Detection of Suboptimal Effort with Symbol Span: Development of a New Embedded Index

J. Christopher Young1,*, Joshua E. Caron2,*, Brandon C. Baughman3, R. John Sawyer1

1Memphis Veterans Affairs Medical Center, Memphis, TN, USA
2Maine Veterans Affairs Medical Center, Augusta, ME, USA
3Semmes-Murphey Neurologic and Spine Institute, Memphis, TN, USA

*Corresponding author: Memphis Veterans Affairs Medical Center, 1030 Jefferson Ave, Memphis, TN 38104, USA. Tel.: +1-901-523-8990; fax: +1-901-577-7371 (J.C. Young) / Tel.: +1-207-623-8411; fax: +1-207-623-5792 (J.E. Caron).
E-mail address: john.young5@va.gov (J.C. Young); joshua.caron@va.gov (J.E. Caron).

Accepted 23 November 2011

Abstract

Developing embedded indicators of suboptimal effort on objective neurocognitive testing is essential for detecting increasingly sophisticated forms of symptom feigning. The current study explored whether Symbol Span, a novel Wechsler Memory Scale-fourth edition measure of supraspan visual attention, could be used to discriminate adequate effort from suboptimal effort. Archival data were collected from 136 veterans classified into Poor Effort (n = 42) and Good Effort (n = 94) groups based on symptom validity test (SVT) performance. The Poor Effort group had significantly lower raw scores (p < .001) and age-corrected scaled scores (p < .001) than the Good Effort group on the Symbol Span test. A raw score cutoff of < 14 produced 83% specificity and 50% sensitivity for detection of Poor Effort. Similarly, sensitivity was 52% and specificity was 84% when employing a cutoff of < 7 for Age-Corrected Scale Score. Collectively, present results suggest that Symbol Span can effectively differentiate veterans with multiple failures on established free-standing and embedded SVTs.

Keywords: Neuropsychological Evaluation; Symptom Validity Testing; Response Bias; Military Veterans

Introduction

Response bias during neuropsychological evaluations is a universal problem encountered by neuropsychologists. Base rates of 25%–40% are not uncommon in settings such as the Veteran’s Administration Healthcare System and medicolegal evaluations (Armistead-Jehle, 2010; Axelrod & Schutte, 2010; Larrabee, 2003; Mittenberg, Patton, Canyock, & Condit, 2002; Young, Roper, & Sawyer, 2011). Many non-neurological factors can influence test performance, even when an obvious incentive to perform poorly is unknown. As such, utilizing freestanding symptom validity tests (SVTs) and embedded measures of effort is now standard practice within clinical neuropsychology (AACN, 2007; Bush et al., 2005; Heilbronner et al., 2009).

It is not enough to assess effort in a limited capacity (Boone, 2009; Bush et al., 2005; Heilbronner et al., 2009). For example, relying on one SVT administered at the beginning of an evaluation will often be insufficient for detecting response bias. Relying on such a method also reduces the ability to assess the nature, severity, and intentionality of the response bias. This assertion has been supported by the literature as several studies show that failing two or more measures of effort significantly increases the ability to identify suboptimal effort while reducing the chance of making a false-positive error (Larrabee, 2003; Meyers & Volbrecht, 2003; Victor, Boone, Serpa, Buehler, & Ziegler, 2009). As Boone (2009) and Victor and colleagues (2009) point out, however, when using several measures of effort it is important to avoid using highly correlated SVTs due to the limited incremental information this would provide. Thus, not only is it important to use multiple tests throughout the assessment, it is also important to employ a multidimensional approach to optimize response bias detection.

The multidimensional approach, however, will necessitate developing additional indicators of effort because there are only a limited number of options at this time to adequately identify suboptimal effort. With new or revised tests continually being...
published and information about effort measures continually being leaked (Boone, 2009; Horwitz & McCaffrey, 2006; Slick, Sherman, & Iverson, 1999; Victor & Abeles, 2004), there appears to be an ever-growing need for new measures to be developed.

The Wechsler scales are some of the most widely used measures of cognitive functioning in neuropsychology, and several measures of effort have been developed utilizing its subtests and index scores. Most of these have been developed using either the Digit Span subtest from the Wechsler Adult Intelligence Scale (WAIS; Digit Span Age-Corrected Scale Score [ACSS]; Iverson & Franzen, 1994; Reliable Digit Span: Griffenstein, Baker, & Gola, 1994) or visual working memory subtests from prior versions of the Wechsler Memory Scale (WMS; Lange, Iverson, Sullivan, & Anderson, 2006; Ylioja, Baird, & Podell, 2009). Recently, the WMS-IV (Pearson, 2008) was released and, while several subtests were retained, a number of new subtests were added to assess visual memory (i.e., Designs) and visual working memory (i.e., Spatial Addition, Symbol Span).

