Developing a Spatial Analogue of the Reliable Digit Span

Shelley G. Ylioja\textsuperscript{a}, Anne D. Baird\textsuperscript{a,b}, Kenneth Podell\textsuperscript{a,b,*}

\textsuperscript{a}Department of Psychology, University of Windsor, Windsor, ON, Canada
\textsuperscript{b}Division of Neuropsychology, Henry Ford Health System, Detroit, MI, USA

Accepted 19 September 2009

Abstract

It is recommended that performance validity be assessed in all neuropsychological cases involving external incentive. The present study sought to develop an embedded performance validity measure based on the Spatial Span task of the Wechsler Memory Scale-III in a sample of litigating persistent postconcussion complainants. The Reliable Spatial Span (RSS) calculation had specificity, sensitivity, and predictive power values within the range of other embedded measures. This finding suggests that RSS is able to distinguish between persistent post-concussion complainants demonstrating valid and invalid performance. Other calculations involving Spatial Span scores had lower classification accuracy. Reliable Digit Span (RDS) classification accuracy within the present sample was lower than that of previous research, as well as of RSS. Potential reasons for lack of RDS replication are discussed, along with the potential use of RSS as an embedded validity performance indicator.

Keywords: Symptom validity testing; Traumatic brain injury; Forensic neuropsychology

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, it is currently acknowledged that incentive to perform well or to perform poorly has a large influence on a test profile (Green, Rohling, Lees-Haley, & Allen, 2001). Failure to exert optimal effort compromises test results and gives an inaccurate picture of true cognitive abilities. Thus, without assessing for effort one cannot have confidence that the data accurately reflect the individual’s optimal performance.

It is, therefore, currently recommended that validity of performance be routinely assessed in neuropsychological evaluations (Bush et al., 2005). Performance validity is especially important to determine in cases of mild traumatic brain injury (TBI) in which there is external motivation for poor performance (financial and work; Binder & Willis, 1991) and no objective medical evidence of brain injury (Inman & Berry, 2002). Given the large influence of effort on test scores and the sometimes subtle nature of cognitive changes in these cases, it is necessary to assess performance validity carefully to get an accurate measurement of neuropsychological performance. Otherwise, legal injustice, improper diagnosis and treatment, and poor outcome may follow.

A number of studies have found compensation-seeking status alone to be a significant variable in test performance (Binder & Willis, 1991; Green, Iverson, & Allen, 1999; Meyers & Volbrecht, 1998; Moore & Donders, 2004). Within the population of individuals who claim mild head trauma and who seek compensation through litigation or other means, estimates of suboptimal effort range from 20% to 60% (Binder, 1993; Constantinou, Bauer, Ashendorf, Fisher, & McCaffrey, 2005; Greiffenstein, Baker, & Gola, 1994; Langeluddecke & Lucas, 2003; Meyers & Volbrecht, 1998).
The present study explored the utility of using the Wechsler Memory Scale-III (WMS-III; Wechsler, 1997) Spatial Span subtest as an embedded measure of performance validity. The usefulness of various scores derived from Spatial Span was explored within a sample of litigants with persistent postconcussion complaints. The structure of the test was preserved; therefore, the role of Spatial Span as a measure of cognitive function was not compromised.

An embedded measure of performance validity is an alternative to using tests specifically developed to assess effort rather than a domain of cognitive ability. There are several possible benefits of using performance validity measures that are built into commonly used neuropsychological tests. First, it is more time efficient if a test can play a dual role in assessing cognitive function as well as performance validity (Langeluddecke & Lucas, 2003; Meyers & Volbrecht, 2003). Second, such measures work as validity checks of the integrity of the results throughout the entire assessment if many of the neuropsychological tests being administered have also been developed into validity measures (Meyers & Volbrecht, 1998). Third, 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, Gola, & Baker, 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. Fourth, embedded validity measures may also be less susceptible to intentional poor effort by clients who have been made aware by their attorneys of tests designed to detect malingering (Mathias, Greve, Bianchini, Houston, & Crouch, 2002) or who have found such information on the internet (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 escape detection by altering responding during these measures (Suhr & Gunstad, 2000).

