

Prevalence of below-criterion Reliable Digit Span scores in a clinical sample of older adults

RyAnna Zenisek¹, Scott R. Millis², Sarah J. Banks³, Justin B. Miller^{3,*}

¹*Department of Psychology, University of Nevada, Las Vegas, Las Vegas, NV, USA*

²*Department of Physical Medicine and Rehabilitation, Wayne State University School of Medicine, Detroit, MI, USA*

³*Cleveland Clinic Lou Ruvo Center for Brain Health, Las Vegas, NV, USA*

*Corresponding author at: Tel.: +1-702-701-7896; fax: +1-702-483-6039.

E-mail address: millerj4@ccf.org (J. B. Miller).

Accepted 5 April 2016

Abstract

Objective: The Reliable Digit Span (RDS) is a well-validated embedded indicator of performance validity. An RDS score of ≤ 7 is commonly referenced as indicative of invalid performance; however, few studies have examined the classification accuracy of the RDS among individuals suspected for dementia. The current study evaluated performance of the RDS in a clinical sample of 934 non-litigating individuals presenting to an outpatient memory disorders clinic for assessment of dementia.

Method: The RDS was calculated for each participant in the context of a comprehensive neuropsychological assessment completed as part of routine clinical care. Score distributions were examined to establish the base rate of below criterion performance for RDS cutoffs of ≤ 7 , ≤ 6 , and ≤ 5 . One-way ANOVA was used to compare performance on a cognitive screening measure and informant reports of functional independence of those falling below and above cutoffs.

Results: A cutoff score of ≤ 7 resulted in a high prevalence of below-criterion performance (29.7%), though an RDS of ≤ 6 was associated with fewer below-criterion scores (12.8%) and prevalence of an RDS of ≤ 5 was infrequent (4.3%). Those scoring below cutoffs performed worse on cognitive measures compared with those falling above cutoffs.

Conclusions: Using the RDS as a measure of performance validity among individuals presenting with a possibility of dementia increases the risk of misinterpreting genuine cognitive impairment as invalid performance when higher cutoffs are used; lower cutoffs may be useful when interpreted in conjunction with other measures of performance validity.

Keywords: Dementia; Malingering/symptom validity testing; Assessment; Reliable Digit Span

Evaluating the validity of neuropsychological profiles is an essential part of clinical care in both litigating and non-litigating contexts (Bush et al., 2005; Heilbronner, Sweet, Morgan, Larrabee, & Millis, 2009; McCrea et al., 2008), and recent research has highlighted the importance of assessing performance validity in research settings as well (An, Zakzanis, & Joordens, 2012; DeRight & Jorgensen, 2015; Silk-Eglit et al., 2014). Given findings indicating effort accounts for a large percentage of variance in neuropsychological test performance (Fox, 2011; Green, Rohling, Lees-Haley, & Allen, 2001; Meyers, Volbrecht, Axelrod, & Reinsch-Boothby, 2011), objective measures of effort have been developed to try to identify inconsistent performance. These tests are designed to ensure that the results obtained from a neuropsychological evaluation are in fact a valid representation of an individual's cognitive abilities (Boone, 2007; Larrabee, 2012). The digit span (DS) subtest from the Wechsler Adult Intelligence Scale-IV (WAIS-IV; Wechsler, 2008) is one of the most frequently used measures of attention and working memory (Hales, Yudofsky, & Roberts, 2014; Rabin, Barr, & Burton, 2005). Reliable Digit Span (RDS; Greiffenstein, Baker, & Gola, 1994) is a well-established embedded measure of performance validity (Boone, 2007; Schroeder, Twumasi-Ankrah, Baade, & Marshall, 2012) that can be derived from standard DS administration. Much of the research to date regarding performance validity, including use of RDS, has been done with adults with an overrepresentation of traumatic brain injury. Substantially less attention

has been paid to how validity measures operate in populations of older adults, particularly those suspected for dementia, as these individuals are often excluded from validation studies of performance validity tests (PVTs). The primary aim of the current study was to explore the prevalence of above and below criterion performance for several published RDS cut scores in a large outpatient dementia sample.

In most instances, PVT failure is thought to reflect suboptimal effort rather than cognitive impairment. However, significant cognitive impairment can lead to difficulty completing PVTs (Howe, Anderson, Kaufman, Sachs, & Loring, 2007; Merten, Bossink, & Schmand, 2007), especially embedded measures such as the RDS, which are more sensitive to actual cognitive functioning, as opposed to stand-alone PVTs that are theoretically so simple as to be robust to cognitive impairment. As such, the level of cognitive impairment typically seen in dementia may interfere with the ability to “pass” PVTs using established cutoffs, despite adequate effort (Teichner & Wagner, 2004). A small number of recent studies have examined the use of various free-standing and embedded PVTs in individuals with dementia, generally finding that using previously established cutoffs resulted in poor specificity rates (Bortnik, Horner, & Bachman, 2013; Dean, Victor, Boone, Philpott, & Hess, 2009; Merten et al., 2007). In one of the most comprehensive reviews of PVTs in older adults, Dean and colleagues (2009) examined the clinical utility of 18 PVTs in a sample of 214 non-litigating individuals diagnosed with various types of dementia. Using established cutoffs, inadequate specificity rates were found for 16 of these PVTs. Additionally, it was found that those with lower scores (i.e., greater impairment) on the Mini Mental Status Examination (MMSE; Folstein, Folstein, & McHugh, 1975), failed a greater number of PVTs (Dean et al., 2009). More recently, Bortnik and colleagues (2013) examined performance on four PVTs in 128 veterans with dementia diagnoses who were separated into “good-effort” and “suspect-effort” groups. They found that three of these PVTs had inadequate specificity rates, and the remaining PVT had adequate specificity but lacked sensitivity. They also found a general pattern of greater cognitive impairment resulting in higher false-positive rates (Bortnik et al., 2013). Given that PVTs are becoming a part of routine clinical practice, it is imperative to examine the operating characteristics of PVTs among individuals being evaluated for dementia. If the prevalence of false-positive errors increases as a function of cognitive impairment, it would be inappropriate to use with dementia samples.