As a novel measure of supraspan visual attention, Symbol Span has yet to be investigated as a potential indicator of response bias. Symbol Span is administered within the delay interval for other WMS-IV verbal and visual memory subtests and may be misperceived as a pure memory test by those attempting to demonstrate memory problems as directions instruct the examinee to “remember” the symbols and their order from left to right. There has been a substantial amount of investigation and validation of the Digit Span subtest based on similar theoretical principals (e.g., Axelrod, Fichtenberg, Millis, & Wertheimer, 2006; Babikian, Boone, Lu, & Arnold, 2006; Heinly, Greve, Bianchini, Love, & Brennan, 2005; Iverson & Tulsky, 2003), so we expect that Symbol Span would operate similarly. Specifically, we hypothesized that Symbol Span would accurately discriminate between Good and Poor Effort in our mixed clinical sample who underwent neuropsychological testing at a large VA hospital in the southeastern USA.

Methods

Participants

A retrospective review of an IRB-approved neuropsychology database identified 143 patients less than 70 years of age for inclusion into the current study. Participants were screened for focal impairments (e.g., hemianopia, hemispatial neglect) that could interfere with performance on the measures utilized in the study. Prior to analyses, seven participants were excluded due to dementia diagnoses. The remaining sample \( n = 136 \) was 85% men with a mean age of 45.2 years \( (SD = 13.3, \text{range } 21–67) \). The racial composition included 61% Caucasian, 37% African American, and 2% other (primarily Latino and Asian American). Ninety-three percent completed 12 or more years of education \( (\text{mean } = 13.0, SD = 2.0) \).

Seventy-four percent of the sample was referred for outpatient evaluation, 22% was referred for Compensation and Pension (C&P) purposes, and 4% was referred during inpatient hospitalizations. For those unfamiliar with the VA healthcare system, C&P exams are performed for the Veterans Benefits Administration in response a claim for impairment associated with injury or disease developed during or attributable to military service. Of those evaluated outside the C&P context, individuals were referred for memory problems \( n = 37 \), history of head injury and/or blast exposure \( n = 27 \), attention problems (e.g., attention-deficit hyperactivity disorder [ADHD]; \( n = 17 \)), possible dementia or non-specific cognitive complaints \( n = 13 \), cognitive decline secondary to general medical condition (e.g., multiple sclerosis, Parkinson’s disease; \( n = 5 \)), primary psychiatric symptoms \( n = 2 \), and a varied number of other reasons (e.g., fitness for duty, appropriateness for service, functional capacity; \( n = 5 \)). Primary diagnoses assigned in the context of the neuropsychological evaluation included post-traumatic stress disorder (PTSD) \( n = 33 \), mood disorder \( n = 32 \), Cognitive Disorder not otherwise specified \( n = 19 \), adjustment disorder \( n = 11 \), ADHD \( n = 9 \), bipolar disorder \( n = 6 \), somatoform disorder \( n = 4 \), substance-related disorder \( n = 3 \), anxiety disorder other than PTSD \( n = 2 \), personality change due to traumatic brain injury or general medical condition \( n = 2 \), and sleep disorder \( n = 1 \). Fourteen individuals (10.3%) did not receive a diagnosis on either Axis I or Axis II (i.e., V71.09).