Although the Digit Span task has received substantial examination as an embedded measure of performance validity, the Spatial Span task has received less scrutiny for this purpose. Reliable Digit Span (RDS) is a calculation of the sum of the highest number of digits forward where both trials are correct plus the highest number of digits backward on which both trials are correct (Greiffenstein et al., 1994). RDS has been cited as one of the best-validated embedded clinical measures of effort (Heinly, Greve, Bianchini, Love, & Brennan, 2005) and has been tested in clinical cases of TBI (Greiffenstein et al., 1994, 1995; Heinly et al., 2005; Mathias et al., 2002) as well as individuals asked to simulate TBI (Inman & Berry, 2002; Strauss et al., 2002). Classification accuracy is lower in cases of documented moderate to severe TBI compared with 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. Other validation samples include patients with medical and/or psychiatric disorders (Babikian, Boone, Lu, & Arnold, 2006), persons reporting chronic pain (Etherton, Bianchini, Greve, & Heinly, 2005), as well as participants in a forensic sample undergoing pre-trial/pre-sentence assessment (Duncan & Ausborn, 2002). A cut-off score of 7 or lower has been shown to differentiate probable valid and invalid responders with acceptable sensitivity, specificity, and predictive power in all of these studies.

The Spatial Span test, purportedly measuring attention and working memory for visuospatial material, is frequently viewed as a non-verbal counterpart to the Digit Span test (Mammarella & Cornoldi, 2005). Research examining the underlying cognitive processes of visuospatial and verbal span tests has documented that both tasks seem to utilize partially overlapping but separable neuronal networks (the visuospatial sketchpad and phonological loop, respectively, of Baddeley’s model of working memory; e.g., Szmalec, Vandierendonck, & Kemps, 2005) and both measures call on executive functions when longer sequences are presented (Vandierendonck, De Vooght, & Van der Goten, 1998; Vandierendonck, Kemps, Fastame, & Szmalec, 2004). In addition, similarities between the span tasks were found by Smyth and Scholey (1996) with respect to serial order and position effects. On the other hand, evidence exists to suggest that the tasks may be less analogous than assumed. For example, recall on Digit Span forward is almost always superior to Digit Span backward, whereas recall on Spatial Span forward (SSf) is equal or inferior to Spatial Span backward (SSb) in a significant minority of individuals (e.g., Hester, Kinsella, & Ong, 2004; Wechsler, 1997; Wilde & Strauss, 2002). In addition, stimulus presentation discrepancy exists because Digit Span forward and backward are comprised of different sequences whereas SSf and SSb sequences are identical (Wilde & Strauss, 2002).

Nevertheless, the tasks seem to be similar enough that given the positive findings regarding the utility of RDS in detecting invalid performance, it was hypothesized in the present study that the non-verbal counterpart of Digit Span would be sensitive to invalid performance and could also act as a simple, yet robust, embedded measure of performance validity.

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 a certain area in which to feign difficulties (Greiffenstein et al., 1995; Iverson & Binder, 2000). It would, therefore, be reasonable to assume that some clients feigning or exaggerating symptoms would do so in the visuospatial domain. Spatial Span could also provide a validity measure to be
used with individuals who have verbal or auditory impairments and are unable to perform the Digit Span task or other performance validity measures that rely on the verbal domain.

In the present study, it was hypothesized that Spatial Span scores would efficiently differentiate an individual’s performance as valid or invalid among a sample of persistent postconcussion complainants seen in the context of civil litigation; thus, Spatial Span could be used as an embedded measure of performance validity.

A second goal of the present study was to replicate earlier research indicating that RDS is a useful technique for classifying performance as valid or invalid (Babikian et al., 2006; Duncan & Ausborn, 2002; Etherton et al., 2005; Greiffenstein et al., 1994, 1995; Heinly et al., 2005; Inman & Berry, 2002; Mathias et al., 2002; Strauss et al., 2002) and compare Spatial Span indices to RDS. Specifically, it was expected that RDS would have acceptable levels of specificity and sensitivity at a cut-off score of 7 in this sample as it has had in other samples and that Spatial Span indices would be equally specific and sensitive.

