OXFORD
UNIVERSITY PRESS

Archives
of
CLINICAL
NEUROPSYCHOLOGY

Archives of Clinical Neuropsychology 30 (2015) 130–138

The Non-Credible Score of the Rey Auditory Verbal Learning Test: Is It Better at Predicting Non-Credible Neuropsychological Test Performance Than the RAVLT Recognition Score?

Kriscinda A. Whitney[1,2,*], Jeremy J. Davis[3]

[1]*Department of Psychiatry, Richard L. Roudebush Veterans Affairs Medical Center, Indianapolis, IN 46202, USA*
[2]*Department of Psychiatry, Indiana University School of Medicine, Indianapolis, IN 46202, USA*
[3]*Division of Physical Medicine and Rehabilitation, University of Utah School of Medicine, Salt Lake City, UT 84132, USA*

*Corresponding author at: Richard L. Roudebush Veterans Affairs Medical Center (116P), 1481 W. 10th Street, Indianapolis, IN 46202, USA.
Tel.: +1-317-988-2006; Fax: +1-317-988-3578
E-mail address: kamarks@iupui.edu (K. Whitney).

Accepted 24 December 2014

Abstract

The ability of both the non-credible score of the Rey Auditory Verbal Learning Test (RAVLT NC) and the recognition score of the RAVLT (RAVLT Recog) to predict credible versus non-credible neuropsychological test performance was examined. Credible versus non-credible group membership was determined according to diagnostic criteria with consideration of performance on two stand-alone performance validity tests. Findings from this retrospective data analysis of outpatients seen for neuropsychological testing within a Veterans Affairs Medical Center ($N = 175$) showed that RAVLT Recog demonstrated better classification accuracy than RAVLT NC in predicting credible versus non-credible neuropsychological test performance. Specifically, an RAVLT Recog cutoff of ≤ 9 resulted in reasonable sensitivity (48%) and acceptable specificity (91%) in predicting non-credible neuropsychological test performance. Implications for clinical practice are discussed. *Note*: The views contained here within are those of the authors and not representative of the institutions with which they are associated.

Keywords: Test validity; Neuropsychology; Malingering; Military veterans; Memory

Introduction

The non-credible score of the Rey Auditory Verbal Learning Test (RAVLT NC) was empirically derived by Boone, Lu, and Wen (2005), who demonstrated that the use of a cutoff of ≤ 12 on this index was associated with 74% sensitivity and approximately 90% specificity in differentiating suspect effort patients from non-suspect effort patients and controls. RAVLT NC is calculated by summing the total of true recognition (i.e., recognition minus false positives) and primacy recognition (i.e., number of words recognized from the first 3rd of the test). This makes it a simple and appealing measure of performance invalidity, primarily because the scores from which it is derived are embedded within the standard administration of the RAVLT and require no complex statistical interpretation. As such, RAVLT NC requires no additional administration time and no co-administration of additional tests, unlike some of the other proposed performance validity indicators of the RAVLT (Barrash, Suhr, & Manzel, 2004; Bernard, Houston, & Natoli, 1993).

Aside from the original study by Boone and colleagues (2005), no other peer-reviewed published study appears to have specifically examined the utility of RAVLT NC. At least 18 other AVLT variables have been examined for use as indicators of performance invalidity (Suhr & Barrash, 2007). One of those variables, in particular, the recognition trial (RAVLT Recog: recognition hits), has shown promise in detecting performance invalidity. For example, results of a study by Binder, Kelly, Villanueva, and Winslow (2003) showed that a cutoff of <6 for RAVLT Recog was associated with 38% sensitivity and 92%– 95% specificity in differentiating mild head injury with poor motivation patients from moderate–severe head injury

doi:10.1093/arclin/acu094 Advance Access publication on 19 January 2015

with good motivation patients and mild head injury with good motivation patients. However, because RAVLT NC was not a variable employed in the latter study, comparison with RAVLT Recog was not possible. The same holds true for a study conducted by Meyers, Morrison, and Miller (2001), where a cutoff of ≤9 for RAVLT Recog identified 50% of simulators and 12% of participants with mild head injury who were involved in litigation, at the same time maintaining 100% specificity among groups with severe traumatic brain injury (TBI), mild TBI non-litigants, and normal controls. The original validation study of RAVLT NC did compare it to the performance of RAVLT Recog in predicting suspected performance invalidity and found that, although the sensitivity of the RAVLT Recog cutoff of ≤9 was slightly inferior to that of the RAVLT NC cutoff of ≤12 (67% vs. 74%), its specificity was about the same (93%– 92% vs. 90–92%) (Boone et al., 2005).