RDS cutoff scores of ≤ 7 and ≤ 6 have been used by clinicians and researchers to identify noncredible performance (Schroeder et al., 2012). However, recent studies have indicated that a cutoff of ≤ 7 typically results in specificity rates under the recommended 90% in a variety of clinical groups, while using a cutoff of ≤ 6 results in adequate specificity rates but lower sensitivity rates (Schroeder et al., 2012; Young, Sawyer, Roper, & Baughman, 2012). Additionally, the Advanced Clinical Solutions (ACS) for the WAIS-IV and the Wechsler Memory Scale-4th edition (Wechsler, 2009b) technical manual also indicates that using a cutoff of ≤ 7 does not produce specificity rates of at least 90% in numerous clinical groups that were part of the standardization sample (Wechsler, 2009a).

Studies examining the classification accuracy of RDS in dementia have generally found specificity that is lower than acceptable standards, which in light of the known deficits in working memory in MCI and dementia syndromes (Calderon et al., 2001; Gagnon & Belleville, 2011; Kessels, Molleman, & Oosterman, 2013; Weintraub, Wicklund, & Salmon, 2012) is not surprising. Merten and colleagues (2007) found only 30% of individuals with probable Alzheimer’s disease obtained a passing score when a cutoff score of ≤ 7 was used. Heinly, Greve, Bianchini, Love, and Brennan (2005) examined the RDS in 228 memory disorder patients using a cutoff of ≤ 7 and found a specificity rate of 68%, which is still substantially below generally accepted levels. Dean and colleagues (2009) found a specificity rate of 70% using a cutoff score of ≤ 6 . More recently, Kiewel, Wisdom, Bradshaw, Pastorek, and Strutt (2012) found that using a cutoff score of ≤ 6 resulted in a specificity of 73% in a sample of individuals with probable Alzheimer’s disease (AD) ($n = 142$). Given the range of cognitive impairment seen in dementia, Dean and colleagues (2009) and Kiewel and colleagues (2012) also examined the relationship between severity of cognitive impairment and specificity of PVTs. Dean and colleagues (2009) stratified their dementia sample into three severity bands based on MMSE scores, and found that specificity on the RDS was 22% in those with severe impairment (MMSE < 15), 60% in those with mild-to-moderate impairment (MMSE = 15–20), and 86% in those with mild impairment (MMSE = 21–30). While the specificity rate for the mild impairment groups was still $< 90\%$, the authors acknowledged that their small sample size after stratification ($n = 44$) resulted in a rather large range of MMSE scores being considered mild. Kiewel and colleagues (2012) also divided their probable AD sample into mild, moderate, and severe impairment groups, but did so according to neurological findings, performance on neuropsychological measures, and functional impairment. They found a similar pattern as that found in Dean and colleagues (2009), with 17% specificity in the severe impairment group ($n = 78$; mean MMSE score = 23.4), 76% specificity in the moderate impairment group ($n = 41$; mean MMSE score = 16.8), and 89% specificity in the mild impairment group ($n = 23$; mean MMSE score = 7.7). Given that the severity was defined differently in the aforementioned studies, it is difficult to make a direct comparison; however, it appears that specificity of RDS approaches 90% in those with dementia who are less cognitively impaired.

The relationship between cognitive impairment severity and the classification accuracy of PVTs in dementia, as well as the clinical characteristics of false-positive errors is important for clinicians and researchers. While individuals with severe cognitive impairment are less likely to be misclassified as exhibiting poor effort, concerns regarding invalid neuropsychological

performance are more likely to occur in individuals reporting more mild cognitive symptoms (Dean et al., 2009). The primary aim of the current study is to investigate the prevalence of below criterion performance on the RDS in a large clinical sample of non-litigating individuals with mild cognitive impairment (MCI) and suspected dementia presenting with mixed etiological considerations. A secondary point of interest is to elucidate the relationship between RDS classifications using various cutoffs and indices of cognitive functioning to gain better understanding of the cognitive characteristics associated with below-criterion performance. It is hypothesized that application of RDS cutoffs currently validated for use in non-dementing populations such as traumatic brain injury will lead to a higher prevalence of below-criterion performance than those reported in non-dementing populations. Moreover, it is further hypothesized that individuals falling below RDS cutoffs will show significantly higher degrees of cognitive impairment than those scoring above RDS cutoffs.