Materials and Methods

Participants

The sample consisted of 62 (25 men, 37 women) clients with persistent postconcussion complaints evaluated between January 2002 and February 2007 by a neuropsychologist at an outpatient clinic in a large urban healthcare system. All were involved in litigation and most were defense referrals. The majority (n = 60) were involved in a motor vehicle accident (58 as driver or passenger, 2 as pedestrian); two had received a blow to the head by other means (object falling on head). The sample was chosen retrospectively using exclusion criteria outlined later from a larger population of 166 TBI cases consisting of independent medical evaluation (80%) and clinical cases. The following criteria were applied to ensure all participants had no worse than a mild TBI: None or loss of consciousness no longer than 1 h, post-traumatic amnesia extending no more than 1 day, and an on-scene emergency medical treatment or emergency room Glasgow Coma Scale (GCS) score of higher than 13 (American Congress of Rehabilitation, 1993). No GCS score was in the medical record of 90% of participants. In these cases, the participants either received no acute medical care or emergency caregivers presumably failed to administer the GCS. For these participants, a GCS of greater than 13 was assumed. We excluded participants with a history of neurosurgical intervention (e.g., craniotomy), documented seizure disorder, brain cancer, encephalitis, stroke, myocardial infarction, mental retardation, substance abuse, and psychiatric history of bipolar disorder or schizophrenia. Those with a documented history of moderate to severe TBI, those who were clinically referred, and those who reported English as their second language also were excluded. In addition, clients were excluded if they had not been administered all the following tests: Test of Memory Malingering (TOMM; Tombaugh, 1996), Word Memory Test (WMT; Green, Allen, & Astner, 1996), WMS-III Spatial Span, and WMS-III/WAIS-III Digit Span.

On the basis of their WMT and TOMM scores, participants were classified as having a valid (passed both, n = 29) or invalid (failed both, n = 33) performance profile. Pass and failure were based on cut-scores outlined in the TOMM and WMT test manuals. Failure on two well-validated performance validity measures has been recommended as being diagnostic of probable malingered neurocognitive dysfunction (Larrabee, Greiffenstein, Greve, & Bianchini, 2007). Those who had passed one and failed the other (n = 19; 79% passed the TOMM) were excluded based on these grouping requirements. This group of clients did not differ significantly from the invalid and valid groups on any demographic variables.

Average age was 38.9 years (SD = 10.1, range 18–55), with an average education level of 12.5 years (SD = 2.1, range 9–20). The participant’s ethnic background was identified as European American in 35 cases and as African American in 25. Two participants were identified as having some other background. Testing was conducted more than 6 months post-injury in 92% (n = 57) of the cases. Four cases were tested between 1.5 and 4.5 months post-injury. Date of injury was missing for one case.

Data for the groups (valid and invalid) were analyzed using an independent t-test or χ² analysis on sex (χ²[1, N = 62] = 0.13, p = .72), age (t[60] = 0.32, p = .75), race (χ²[1, N = 60] = 1.19, p = .28), time since injury (t[60] = −0.11, p = .91), and education (r[60] = −0.73, p = .47). No significant differences were observed between the groups on these variables (Table 1). Full-scale intelligence (FSIQ) was significantly different between the invalid and valid groups (t[59] = 2.22, p = .03). This is likely secondary to invalid performance because years of education and FSIQ show an expected significant relationship in the valid group (r = 0.59, p = .001), but fail to do so in the invalid group (r = 0.25, p = .15).

Measures

Participants completed a comprehensive neuropsychological battery including the WMS-III (Wechsler, 1997), Wechsler Adult Intelligence Scale-III (The Psychological Corporation, 1997) or Wechsler Abbreviated Scale of Intelligence
The measures used in the present study included the TOMM, WMT, Spatial Span subtest of the WMS-III, and Digit Span subtest of the WMS-III or WAIS-III. Measures were administered in a standardized manner as outlined in the instruction manuals. The retention trial of the TOMM was always completed regardless of performance on trials 1 and 2. Order of administration was pseudo-randomized with the TOMM and WMT administered in different parts of the day (e.g., one in the morning and the other in the afternoon).

