drane, D.L., Williamson, D.J., Stroup, E.S., Holmes, M.D., Jung, M., Koerner, E., Chaytor, N., Wilensky, A.J., & Miller, J.W. (2006). Cognitive impairment is not equal in patients with epileptic and psychogenic nonepileptic seizures. *Epilepsia, 47*, 1879-1886.

Duncan, S.A., & Ausborn, D.L. (2002). The use of Reliable Digits to detect malingering in a criminal forensic pretrial population. *Assessment, 9*, 56-61.

Echemendia, R.J., Putukian, M., Mackin, R.S., Julian, L., & Shoss, N. (2001). Neuropsychological test performance prior to and following sports-related mild traumatic brain injury. *Clinical Journal of Sports Medicine, 11*, 23-31.

Essig, S.M., Mittenberg, W., Petersen, R.S., Strauman, S., & Cooper, J.R. (2001). Practices in forensic neuropsychology: Perspectives of neuropsychologists and trial attorneys. *Archives of Clinical Neuropsychology, 16*, 271-291.

Etherton, J.L., Bianchini, K.J., Greve, K.W., & Heinly, M.T. (2005). Sensitivity and specificity of Reliable Digit Span in malingered pain-related disability. *Assessment, 12,* 130-136.

Farrand, P. & Jones, D. (1996). Direction of report in spatial and verbal serial short-term memory. *The Quarterly Journal of Experimental Psychology, 49*, 149-158.

Faust, D. & Ackley, M.A. (1998). Did you think it was going to be easy? Some methodological suggestions for the investigation and development of malingering detection techniques. In C.R. Reynolds (Eds.), *Detection of malingering during head injury litigation* (pp. 1-54). New York: Plenum Press.

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: SAGE Publications.

Frederick, R.I., Crosby, R.D., & Wynkoop, T.F. (2000). Performance curve classification of invalid responding on the validity indicator profile. *Archives of Clinical Neuropsychology, 15,* 281-300.

Giggey, P.P., Spencer, R.J., Rice, S., Katzel, L.I., & Waldstein, S.R. (2006). Digit, word, and visual spans are not interchangeable measures in healthy older adults [Abstract]. *Archives of Clinical Neuropsychology, 21*, 574.

Gorny, I. & Merten, T. (2005). Symptom information- warning- coaching: How do they affect successful feigning in neuropsychological assessment? *Journal of Forensic Neuropsychology, 4,* 71-97.

Green, P. (2005, October). Your choice of SVTs is fundamental to the Slick et al. criteria. Lecture presented at the National Academy of Neuropsychology annual meeting, Tampa, FL. Retrieved June 25, 2007, from http://www.wordmemorytest.com.

Green, P. (2006). The pervasive influence of effort on neuropsychological tests. *International Journal of Forensic Psychology, 1,* 1-21.

Green, P., Allen, L.M., & Astner, K. (1996). *The Word Memory Test: A user's guide to the oral and computer-administered forms.* Durham, NC: CogniSyst, Inc.

Green, P., Iverson, G.L., & Allen, L. (1999). Detecting malingering in head injury litigation with the Word Memory Test. *Brain Injury, 13,* 813-819.

Green, P., Rohling, M.L., Lees-Haley, P.R., & Allen, L.M. (2001). Effort has a greater effect on test scores than severe brain injury in compensation claimants. *Brain Injury. 15,* 1045-1060.

Greene, R.L., Brown, R.C., & Kovan, R.E. (1998). *MMPI-2 adult interpretive system professional manual.* Lutz, FL: Psychological Assessment Resources.

Gregoire, J., & Van der Linden, M. (1997). Effects of age on forward and backward digit spans. *Aging, Neuropsychology, and Cognition, 4,* 140-149.

Greiffenstein, M.F., Baker, W.J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment, 6,* 218-224.

Greiffenstein, M.F, Gola, T., & Baker, W.J. (1995). MMPI-2 validity scales versus domain specific measures in detection of factitious traumatic brain injury. *The Clinical Neuropsychologist, 9,* 230-240.

Greve, K.W., & Bianchini, K.J. (2004). Setting empirical cut-offs on psychometric indicators of negative response bias: A methodological commentary with recommendations. *Archives of Clinical Neuropsychology, 19,* 533-541.

Greve, K.W., Bianchini, K.J., Mathias, C.W., Houston, R.J., & Crouch, J.A. (2002). Detecting malingered performance with the Wisconsin Card Sorting Test: A preliminary investigation in traumatic brain injury. *The Clinical Neuropsychologist, 16*, 179-191.

Handley, S.J., Capon, A., Copp, C., & Harper, C. (2002). Conditional reasoning and the Tower of Hanoi: The role of spatial and verbal working memory. *British Journal of Psychology, 93*, 501-518.

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

Heinley, M.T., Greve, K.W., Bianchini, K.J., Love, J.M., & Brennan, A. (2005). WAIS digit span-based indicators of malingered neurocognitive dysfunction: Classification accuracy in traumatic brain injury. *Assessment, 12,* 429-444.

Hester, R.L., Kinsella, G.J., & Ong, B. (2004). Effect of age on forward and backward span tasks. *Journal of International Neuropsychological Society, 10*, 475-481.