Methods

Participants

The present study was reviewed and approved by the Cleveland Clinic Institutional Review Board. Data were drawn from archival records of 934 non-litigating consecutive referrals to an outpatient neurology clinic specializing in diagnosis and treatment of neurodegenerative disease. All patients in the current sample were administered the Montreal cognitive assessment (MoCA) during their initial consultation with neurology and subsequently referred for a comprehensive neuropsychological evaluation as part of routine clinical care. The main criteria for case selection were: (i) patients presenting to a memory disorders clinic; (ii) completed MoCA at time of neurology appointment; (iii) referral for, and completion of, neuropsychological testing. Cases were excluded from analysis if the patient reported being involved in litigation at any point during their diagnostic work up by neurology or neuropsychology, or if there was an obvious external or secondary gain at the time of clinical interview with neuropsychology (e.g., disability application). Additional records were excluded if diagnostic information were not available, if the primary diagnosis was of a psychiatric disorder, the patient was deemed normal by their neurologist, or if the patient was <55 years old.

The analyzed sample ($n = 579$) was 53.5% male and predominantly Caucasian (90.7%), with an average age of 72.8 years ($SD = 7.9$; $range = 55–93$ years) and an average education of 14.8 years ($SD = 2.9$; $range = 3–20$ years). Given that neuropsychological assessment was part of the diagnostic workup, by definition patients in the present sample did not have definitive diagnoses. Primary differential diagnoses were made by the neurologists on the basis of all data available, including the neuropsychological assessment. At the time data were extracted from clinical records, primary diagnoses included: AD ($n = 133$), Vascular dementia (VAD; $n = 8$), dementia with Lewy bodies ($n = 27$), Frontotemporal dementia (FTD; $n = 15$), Parkinsonian syndromes including Parkinson's disease, corticobasal syndrome, and progressive supranuclear palsy ($n = 20$), MCI ($n = 168$), unspecified cognitive disorder (COGNOS; $n = 172$), and other neurological conditions including alcohol-related dementia, traumatic/hypoxic brain injury, normal pressure hydrocephalus, stroke, or other alteration of consciousness ($n = 36$).

Measures

Neuropsychological battery. The neuropsychological battery included the DS subtest of the WAIS-IV (Wechsler, 2008), generating scores for all three components (digits forward, backward, and sequencing). RDS was calculated for each participant based on methods outlined in the ACS for Wechsler tests (Wechsler, 2009a). Cut scores of ≤ 5 , ≤ 6 , and ≤ 7 were examined. Memory was assessed with the immediate and delayed recall trials of the Hopkins verbal learning test, Revised (HVLTR; Brandt & Benedict, 2001) and Brief Visuospatial Memory Test, Revised (BVMT-R; Benedict, 1997). Language was assessed with the Boston Naming Test (Kaplan, Goodglass, & Weintraub, 2001). In addition to DS, the Block Design and Coding subtests of the WAIS-IV (Wechsler, 2008) provided measures of visuospatial reasoning and processing speed, respectively. Executive functioning was assessed with subtests from the Delis–Kaplan executive functioning system (DKEFS; Delis, Kaplan, & Kramer, 2001) including Verbal Fluency, Color-word Interference, and Trail Making Test.

MoCA (Nasreddine et al., 2005). The MoCA is a paper-and-pencil cognitive screening tool that takes ~ 10 min to administer. It consists of 12 individual tasks that are grouped in order to assess the following cognitive domains: visuospatial and executive functioning, naming, attention, language, abstraction, delayed memory recall, and orientation. Item scores are summed for a total of 30 possible points. Additionally, an education correction for individuals with <12 years of education is used. Although recent factor analysis in a similar population has demonstrated the construct validity of MoCA in this patient population (Vogel, Banks, Cummings, & Miller, 2015), validated clinical use is restricted to interpretation of the total score, which will be used in the present analyses.

Activities of Daily Living Questionnaire (ADL-Q; Johnson, Barion, Rademaker, Rehkemper, & Weintraub, 2004). The ADL-Q is an informant report questionnaire that is completed by a caregiver having frequent contact with the individual and familiarity

with their daily functioning. Content items are divided into six domains of functioning specific to: (1) self-care, (2) household care, (3) employment and recreation, (4) shopping and money, (5) travel, and (6) communication. Individual items are endorsed based on a 4-point scale, ranging from a score of 0 indicating no problem, to a score of 3 which indicates the individual is no longer capable of performing the activity. A total score representing the overall percentage of functional impairment is calculated such that scores ranging from 0% to 33% indicate minimal impairment, 34–66% moderate impairment, and $\geq 67\%$ severe impairment.