**Spatial Span indices.** The dependent measures in this study were five indices derived from the raw scores on the WMS-III Spatial Span subtest. These measures included an index called Reliable Spatial Span (RSS), so labeled because it was calculated in a manner identical to RDS by summing the length of the longest spatial span forward and backward correctly repeated twice within the same trial. For example, a client correctly repeating both trials of five forward and one trial of six forward along with both trials of three backward and one trial of four backward would have an RSS score of 8 by adding 5 and 3. The other indices included an RSS forward (longest string forward correctly repeated within the same trial), an RSS backward (longest string backward correctly repeated within the same trial), SSf (longest string forward), and SSb (longest string backward).

**Reliable Digit Span.** RDS was calculated by summing the longest string forward and backward correctly repeated within the same trial, as outlined by Greiffenstein and colleagues (1994).

**Data Analysis and Results**

**WMS-III Spatial Span as a Within-Test Performance Validity Measure**

**Sensitivity, specificity, and cut-off scores.** Sensitivity and specificity calculations were performed using the valid and invalid groups to determine which Spatial Span index has the best utility in detecting invalid performers. These values are presented in Table 2. For each Spatial Span index, 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). An invalid performance profile would be suspected if a person’s score falls below this cut-off value. It has 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 five Spatial Span indices, RSS provides the best compromise between specificity and sensitivity to give relatively high values of both. A cut-off of 6
or less on RSS correctly classifies approximately 55% of invalid performers and misclassifies 14% of valid performers. A cut-off of 7 or less correctly classifies approximately 70% of invalid performers with a slight rise in misclassification of valid performers to 20% (Table 3).

**Table 2.** Specificity and sensitivity for the Spatial Span indices and RDS

<table>
<thead>
<tr>
<th>Cut-off score</th>
<th>RSS</th>
<th>RSSf</th>
<th>RSSb</th>
<th>SSf</th>
<th>SSb</th>
<th>RDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>96.6</td>
<td>96.6</td>
<td>96.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>3.0</td>
<td>6.1</td>
<td>18.2</td>
<td>30.3</td>
<td>21.2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>96.6</td>
<td>97.3</td>
<td>87.9</td>
<td>87.9</td>
<td>87.9</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>15.2</td>
<td>57.6</td>
<td>87.9</td>
<td>34.5</td>
<td>21.2</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>89.7</td>
<td>82.8</td>
<td>34.5</td>
<td>41.4</td>
<td>62.1</td>
<td>0</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>30.3</td>
<td>54.5</td>
<td>34.5</td>
<td>62.1</td>
<td>62.1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

**Table 3.** Sensitivity, specificity, false positives, and PPP by base rate for RSS

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>False positives</th>
<th>PPP BR = .2</th>
<th>BR = .3</th>
<th>BR = .4</th>
<th>BR = .5</th>
<th>BR = .6</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>54.5</td>
<td>86.2</td>
<td>13.8</td>
<td>.50</td>
<td>.63</td>
<td>.72</td>
<td>.80</td>
<td>.86</td>
</tr>
<tr>
<td>7</td>
<td>69.7</td>
<td>79.3</td>
<td>20.7</td>
<td>.46</td>
<td>.59</td>
<td>.69</td>
<td>.77</td>
<td>.83</td>
</tr>
<tr>
<td>8</td>
<td>81.8</td>
<td>58.6</td>
<td>41.4</td>
<td>.33</td>
<td>.46</td>
<td>.57</td>
<td>.66</td>
<td>.75</td>
</tr>
</tbody>
</table>

Notes: RSS = Reliable Spatial Span; RSSf = Reliable Spatial Span forward; RSSb = Reliable Spatial Span backward; SSf = Spatial Span forward one trial correct; SSb = Spatial Span backward one trial correct; RDS = Reliable Digit Span.

or less on RSS correctly classifies approximately 55% of invalid performers and misclassifies 14% of valid performers. A cut-off of 7 or less correctly classifies approximately 70% of invalid performers with a slight rise in misclassification of valid performers to 20% (Table 3).