Measures

Symbol Span. The Symbol Span subtest was administered as part of a larger neuropsychological test battery that included selected subtests from the WMS-IV. It measures visual working memory and was developed to be the analog to the Digit Span subtest from the WAIS-IV (Wechsler, 2008). During the administration, patients are shown an increasing number of simple visual designs for a period of 5 s. After the display is removed, the patient must identify the correct designs from an array of target and foil designs while also stating their correct presentation order from left to right. Psychometrically, the test appears to have acceptable international consistency \( (0.76–0.92) \) and modest test–retest reliability \( (0.72) \). Concurrent validity revealed modest correlations with WAIS-IV Digit Span \( (0.47) \).
Symptom Validity Measures. Freestanding symptom validity measures included the Word Memory Test (WMT; Green, 2003), Test of Memory Malingering (TOMM; Tombaugh, 1996), Computerized Assessment of Response Bias (CARB; Allen, Conder, Green, & Cox, 1997). On these measures, recommended cutoffs were used to identify suboptimal effort (WMT Immediate Recognition, Delayed Recognition, or Consistency $\leq 82.5$%; TOMM Trial 2 <45; CARB Trial 1 <96%). In addition to the freestanding SVTs listed above, several embedded validity markers were also employed. These measures and their cutoffs include the California Verbal Learning Test-II Long-Delay Forced Choice Recognition (<15; Millis, Putnam, Adams, & Ricker, 1995), Wisconsin Card Sorting Test Failure to Maintain Set (<2; Suhr & Boyer, 1999), Reliable Digit Span as applied to the WAIS-IV (<7; Spencer, Tree, Drag, Pangilinan, & Bieliauskas, 2010; Young, Roper, Baughman, & Yehyawi, 2011), and the Digit Span age-corrected scaled score on the WAIS-IV (<6; Young, Roper, Baughman, et al., 2011).

Procedures

Effort group classification was based on Criterion B for determining “Malingered Neurocognitive Dysfunction” (MND; Slick et al., 1999). Specially, participants failing two or more freestanding and/or embedded effort measures were categorized as providing insufficient/non-credible effort (Poor Effort group). While both Reliable Digit Span and Digit Span ACSS were used as embedded SVTs, performance below cutoffs on both was considered a single failure given their multicollinearity and failure on another SVT were required for inclusion into the Poor Effort group. Consistent with Criterion B1, patients performing at or below chance on a freestanding SVT (regardless of whether they had a second failed effort index) were automatically included into the Poor Effort group ($n = 4$).

Results

The Good Effort ($n = 94$) and Poor Effort ($n = 42$) groups did not significantly differ in age, gender, ethnicity, or education. Regarding performance on the Symbol Span subtest, the Good Effort group had significantly higher raw scores and age-corrected scores ($t = 5.22$, $p < .001$ and $t = 5.41$, $p < .001$, respectively). A summary of group demographics and Symbol Span scores can be found in Table 1.

Given observed differences in performance, logistic regressions were performed to examine the predictive validity of Symbol Span raw scores and ACSS. Analyses revealed that raw scores—$\chi^2(1, n = 136) = 22.36$, $p < .001$, OR = 1.15—and ACSS—$\chi^2(1, n = 136) = 24.93$, $p < .001$, OR = 1.49—effectively differentiated the Good and Poor Effort groups, with the raw score performing slightly better. Predictive performance of raw scores and ACSS was further examined using receiver operating characteristic (ROC) analysis, noting that greater distance from 0.50 yields a larger area under the curve (AUC) and improved prediction. Observed AUCs were moderately large with Symbol Span raw scores (AUC = 0.746, CI = 0.656–0.837) and ACSS (AUC = 0.753, CI = 0.663–0.842) displaying relatively equivalent performance (Figure 1).

Exploratory analyses examining the operational characteristics of the Symbol Span raw scores and ACSS were conducted to determine recommended cutoffs. Employing a raw score cutoff of <14 produced moderate sensitivity (0.50) and reasonably adequate specificity (0.83). Using the more stringent cutoff of <13 raised specificity to 0.89 while reducing sensitivity (0.31) significantly. The ACSS cutoff of <7 performed comparably to the raw score cutoff of <14 with moderate sensitivity (0.52) and adequate specificity (0.84). Alternatively, employing a cutoff of <6 raised specificity to an excellent level (0.95); however,