Data Analysis

Frequency distributions using each of the three RDS cutoff scores (≤ 7 , ≤ 6 , ≤ 5) were generated to ascertain the proportions of above and below criterion in the overall sample. Demographic data for those falling above and below each of the cutoffs were compared using one-way analyses of variance to assess differences in age and years of education. Chi-square tests were used to assess differences in sex and ethnicity distribution between groups. Age and years of education were included as covariates in subsequent analyses in an effort to account for these influences on cognitive testing. One-way analysis of covariance (ANCOVA) was also used to compare MoCA performance and ADL-Q impairment ratings between those falling above and below the established cutoffs. Given the size of the present sample and the associated statistical power, no additional significance testing was carried out on individual cognitive variables. The magnitude of group differences of those scoring above and below criterion was instead evaluated using effect sizes (Cohen's d) calculated based on group means and pooled standard deviations (Cohen, 1988).

Results

Demographic characteristics and cognitive performance data for the overall sample and for each RDS criterion are presented in Table 1. No significant differences in age were found between groups at any criterion, and gender or ethnicity distributions were not significantly different between groups. Years of education were significantly different between those falling above and below an RDS ≤ 6 , as well as an RDS ≤ 7 ; no differences in education were noted for an RDS ≤ 5 cutoff. For continuity across analyses, years of education were included as covariates in subsequent analyses to statistically account for these effects. RDS values ranged from 3 to 17 ($M = 8.6$, $SD = 1.9$) for the overall sample. Frequency distributions of RDS scores revealed that using RDS cutoff scores of ≤ 7 , ≤ 6 , and ≤ 5 resulted in 29.7%, 12.8%, and 4.3% of the sample falling below the cutoff score, respectively. The frequency distribution of RDS scores for the overall sample is presented in Fig. 1, and the prevalence of below criterion performance by

Table 1. Demographic and cognitive performance data by reliable digit span criterion

	Total sample	RDS ≤ 7			RDS ≤ 6			RDS ≤ 5		
	<i>M</i> (<i>SD</i>)	<i>M</i> (<i>SD</i>)		<i>d</i>	<i>M</i> (<i>SD</i>)		<i>d</i>	<i>M</i> (<i>SD</i>)		<i>D</i>
	(<i>n</i> = 579)	Above	Below		Above	Below		Above	Below	
Age	72.8 (7.9)	72.6 (7.8)	73.4 (8.1)	—	72.7 (7.8)	73.5 (8.3)	—	72.8 (7.9)	73.5 (8.1)	—
Education (years)	14.8 (2.9)	15.2 (2.8)	13.9 (2.9)	—	14.9 (2.9)	13.8 (2.5)	—	14.8 (2.9)	14.0 (2.8)	—
Sex (% female)	46.5	46.2	47.1	—	47.3	40.5	—	46.9	36.0	—
Race (% Caucasian)	90.7	91.2	89.5	—	91.1	87.8	—	91.0	84.0	—
MoCA total	21.0 (4.8)	22.2 (4.2)	18.2 (5.1)	0.86	21.6 (4.4)	17.1 (5.3)	0.92	21.3 (4.7)	15.5 (4.6)	1.25
ADL-Q total	19.1 (19.8)	16.4 (18.1)	25.0 (21.9)	0.43	17.8 (18.9)	26.4 (22.6)	0.41	18.6 (19.6)	24.3 (19.9)	0.29
HVLT-R immediate	17.6 (6.5)	18.7 (6.2)	15.2 (6.4)	0.56	18.1 (6.3)	14.3 (6.4)	0.60	17.8 (6.4)	13.1 (5.7)	0.78
HVLT-R delayed	4.1 (3.8)	4.5 (3.8)	3.4 (3.5)	0.30	4.2 (3.8)	3.6 (3.5)	0.16	4.2 (3.8)	3.8 (3.4)	0.11
BVMT-R immediate	11.9 (7.0)	13.1 (7.2)	9.0 (5.6)	0.64	12.4 (7.1)	8.3 (5.0)	0.67	12.0 (7.1)	8.7 (5.1)	0.53
BVMT-R delayed	4.2 (3.3)	4.8 (3.3)	2.9 (2.8)	0.62	4.4 (3.3)	2.8 (2.7)	0.53	4.3 (3.3)	2.9 (2.7)	0.46
Boston naming Test	50.0 (9.7)	51.2 (8.9)	47.0 (10.9)	0.42	50.5 (9.4)	46.1 (10.7)	0.44	50.2 (9.5)	44.0 (10.8)	0.61
WAIS-IV block design	27.1 (10.1)	28.7 (9.7)	23.3 (10.0)	0.55	27.8 (10.0)	22.2 (9.1)	0.59	27.4 (10.1)	21.6 (7.6)	0.65
FAS total	30.1 (12.2)	32.3 (12.1)	24.5 (10.7)	0.68	31.3 (11.8)	21.0 (11.0)	0.90	30.5 (12.1)	20.3 (10.5)	0.90
Animals	14.0 (5.6)	15.0 (5.5)	11.5 (4.8)	0.68	14.4 (5.5)	10.7 (5.0)	0.70	14.1 (5.5)	9.7 (4.2)	0.90
DKEFS trails switching	135.1 (54.9)	129.0 (52.7)	155.6 (57.6)	0.48	132.4 (54.3)	165.4 (53.2)	0.61	134.1 (54.6)	178.2 (53.9)	0.81
DKEFS inhibition	83.9 (29.8)	79.6 (27.9)	97.5 (31.9)	0.60	81.8 (28.5)	107.7 (34.3)	0.82	82.2 (28.3)	125.1 (38.7)	1.27
WAIS-IV coding	44.3 (14.2)	46.7 (13.3)	38.0 (14.6)	0.62	45.6 (13.7)	34.4 (14.5)	0.79	45.0 (13.9)	27.3 (12.2)	1.35

Note: All values reflect raw scores. MoCA = Montreal Cognitive Assessment; ADL-Q = Activities of Daily Living Questionnaire; WAIS-IV = Wechsler Adult Intelligence Scale, 4th Ed.; HVLT-R = Hopkins Verbal Learning Test, Revised; BVMT-R = Brief Visuospatial Memory Test, Revised.