**Predictive power and base rates.** Positive predictive power (PPP) enables the clinician to determine the probability that a client’s profile is invalid, given the individual’s test score and the base rate in that particular clinic. PPP values depend on the assumed base rate, as well as the specificity and sensitivity of the test in question. Given that the base rate of insufficient effort in samples of litigation evaluations has varied, PPP in the present study was calculated for a range of RSS scores and base rates (Table 3). For example, in a clinic with a 50% base rate of invalid performance among those with persistent postconcussion complaints, a client with an RSS value of 6 can be assumed to be performing in an invalid manner with a probability of 80%. PPP is 50% or higher for even the lowest base rate of 20% at this cut-off. PPP was calculated using the following formula (Slick, 2006): PPP = base rate × sensitivity/(base rate × sensitivity + [(1 − base rate) × (1 − specificity)])

**Incremental validity.** Incremental validity measures to what degree the test in question is able to classify invalid performance above base-rate guessing. This was calculated by subtracting the malingering base rate of the present sample (53%) from the overall hit rate of RSS (true positives + true negatives/sample size). RSS improves base-rate guessing by 16% at a cut-off score of 6 and 21% at a cut-off score of 7.
Sensitivity and specificity were calculated for RDS to allow direct comparison of values to previous research and to RSS. These values are presented in Table 2. The current RDS classification accuracy is lower than RSS and a cut-off score of 7 is not replicated. An RDS score of 7 results in a sensitivity of 48.5 and a specificity of 58.6.

Auxiliary Analyses

**RSS and RDS in cases of inconsistent performance.** We were interested in seeing how the recommended RSS and RDS values would perform in cases in which the client passed one validity measure, usually the TOMM, but failed the other. This might indicate how robust RSS and RDS are in cases where performance validity was inconsistent. The percentage of the group with inconsistent performance that would be classified as valid at RSS of 6 is 73.7, RSS of 7 is 47.4, RDS of 6 is 78.9, and RDS of 7 is 52.6. These classification rates are intermediate to those obtained with the same cut-off scores for the valid and invalid groups.

**RSS and RDS versus Spatial Span and Digit Span Age-Corrected Scaled Scores.** Some research has found similar or higher classification accuracy with Digit Span age-scaled score compared with RDS (Axelrod, Fichtenberg, Millis, & Wertheimer, 2006; Babikian et al., 2006; Greve et al., 2007; Heinly et al., 2005). Therefore, we briefly examined the performance of Digit Span and Spatial Span age-scaled scores in classification of the valid and invalid groups (Table 4).

In line with prior research, Digit Span age-scaled scores had somewhat superior classification than RDS, though rates were much lower than RSS. In contrast, comparison of values in Tables 2 and 4 shows that Spatial Span age-scaled scores generally resulted in lower classification rates than RSS. The one exception to this occurs when comparing the effects of using a cut-off score of 3 on the Spatial Span age-scaled score and a cut-off of 5 for RSS. These cut-off scores result in the same sensitivity and specificity values. Thus, RSS has superior classification rates overall to both Digit Span and Spatial Span age-scaled scores.

Discussion

The main goal of the present study was to explore the utility of the Spatial Span subtest as an embedded measure of performance validity for use with clients presenting with persistent postconcussion complaints. 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 when compared with other embedded measures of validity, 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 an embedded measure of performance validity. This hypothesis was not supported since classification accuracy rates of RDS were 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. However, this finding would need to be replicated.

**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 clients with persistent postconcussion complaints. A cut-off score of 6 or less correctly detects
55% of invalid performers with a low false positive rate of valid performers at 14%. Alternately, a score of 7 or less correctly detects a substantially greater proportion of invalid performers (70%) while still maintaining a relatively low percentage of false positives (20%). These results are both consistent with studies of other embedded measures of performance validity. For example, RDS of 7 or less in previous TBI samples has shown specificity from 68% to 93% and sensitivity from 43% to 89% (Babikian et al., 2006; Greiffenstein et al., 1994, 1995; Heinly et al., 2005; Mathias et al., 2002). In fact, RSS classification accuracy at a score of 7 or less in the present study is almost identical to that of RDS in Heinly and colleagues’ study (2005; specificity 83% and sensitivity 71%). A second example is the Mittenberg discriminant function score indicator from the WAIS which has been shown to have specificity of 81% and sensitivity of 50%–54% in a validation study using a mild TBI sample (Greve, Bianchini, Mathias, Houston, & Crouch, 2003). PPP and incremental validity of RSS in the current sample are also adequate. Other Spatial Span indices examined in the present study failed to show as high classification accuracy as RSS. Therefore, only the RSS alternative is recommended for potential use as an embedded measure of performance validity.