Hux, K., Schram, C.D., & Goeken, T. (2006). Misconceptions about brain injury: A survey replication study. *Brain Injury, 20,* 547-553.

Imbo, I., Vandierendonck, A., & De Rammalaere, S. (2007). The role of working memory in the carry operation of mental arithmetic: Number and value of the carry. *The Quarterly Journal of Experimental Psychology, 60*, 708-731.

Immediate Post-concussion Assessment and Cognitive Testing (2006). ImPACT: The best approach to concussion management. Retrieved June 22, 2007, from http://www.impacttest.com.

Inman, T.H. & Berry, D.T.R. (2002). Cross-validation of indicators of malingering: A comparison of nine neuropsychological tests, four tests of malingering, and behavioural observations. *Archives of Clinical Neuropsychology, 17*, 1-23.

Iverson, G.L., & Binder, L.M. (2000). Detecting exaggeration and malingering in neuropsychological assessments. *Journal of Head Trauma Rehabilitation, 15*, 829-858.

Iverson, G.L., Brooks, B.L., Collins, M.W., & Lovell, M.R. (2006). Tracking neuropsychological recovery following concussion in sport. *Brain Injury, 20*, 245-252.

Kirk, R.E. (1995). *Experimental design: Procedures for the behavioural sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Langeluddecke, P.M. & Lucas, S.K. (2003). Quantitative measures of memory malingering on the Wechsler Memory Scale – Third edition in mild head injury litigants. *Archives of Clinical Neuropsychology, 18*, 181-197.

Larrabee, G.J., & Curtiss, G. (1995). Construct validity of various verbal and visual memory tests. *Journal of Clinical and Experimental Neuropsychology, 17*, 536-547.

Mackinnon, A.J. (2000). DAG_Stat [Computer program]. Retrieved from http://www.mhri.edu.au/biostats/DAG_Stat

Mammarella, I.C. & Cornoldi, C. (2005). Sequence and space: The critical role of a backward spatial span in the working memory deficit of visuospatial learning disabled children. *Cognitive Neuropsychology, 22,* 1055-1068.

Mathias, C.W., Greve, K.W., Bianchini, K.J., Houston, R.J., & Crouch, J.A. (2002). Detecting malingered neurocognitive dysfunction using the Reliable Digit Span in traumatic brain injury, *Assessement, 9*, 301-308.

Matthews, C.G. & Klove, K. (1964). *Instruction manual for the adult neuropsychology test battery.* Madison, WI: University of Wisconsin Medical School.

McHugh, R., Laforce, R., Gallagher, P., Quinn, S., Diggle, P., & Buchanan, L. (2006). Natural history of the long-term cognitive, affective, and physical sequelae of mild traumatic brain injury [Abstract]. *Brain and Cognition, 60,* 209-211.

Mertens, V.B., Gagnon, M., Coulombe, D., & Messier, C. (2006). Exploratory factor analysis of neuropsychological tests and their relationship to the Brown-Peterson task. *Archives of Clinical Neuropsychology, 21,* 733-739.

Meyers, J.E., Morrison, A.L., & Miller, J.C. (2001). How low is too low, revisited: Sentence repetition and AVLT recognition in the detection of malingering. *Applied Neuropsychology, 8,* 234-241.

Meyers, J.E. & Volbrecht, M. (1998). Validation of Reliable Digits for detection of malingering. *Assessment, 5,* 301-305.

Meyers, J.E. & Volbrecht, M.E. (2003). A validation of multiple malingering detection methods in a large clinical sample. *Archives of Clinical Neuropsychology, 18,* 261-276.

Millis, S.R. (1992). Recognition Memory Test in the detection of malingered and exaggerated memory deficits. *The Clinical Neuropsychologist, 7*, 406-414.

Millis, S.R., Putnam, S.H., Adams, K.H., & Ricker, J.H. (1995). The California Verbal Learning Test in the detection of incomplete effort in neuropsychological testing. *Psychological Assessment, 7,* 463-471.

Moore, B.A. & Donders, J. (2004). Predictors of invalid neuropsychological test performance after traumatic brain injury. *Brain Injury, 18,* 975-984.

Morey, L.C. (1996). *Personality Assessment Inventory professional manual.* Odessa, FL: Psychological Assessment Resources, Inc.

Moss, A., Jones, C., Fokias, D., & Quinn, D. (2003). The mediating effects of effort upon the relationship between head injury severity and cognitive functioning, *Brain Injury, 17,* 377-387.

O'Bryant, S.E., Duff, K., Fisher, J., & McCaffrey, R.J. (2004). Performance profiles and cut-off scores on the Memory Assessment Scales. *Archives of Clinical Neuropsychology, 19,* 489-496.

Psychological Corporation (The) (1997). *WAIS-III-WMS-III technical manual.* San Antonio, TX: Author.

Psychological Corporation (The) (1999). *Wechsler Abbreviated Scale of Intelligence (WASI) manual.* San Antonio, TX: Author.

Reitan, R.M., & Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test Battery: Theory and interpretation.* Tucson, AZ: Neuropsychology Press.