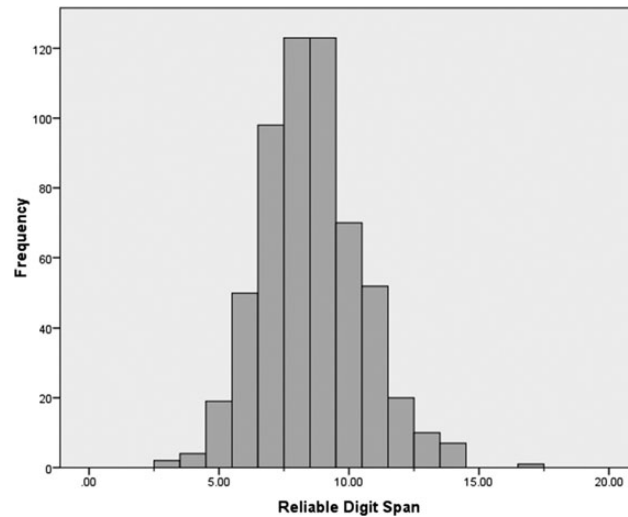


Fig. 1. Distribution of Reliable Digit Span in the overall sample.

Table 2. Percentage of below criterion performance by diagnostic group

	RDS ≤ 7	RDS ≤ 6	RDS ≤ 5
Alzheimer's disease ($n = 133$)	39.1	19.5	8.3
Vascular dementia ($n = 8$)	62.5	25.0	0.0
Dementia with Lewy bodies ($n = 27$)	37.0	14.8	0.0
Frontotemporal dementia ($n = 15$)	53.3	26.7	13.3
Parkinsonian syndromes ($n = 20$)	35.0	20.0	5.0
Mild cognitive impairment ($N = 168$)	19.0	5.4	1.2
Unspecified cognitive disorder ($n = 172$)	26.2	12.2	4.7
Other ($n = 36$)	36.1	11.1	2.8

diagnostic group is presented in Table 2. Among the three largest diagnostic groups (AD, MCI, and COGNOS), patients with AD had the highest prevalence of below criterion performance at all three cutoffs and patients with MCI had the lowest. Although some diagnostic groups (e.g., VAD, FTD) had higher prevalence rates than those observed in AD, MCI, or COGNOS, these groups were relatively small and the observed prevalence rates may not generalize to larger samples.

Group differences in the MoCA and ADL-Q total scores were examined via ANCOVA on observed scores using education as covariates for each of the three RDS cutoffs; α levels were adjusted down to 0.005 to account for multiple comparisons. At an RDS cutoff of ≤ 7 , individuals scoring below criterion had significantly lower scores than those falling above the cutoff score on the MoCA ($F(1, 536) = 69.38, p < .001$) and had significantly higher impairment ratings on the ADL-Q ($F(1, 536) = 15.00, p < .001$) with effect sizes (Cohen's d) of 0.86 and 0.43, respectively. Using a cutoff score of RDS ≤ 6 , it was similarly found that those falling below the criterion scored significantly lower than those falling above the cutoff score on the MoCA ($F(1, 536) = 45.61, p < .001$). Functional impairment ratings by caregivers on the ADL-Q ($F(1, 536) = 8.05, p = .005$), however, were not significantly different after adjusting for multiple comparisons. Effect sizes were 0.92 (MoCA) and 0.41 (ADL-Q). A similar same pattern emerged using a cutoff score of RDS ≤ 5 such that the MoCA score of those scoring below criterion was significantly lower ($F(1, 536) = 31.36, p < .001$) with an effect size of 1.25, though impairment ratings on the ADL-Q did not differ ($F(1, 536) = 1.14, p = .286, \text{Cohen's } d = 0.29$).

Effect sizes for cognitive data at an RDS cutoff of ≤ 7 ranged from 0.30 (HVLTR Delayed Recall) to 0.68 (FAS, Animals), with a mean overall effect size of 0.56. At an RDS cutoff of ≤ 6 , effect sizes ranged from 0.16 (HVLTR Delayed Recall) to 0.90 (FAS), with a mean of 0.62. For the RDS cutoff of ≤ 5 , the smallest effect size was 0.11 (HVLTR Delayed) and the largest of was 1.35 (WAIS-IV Coding), with a mean of 0.76. Across cutoffs, the mean effect size increased as the RDS criterion was lowered, suggesting that those scoring below the less stringent criteria were more cognitively impaired overall. Conversely, the largest mean differences for both delayed memory recall measures were observed at the highest RDS criteria. Delayed verbal memory in particular showed relatively small mean differences at the two lower cutoffs and the magnitude of the difference was highly comparable.