A substantial number of clients failed both the WMT and TOMM yet received a valid classification on the RSS using our suggested cut-off scores. This result may reflect lower sensitivity of the RSS to effort or else variable client effort across tests for diverse reasons, for example, a deliberate choice (Greiffenstein et al., 1995; Iverson & Binder, 2000) or fluctuating arousal.

In addition, some clients passed both validity measures and received an invalid classification with RSS cut-off scores of 7 and 6. This might be a result of fluctuating arousal, cognitive deficits impacting Spatial Span performance, test anxiety affecting working memory (e.g., Darke, 1988), or poor effort on the Spatial Span task.

Of course, it must be reiterated that this initial positive finding is the result of a single study, and further research with RSS would be required for the index to be recommended for clinical use. Also, even with replication, RSS would 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 (Greve & Bianchini, 2004).

Replication of RDS

The present results found RDS to have substantially lower classification accuracy than that of the majority of previous research. 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 adequate at 79%, but this corresponded to a sensitivity of only 27%. Past research has recommended an RDS cut-off score of 7 or lower, which corresponds to a specificity of 59% and a sensitivity of 49% in the current sample. 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).

Methodological differences from previous research may account for the failure of the present study to replicate strongly the RDS classification accuracy noted in prior clinical TBI research. The first, and likely the largest, factor is the set of criteria used to initially classify subjects as invalid or valid responders. Most previous studies used a combination of objective (test scores) and clinical judgment methods to classify subjects, whereas the present study relied solely on objective test scores. Secondly, no previous study used objective validity measures identical to those in the current study. Thirdly, other methodological differences, such as TBI severity and financial incentive, may be relevant. For example, two prior studies (Babikian et al., 2006; Mathias et al., 2002) had a valid group without financial incentive. These factors obviously are important for clinicians to consider.

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 under certain criteria. RSS specificity (80%–86%) and sensitivity (55%–70%) values at scores of 6 and 7 are within the range of values reported for RDS in previous TBI samples of 68%–93% specificity and 43%–89% sensitivity. Of course, such comparisons are subject to the methodology differences just mentioned 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 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 TBI 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 a total span score of 8 or lower (Langeluddecke & Lucas, 2003). Additionally, a study employing simulated malingering using the WMS-Revised (WMS-R) found the difference between simulated
Values of RSS and RDS are significantly correlated ($r = .43, p = .000$) in the present sample. However, upon individual examination of each group, a marked divergence in the correlation emerges with the invalid group’s RSS and RDS correlation improving ($r = .57, p = .001$), whereas the valid group’s correlation declined drastically and was no longer significant ($r = .25, p = .192$). The lack of significant relation in those demonstrating valid performance provides evidence of dissimilarity between the cognitive constructs being measured by the Spatial Span and Digit Span tasks. Agreement in the classification of valid or invalid performance between these methods was 69.4% using an RDS and RSS score of $\leq 7$, and 67.7% using an RDS score of $\leq 7$ and RSS score of $\leq 6$. Agreement was more often present in classifying valid performance.

Given the robustness of RDS as an embedded measure of performance validity, the present study suggests RSS has at least similar 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 (valid and invalid) membership: Two pieces of objective evidence as used in the present study versus one piece each of objective and clinical evidence as used in previous RDS research.

**Generalizability**

The 53% base rate of invalid performance among persistent postconcussion complainants involved in litigation in the current study is very similar to that recently reported by Greiffenstein and Baker (2006) in a sample of chronic whiplash and mild head injury litigants, in which a 57% base rate of probable invalid performance was found. It is also quite similar to the 40% rate reported by Larrabee (2003) in a compilation of 11 studies of performance validity in which secondary gain was present. Thus, the base rate in the current study provides further evidence that the expected base rate of invalid performance among mild TBI litigants is in the range of 40%–60%.