Reitan, R.M., & Wolfson, D. (2000). *Mild head injury: Intellectual, cognitive, and emotional consequences.* Tucson, AZ: Neuropsychology Press.

Rogers, R., Harrell, E.H., & Liff, C.D. (1993). Feigning neuropsychological impairment: A critical review of methodological and clinical considerations. *Clinical Psychology Review, 13*, 255-274.

Ross, S.R., Putnam, S.H., & Adams, K.M. (2006). Psychological disturbance, incomplete effort, and compensation-seeking status as predictors of neuropsychological test performance in head injury. *Journal of Clinical and Experimental Neuropsychology, 28*, 111-125.

Ruff, R.M. (1988). *Ruff Figural Fluency Test professional manual.* Odessa, FL: Psychological Assessment Resources, Inc.

Ruff, R.M., & Allen, C.C. (1996). *Ruff 2 & 7 Selective Attention Test professional manual.* Odessa, FL: Psychological Assessment Resources, Inc.

Ruff, R.M., Wylie, T., & Tennant, W. (1993). Malingering and malingering-like aspects of mild closed head injury. *Journal of Head Trauma Rehabilitation, 8*, 60-73.

Slick, D., Hopp, G., Strauss, E., & Thompson, G.B. (1997). *The Victoria Symptom Validity Test: Version 1.0.* Odessa, FL: Psychological Assessment Resources.

Slick, D.J., Sherman, E.M.S., & Iverson, G.L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist, 13,* 545-561.

Smyth, M.M., & Scholey, K.A. (1992). Determining spatial span: The role of movement time and articulation rate. *Quarterly Journal of Experimental Psychology, 45A*, 479-501.

Smyth, M.M., & Scholey, K.A. (1996). Serial order in spatial immediate memory. *Quarterly Journal of Experimental Psychology, 49A,* 159-177.

Strauss, E., Slick, D.J., Levy-Bencheton, J., Hunter, M., MacDonald, S.W.S., & Hultsch, D.F. (2002). Intradindividual variability as an indicator of malingering in head injury. *Archives of Clinical Neuropsychology, 17*, 423-444.

Suhr, J.A., & Gunstad, J. (2000). The effect of coaching on the sensitivity and specificity of malingering measures. *Archives of Clinical Neuropsychology, 15,* 415-424.

Szmalec, A., Vandierendonck, A., & Kemps, E. (2005). Response selection involves executive control: Evidence from the selective interference paradigm. *Memory & Cognition, 33*, 531-541.

Tabachnick, B.G., & Fidell, L.S. (2001). *Using multivariate statistics* ($4^{th}$ ed.). Needham Heights, MA: Allyn and Bacon:

Tombaugh, T. (1996). *Test of Memory Malingering manual.* New York: MultiHealth Systems.

Tubi, N. & Calev, A. (1989). Verbal and visuospatial recall by younger and older subjects: Use of matched tasks. *Psychology and Aging, 4*, 493-495.

Vanderploeg, R.D., Curtiss, G., & Belanger, H.G. (2005). Long-term neuropsychological outcomes following mild traumatic brain injury. *Journal of the International Neuropsychological Society, 11,* 228-236.

Vandierendonck, A., De Vooght, G., & Van der Goten, K. (1998). Does random time interval generation interfere with working-memory executive functions? *European Journal of Cognitive Psychology, 10,* 413-442.

Vandierendonck, A., Kemps, E., Fastame, M.C., & Szmalec, A. (2004). Working memory components of the Corsi blocks task. *British Journal of Psychology, 95,* 57-79.

Warrington, E.K. (1984). *Recognition Memory Test manual.* Berkshire, United Kingdom: NFER-Nelson.

Wechsler, D.A. (1997). *Wechsler Memory Scale-III.* New York: Psychological Corporation.

Wetter, M.W. & Corrigan, S.K. (1995). Providing information to clients about psychological tests: A survey of attorney's and law students' attitudes. *Professional Psychology: Research and Practice*, 26, 474-477.

Wilde, N. & Strauss, E. (2002). Functional equivalence of WAIS-III/WMS-III digit and spatial span, under forward and backward recall conditions. *Clinical Neuropsychologist, 16*, 322-330.

Wilde, N.J., Strauss, E., & Tulsky, D.S. (2004). Memory span on the Wechsler scales. *Journal of Clinical and Experimental Neuropsychology, 26*, 539-549.

Wilkinson, G.S. (1993). *Wide Range Achievement Test 3.* Wilmington, DE: Wide Range Inc.

Youngjohn, J.R. (1995). Confirmed attorney coaching prior to neuropsychological evaluation. *Assessment, 2*, 279-283.

Vita Auctoris


| | |
|---|---|
| Name: | Shelley Ylioja |
| Place of Birth: | Outlook, Saskatchewan |
| Year of Birth: | 1980 |
| Education: | Lucky Lake High School, Lucky Lake, Saskatchewan<br>1992-1998 |
| | University of Regina, Regina, Saskatchewan<br>1998-2002 B.A.(hons) |
| | University of Windsor, Windsor, Ontario<br>2005-2007 M.A. |