Discussion

The present findings raise concern about the use of the RDS as a measure of performance validity in persons with neurodegenerative disease. Most importantly, using the widely accepted classification standard of ≤ 7 resulted in a very high rate of below criterion performance, which in the present sample would all be highly suspicious for false-positive classification errors if the RDS were used as a measure of performance validity, as there was no obvious incentive for underperformance. Using a cutoff of ≤ 6 resulted in considerably fewer below criterion scores, with 87% of patient's scoring above this criterion. Only when cutoffs were set to nearly floor levels (i.e., ≤ 5) was an acceptable rate of below criterion performance achieved. Those scoring below cutoff, regardless of the criterion applied, were significantly more impaired on average on cognitive screening measures, and at the highest criterion, rated as more functionally impaired by caregivers than those falling above the cutoff scores. Moreover, the magnitude of the discrepancy in cognitive functioning between groups increased as the criterion was lowered, suggesting that using the higher established cutoff scores in individuals suspected of neurodegenerative disease increases the risk of misinterpreting genuine cognitive impairment as invalid performance on this index.

To better understand the cognitive characteristics of those falling below criterion, comparing performance on the more comprehensive neuropsychological battery reveals a pattern suggestive of a dose–response relationship between RDS failure and cognitive functioning. Except for memory measures, effect sizes generally increased as the criterion was lowered. At the highest RDS criterion of ≤ 7 , the largest effect sizes were observed for letter and animal fluency, nonverbal learning and recall, and processing speed, with moderate effects noted for response inhibition, verbal learning, and visuospatial construction. The smallest observed effect at this cutoff was for delayed verbal recall, though the magnitude was still considered moderate by conventional interpretation guidelines. Using the less stringent RDS criterion of ≤ 6 showed a general trend of larger effect sizes when compared with an RDS of ≤ 7 , with letter fluency and response inhibition showing large group differences, and moderate differences between measures of processing speed, animal naming, verbal and nonverbal learning, and cognitive set-shifting. Similar to the higher cutoff, delayed verbal recall again showed the smallest effect; however, the magnitude was even smaller than that observed at the higher cutoff. A comparable trend in effect sizes emerged at the lowest cutoff, suggesting that verbal memory differs considerably between those scoring above and below an RDS of 7, but the difference is negligible when the criterion is reduced. Of additional interest, at the lowest cutoff, the largest effects were observed for measures of executive functioning and processing speed, while delayed memory for verbal information was associated with the smallest effects across cut scores. This suggests that below criterion performance at this level was much more strongly characterized by executive impairments rather than episodic memory impairments, which is not surprising given the role of working memory in completion of DS.

The pattern of increasing effect sizes and consistency of this observation across cognitive measures, cognitive screenings, and functional ratings suggest that a higher burden of cognitive impairment is more likely to lead to a below criterion score on RDS when a cutoff of 7 is employed, and moreover, that below criterion performance is related to the extent of cognitive impairment. This is further evidenced by comparing the higher prevalence of below-criterion scores in AD to the lower prevalence among individuals with less severe disease (i.e., MCI), which is consistent with existing literature (Dean et al., 2009; Kiewel et al., 2012; Walter, Morris, Swier-Vosnos, & Pliskin, 2014). Based on these findings, the use of the RDS as measure of performance validity in older individuals suspected for neurodegenerative disease may be misleading, particularly among those with evidence of significant executive impairment. Given the observed rate of below criterion performance, using an RDS cutoff of 7 or less should be avoided entirely; however, adjusting the cutoff to 6 or less could provide meaningful information to help contextualize an obtained cognitive profile if interpreted in the presence of additional supporting evidence pointing towards underperformance.

It is critical to reiterate that any measure of performance validity should never be used in isolation, nor as a litmus test for poor effort, as there are multiple reasons why an individual could generate an abnormal score. When a validity indicator is deemed a failure, it should alert clinicians to explore additional influences on test performance beyond cognitive impairment (e.g., poor sleep, somnolence during the exam, metabolic disturbances, medication side effects, and fluctuations). It is also important to bear in mind that the base rate (i.e., prior probability) of underperformance is likely quite low to begin with in a neurodegenerative disease population. A positive test result on the RDS, or any PVT for that matter, will therefore have a minor effect on the posterior probability (i.e., probability of making a true positive classification decision given a positive test result) of poor effort classification. Minimizing the false-positive rate is important, and in many settings, a false-positive rate of $< 10\%$ is desirable (e.g., forensic contexts); however, the 12% false-positive rate associated with an RDS criterion of ≤ 6 is certainly close to this and may be acceptable in some instances, especially if other measures of validity are employed. Ultimately, what is considered an acceptable false-positive rate is dependent upon the assessment parameters and context of the evaluation.