It is important to note that the generalizability of these findings is currently limited. These results may not be applicable to clients outside an age range of 18–55, or clients with a history of substance abuse, other neurological disorders, prior moderate to severe TBI, and diagnosis of schizophrenia, nor are these data applicable to individuals whose first language is other than English. It would be of particular 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.

Classification accuracy of RSS requires replication prior to clinical use. Two elements of future research that would be especially enlightening in comparison to the current study include changes to initial grouping criteria and addition of a non-litigating TBI comparison group. The majority of research on validity tests currently utilizes the diagnostic criteria outlined by Slick, Sherman, and Iverson (1999). However, Larrabee and colleagues (2007) recently recommended that diagnostic criteria for probable malingered neurocognitive dysfunction be refined to include failure on two well-validated symptom validity tests, such as the TOMM and WMT. These authors also stated that the clinical criteria, or those included in the C category in the Slick outline, are the most subjective, as they are very dependent on a user’s neurological knowledge. Thus, we decided to use performance on two well-validated symptom validity measures for initial grouping with the goal of focusing on objective data-driven group classification. This method also has been used recently by others (Blaskewitz, Merten, & Brockhaus, 2009; Victor, Boone, Serpa, Buehler, & Ziegler, 2009). With this grouping strategy, the resulting sensitivity and specificity values represent the degree of agreement between the span measures used in this study and the symptom validity tests used for initial classification of patient protocols. We saw this as a good initial step to show the agreement of RSS with well validated objective measures of effort. Future research can focus on validating RSS using the broader and more commonly used Slick and colleagues’ (1999) set of criteria for malingering. It is suspected that RSS would demonstrate even more promise in classifying patient protocols if this measure were to be compared against the Slick et al.’s criteria rather than the narrower criteria used for patient grouping in the current study.

Also, use of RSS as a performance validity measure would have greater application if it was replicated using a valid group comprised of clinically referred non-litigating cases with moderate to severe well-documented TBI and valid performance on the TOMM and WMT. Comparison of this group with a group like the invalid group in the current study (individuals who have
sustained at most a mild TBI but who have failed the TOMM and WMT) can provide strong evidence of invalid performance if the more severely injured valid group has a significantly higher RSS.

Owing to our decision to use two objective validity tests for initial group classification, the group that failed only one of the WMT and TOMM was removed. This may limit generalizability of our findings due to creation of extreme performance groups. However, since the use of a single validity measure to diagnose invalid performance has not been justified in the literature, it was decided that this group would not be included, especially for an initial exploratory study.

Future research exploring reasons for inconsistent performance may be warranted. For example, Gervais, Rohling, Green, and Ford (2004) examined performance on the TOMM, WMT, and Computerized Assessment of Response Bias as well as performance on a neuropsychological test battery. Individuals failing all three had the lowest neuropsychological test scores, and those failing the WMT but passing the other two validity measures had lower average scores on the neuropsychological tests than those passing all three. Similar to our results, this finding suggests that inconsistent performers may be in the middle of a continuum from valid to invalid performers. Whereas this inconsistent performance group is not typically used in the development of symptom validity tests, change in such a practice may be required in light of the frequency of occurrence.

This initial finding that Spatial Span is able to classify invalid and valid performance with a high level of accuracy shows promise for RSS as an additional embedded performance validity measure. The move toward the development of embedded measures has been propelled by the need to decrease time and money spent on neuropsychological examination (Langeluddecke & Lucas, 2003; Meyers & Volbrecht, 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). Since performance validity measures currently in use are mainly in the verbal modality, they could provide inaccurate results in a client with verbal or auditory impairments, and a client exaggerating or faking non-verbal deficits could avoid detection. In contrast, the results of the present study provide an embedded measure of performance validity in the non-verbal domain, thereby making the findings especially useful.

Funding

This research was supported by a Canada Graduate Scholarship Master’s awarded to the first author by the National Sciences and Engineering Research Council of Canada.

Conflict of Interest

None declared.

References