Table 1. Group Demographics and Symbol Span scores

<table>
<thead>
<tr>
<th></th>
<th>Good Effort ($n = 94$)</th>
<th>Poor Effort ($n = 42$)</th>
<th>$t$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>45.3 (13.9)</td>
<td>45.1 (12.0)</td>
<td>$t = 0.06$</td>
<td>ns</td>
</tr>
<tr>
<td>Gender (n [%])</td>
<td></td>
<td></td>
<td>$\chi^2 = 1.30$</td>
<td>ns</td>
</tr>
<tr>
<td>Men</td>
<td>78 (83)</td>
<td>38 (90)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>16 (17)</td>
<td>4 (10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnicity (n [%])</td>
<td></td>
<td></td>
<td>$\chi^2 = 4.03$</td>
<td>ns</td>
</tr>
<tr>
<td>Caucasian</td>
<td>61 (65)</td>
<td>22 (52)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>30 (32)</td>
<td>20 (48)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian American</td>
<td>2 (2)</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latino</td>
<td>1 (1)</td>
<td>—</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>13.2 (2.1)</td>
<td>12.6 (1.7)</td>
<td>$t = 1.90$</td>
<td>ns</td>
</tr>
<tr>
<td>Symbol Span Raw Score</td>
<td>21.3 (7.2)</td>
<td>15.0 (7.1)</td>
<td>$t = 4.77$</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Symbol Span Scaled Score</td>
<td>9.0 (2.6)</td>
<td>6.6 (2.6)</td>
<td>$t = 5.07$</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>
sensitivity was reduced to a level that would limit clinical usefulness (0.26). Operational characteristics for the Symbol Span raw scores and ACSS scores are summarized in Table 2.

**Discussion**

This study utilized a criterion-group design to evaluate whether performance on the WMS-IV Symbol Span subtest could be used as an embedded measure of symptom validity in an adult veteran sample. As expected, findings indicated that poor performance on the Symbol Span subtest was associated with multiple SVT failures. While raw score and ACSS cutoffs displayed comparable predictive validity, the raw scores performed slightly better and allow for a wider range of cutoff applications. For example, a raw score cutoff of <14 can be used to provide converging evidence of Poor Effort when used with other SVTs; however, as is the case with all SVTs and embedded measures of effort, it is not recommended that determination of response bias be made solely on the basis of this cutoff. Of course, utilization of Symbol Span as an SVT should also be considered in context with other competing clinical deficits that might interfere with performance on the task (e.g., hemispatial neglect, delirium, vision loss).

Given that Symbol Span was developed to be a visual analog to the Digit Span subtest, it was expected that identified cutoffs on Symbol Span scores would display comparable operational characteristics to effort indices derived from Digit Span. When comparing Symbol Span cutoffs to WAIS-III Digit Span ACSS, our findings suggest that Symbol Span (raw or ACSS) has slightly weaker sensitivity than the Digit Span ACSS, which has been reported to be between 0.36 and 0.42 when adjusting cutoffs to maintain specificity at or above 0.90 (Axelrod et al., 2006; Babikian et al., 2006). However, the Digit Span subtest underwent a substantial revision for the WAIS-IV and initial investigations of the WAIS-IV Digit Span ACSS...
suggest that Symbol Span cutoffs perform as well or better than the WAIS-IV Digit Span ACSS (Young, Roper, Baughman, et al., 2011).

While extensive investigation of suboptimal effort within the VA healthcare system was not an aim of this study, we found it important to highlight the differences in SVT failure between veteran groups and those studies of private sector SVT performance (23% when excluding C&P evaluations, 31% otherwise). Explanation for base rate differences between veteran and non-veteran samples is still poorly understood; however, numerous researchers have suggested that the structure of the VA system promotes a culture where consumers of VA services are always mindful that healthcare visits may impact future determinations of benefits (Armistead-Jehle, 2010; Laffaye, Rosen, Schurr, & Friedman, 2007; Young, Kearns, & Roper, 2011). Considering these factors, there is a clearly demonstrated importance in developing validated measures of capturing symptoms effort in all clinical contexts, but especially in veterans groups.

The present findings have several limitations that warrant consideration in future investigations. First, the sample was composed of solely of veterans, and the results may not necessarily generalize to other populations or settings. As such, there is need for further investigation of cutoffs in different clinical and medicolegal contexts. Related to the prior limitation, members of our Poor Effort group were classified using Slick and colleagues Criteria B1 and B2; however, this does not equate to malingering as other Criteria for determining MND were not considered. Resultantly, cross-validation using known-groups of malingerers is suggested.

In conclusion, the present investigation identifies a new embedded index of effort within one of the most commonly employed neuropsychological assessment measures. Consequently, it is expected that examining Symbol Span performance will prove quite useful to clinicians using WMS-IV and when used in conjunction with other indices should enhance our ability to detect suboptimal effort.

Conflict of Interest

None declared.

Acknowledgements

This material is the result of work supported with resources and the use of facilities at the Memphis VA Medical Center.

References


Young, J. C., Roper, B. L., Baughman, B. C., & Yehyawi, N. T. (2011). Correspondence of Digit Span derived effort indices with Word Memory Test performance. *Poster presented at the 39th annual meeting for the International Neuropsychological Society*, Boston, MA.