The present findings also provide highly relevant base rate information of below criterion performance in a large, non-litigating clinical sample with suspected neurological disease. When exceptionally poor performance on the RDS (i.e., scores falling at or < 5) is encountered in persons where the probability of neurodegenerative disease is low (e.g., a young adult status-post mild traumatic brain injury), significant concern is warranted. This level of performance was quite rare in the present sample, with $\sim 96\%$ of

individuals scoring above this criterion, most of whom demonstrated some degree of cognitive impairment. Thus, if an RDS of 5 is encountered in an individual with less severe cognitive complaints than those reported by a patient with dementia, or if the probability of a neurodegenerative disease is low, this would be considered an extremely unusual score and cause for investigation of additional indices of performance validity.

The most notable limitation of this study is the lack of external criterion against which additional determinations of performance validity could be made. Although there was no apparent incentive to feign cognitive impairment, no other measures of performance validity were administered to support the assumption that all individuals in the present sample were putting forth full effort. As such, the present data cannot disentangle whether or not genuine cognitive dysfunction is driving low RDS scores, or if low RDS scores are actually reflective of underperformance. If present, the prevalence of underperformance in this sample is anticipated to be quite low. And given that individuals falling below any criterion in the present study carry a diagnosis of some form of neurological disease, the likelihood that these observations reflect actual cognitive impairment is substantially higher. If deemed invalid, these observations would thus more likely be false-positive classification errors. Regardless, validation of the present findings against other performance validity measures that have established classification accuracy among patients with neurodegenerative diseases is warranted.

An additional limitation of the present study is the clinical heterogeneity of the sample, as it does not allow for refined exploration of the RDS operating characteristics in narrowly defined clinical groups. However, the diversity of the present sample and its size strengthen the generalizability of the present findings to broadly include older individuals presenting with reasonable suspicion of neurodegenerative disease. Moreover, the clinical diversity of this sample is likely a closer approximation of the heterogeneity that clinicians are likely to encounter in practice. As a future aim, however, comparing RDS performance among specific clinical populations would be of particular interest. The present sample is also relatively highly educated, though education was included in all models and not shown to have significant effect, and therefore is of less concern. The lack of racial diversity in our sample, however, is a limiting factor, and the observed findings may not generalize to more diverse patient populations.

In conclusion, although the RDS has a significant amount of empirical support in many populations, use of the standard criterion of 7 in older adults with reasonable suspicion of neurodegenerative disease would be inappropriate as suggested by the present findings. Reducing the cutoff to 6 partially mitigates concerns of excessive below-criterion performance, and in the context of additional evidence of invalidity or underperformance, the RDS may be viable in some circumstances and provide a unique source of information to help frame the remaining cognitive profile; isolated use, however, is ill-advised. Future research should explore the test operating characteristics against a validated external criterion of performance validity in more narrowly defined patient groups, as well as more racially diverse populations.

Acknowledgements

The authors wish to thank January Durant for her contribution in preparing the tables included in the present manuscript.

References

- An, K. Y., Zakzanis, K. K., & Joordens, S. (2012). Conducting research with non-clinical healthy undergraduates: Does effort play a role in neuropsychological test performance? *Archives of Clinical Neuropsychology*, 27 (8), 849–857.
- Benedict, R. H. (1997). *Brief visuospatial memory test-revised: Professional manual*. Odessa, FL: Psychological Assessment Resources.
- Boone, K. B. (2007). *Assessment of feigned cognitive impairment: A neuropsychological perspective*. New York: Guilford Press.
- Bortnik, K. E., Horner, M. D., & Bachman, D. L. (2013). Performance on standard indexes of effort among patients with dementia. *Applied Neuropsychology: Adult*, 20, 233–242.
- Brandt, J., & Benedict, R. H. (2001). *Hopkins verbal learning test-revised: Professional manual*. Lutz, FL: Psychological Assessment Resources.
- Bush, S. S., Ruff, R. M., Troster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., et al. (2005). Symptom validity assessment: Practice issues and medical necessity NAN policy & planning committee. *Archives of Clinical Neuropsychology*, 20 (4), 419–426.
- Calderon, J., Perry, R., Erzinclioglu, S., Berrios, G., Dening, T., & Hodges, J. (2001). Perception, attention, and working memory are disproportionately impaired in dementia with Lewy bodies compared with Alzheimer's disease. *Journal of Neurology, Neurosurgery, & Psychiatry*, 70 (2), 157–164.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: L. Erlbaum Associates.
- Dean, A. C., Victor, T. L., Boone, K. B., Philpott, L. M., & Hess, R. A. (2009). Dementia and effort test performance. *The Clinical Neuropsychologist*, 23 (1), 133–152.
- Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *Delis-Kaplan executive function system*. San Antonio, TX: The Psychological Corporation.
- DeRight, J., & Jorgensen, R. S. (2015). I just want my research credit: Frequency of suboptimal effort in a non-clinical healthy undergraduate sample. *The Clinical Neuropsychologist*, 29 (1), 101–117.
- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state." A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12 (3), 189–198.
- Fox, D. D. (2011). Symptom validity test failure indicates invalidity of neuropsychological tests. *The Clinical Neuropsychologist*, 25 (3), 488–495.

- Gagnon, L. G., & Belleville, S. (2011). Working memory in mild cognitive impairment and Alzheimer's disease: Contribution of forgetting and predictive value of complex span tasks. *Neuropsychology, 25* (2), 226–236.
- Green, P., Rohling, M. L., Lees-Haley, P. R., & Allen, L. M., III. (2001). Effort has a greater effect on test scores than severe brain injury in compensation claimants. *Brain Injury, 15* (12), 1045–1060.
- Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment, 6* (3), 218–224.
- Hales, R. E., Yudofsky, S. C., & Roberts, L. W. (Eds.). (2014). *The American psychiatric publishing textbook of psychiatry* (6th ed.). Arlington, VA: American Psychiatric Publishing.
- Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., & Millis, S. R. (2009). American Academy of Clinical Neuropsychology Consensus Conference Statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 23* (7), 1093–1129.
- Heinly, 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* (4), 429–444.
- Howe, L. L., Anderson, A. M., Kaufman, D. A., Sachs, B. C., & Loring, D. W. (2007). Characterization of the Medical Symptom Validity Test in evaluation of clinically referred memory disorders clinic patients. *Archives of Clinical Neuropsychology, 22* (6), 753–761.
- Johnson, N., Barion, A., Rademaker, A., Rehkemper, G., & Weintraub, S. (2004). The Activities of Daily Living Questionnaire: A validation study in patients with dementia. *Alzheimer Disease and Associated Disorders, 18* (4), 223–230.
- Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston naming test* (2nd ed.). Philadelphia, PA: Lippincott, Williams & Wilkins.
- Kessels, R. P. C., Molleman, P. W., & Oosterman, J. M. (2013). Assessment of working-memory deficits in patients with mild cognitive impairment and Alzheimer's dementia using Wechsler's Working Memory Index. *Aging Clin Exp Res, 23* (5), 487–490.
- Kiewel, N. A., Wisdom, N. M., Bradshaw, M. R., Pastorek, N. J., & Strutt, A. M. (2012). A retrospective review of digit span-related effort indicators in probable Alzheimer's disease patients. *The Clinical Neuropsychologist, 26* (6), 965–974.
- Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society, 18* (4), 625–630.
- McCrea, M., Pliskin, N., Barth, J., Cox, D., Fink, J., French, L., et al. (2008). Official position of the military TBI task force on the role of neuropsychology and rehabilitation psychology in the evaluation, management, and research of military veterans with traumatic brain injury. *The Clinical Neuropsychologist, 22* (1), 10–26.
- Merten, T., Bossink, L., & Schmand, B. (2007). On the limits of effort testing: Symptom validity tests and severity of neurocognitive symptoms in nonlitigant patients. *Journal of Clinical and Experimental Neuropsychology, 29* (3), 308–318.
- Meyers, J. E., Volbrecht, M., Axelrod, B. N., & Reinsch-Boothby, L. (2011). Embedded symptom validity tests and overall neuropsychological test performance. *Archives of Clinical Neuropsychology, 26* (1), 8–15.
- Nasreddine, Z. S., Phillips, N. A., Bedirian, V., Charbonneau, S., Whitehead, V., & Collin, I., . . . , et al. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society, 53* (4), 695–699.
- Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20* (1), 33–65.
- Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2012). Reliable Digit Span: A systematic review and cross-validation study. *Assessment, 19* (1), 21–30.
- Silk-Eglit, G. M., Stenclik, J. H., Gavett, B. E., Adams, J. W., Lynch, J. K., & McCaffrey, R. J. (2014). Base rate of performance invalidity among non-clinical undergraduate research participants. *Archives of Clinical Neuropsychology, 29* (5), 415–421.
- Teichner, G., & Wagner, M. T. (2004). The Test of Memory Malingering (TOMM): Normative data from cognitively intact, cognitively impaired, and elderly patients with dementia. *Archives of Clinical Neuropsychology, 19* (3), 455–464.
- Vogel, S. J., Banks, S. J., Cummings, J. L., & Miller, J. B. (2015). Concordance of the Montreal cognitive assessment with standard neuropsychological measures. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 1* (3), 289–294.
- Walter, J., Morris, J., Swier-Vosnos, A., & Pliskin, N. (2014). Effects of severity of dementia on a symptom validity measure. *The Clinical Neuropsychologist, 28* (7), 1197–1208.
- Wechsler, D. (2008). *Wechsler adult intelligence scale-fourth edition: Administration and scoring manual*. San Antonio, TX: Pearson Assessment.
- Wechsler, D. (2009a). *Advanced clinical solutions for the WAIS-IV and WMS-IV*. San Antonio, TX: Pearson Assessment.
- Wechsler, D. (2009b). *Wechsler memory scale - fourth edition*. San Antonio, TX: Pearson Assessment.
- Weintraub, S., Wicklund, A. H., & Salmon, D. P. (2012). The Neuropsychological Profile of Alzheimer Disease. *Cold Spring Harbor Perspectives in Medicine, 2* (4), a006171.
- Young, J. C., Sawyer, R. J., Roper, B. L., & Baughman, B. C. (2012). Expansion and re-examination of Digit Span effort indices on the WAIS-IV. *The Clinical Neuropsychologist, 26* (1), 147–159.