Theoretically, RAVLT NC captures more potentially useful information than that provided by RAVLT Recog in the sense that RAVLT NC is computed by subtracting false positives on the recognition list from the RAVLT Recog score and then adding the number of recognition hits on the first 5 items of the word list. In this way, RAVLT NC not only captures the information provided by RAVLT Recog but also includes information regarding whether or not the test taker showed the normal primacy effect by correctly recognizing items from the beginning of the list, and, also, whether or not the examiner made an excessive amount of false-positive errors on the recognition trial. The latter additional information is potentially useful because, although findings have been inconsistent, some research suggests that individuals simulating memory impairment fail to display a primacy effect on the AVLT (Bernard, 1991; Haines & Norris, 2001; Suhr, Tranel, Wefel, & Barrash, 1997; Suhr, 2002), and make more false-positive errors on the recognition trial (Suhr et al., 1997).

Notably, however, the RAVLT recognition trial employed by Boone and colleagues (2005) was in a story format (see Schmidt, 1996, p. 73), whereas the RAVLT recognition trial employed in the current study was in the more popularly administered list format. Boone and colleagues (2005) reported that it was unknown whether the cutoff scores generated from their use of the paragraph recognition format would generalize to other recognition versions. However, because their review of the literature revealed that mean scores from story and list versions of the recognition trial were similar across "real-world" samples, the authors suggested, "that the cut-offs might be able to be imported for use in list recognition formats" (Boone et al., 2005, p. 316).

Given the lack of validation research relevant to RAVLT NC, the goal of the current investigation was to examine the utility of the RAVLT NC, in comparison to RAVLT Recog, in predicting credible versus non-credible neuropsychological test performance. Non-credible group performance was identified through the use of multiple performance validity tests (PVTs) and application of the Slick, Sherman, and Iverson (1999) criteria for malingered neurocognitive dysfunction. Due to the larger scope of potentially useful data captured by RAVLT NC than RAVLT Recog, it was predicted that RAVLT NC would show better classification accuracy in predicting credible versus non-credible neuropsychological test performance.

Materials and Methods

Participants

Data were collected from the files of 194 outpatients who were consecutively referred to the lead author for neuropsychological testing within a VA Medical Center. Fifteen patients were excluded due to carrying a final diagnosis of dementia. Four more patients were excluded from the study based on their performances on the PVTs administered as part of the study. Details concerning their exclusion are available in the procedures section. The final sample consisted of 175 individuals. No patients were diagnosed with mental retardation. All patients were either active duty or veteran soldiers primarily referred for neuropsychological evaluation to assess the potential presence of cognitive dysfunction, not primarily to assess for the presence of psychiatric disorder. Consecutive referrals were reviewed for cases that were administered the test of memory malingering (TOMM; Tombaugh, 1996) and the Medical Symptom Validity Test (MSVT; Green, 2004). Files reviewed for this study overlap with files reported in a previous study (Whitney, 2013). Generally speaking, all patients referred for testing are given the latter measures.

At the time of the study, participants' medical records were retrospectively reviewed. For the final sample ($N = 175$), referral sources included psychiatry ambulatory care (37%), primary care (34%), neurology (12%), polytrauma (7%), and other clinics (4%). A small number of participants were evaluated in association with an active claim for injury through the Veterans Benefits Association and were referred through the compensation and pension clinic (6%). Patients from other referral sources were referred solely for clinical reasons. However, because "many neuropsychological evaluations conducted within the general clinical framework of Veterans Affairs healthcare may be impacted by patient concerns regarding the attainment and/or maintenance of disability," incentive to underperform and discrepancies in patient behavior versus test scores were always clinically assessed (Young, Kearns, & Roper, 2011, p. 195).

Referral reasons included history of TBI (34%), major neurologic condition (19%), and memory or concentration problem of unknown etiology (47%). Among participants ($n = 60$) reporting a history of TBI, the majority ($n = 45$) reported possible mild TBI and a minority ($n = 15$) reported moderate-to-severe TBI (Malec, Brown, Leibson, Flaada, & Mandrekar, 2007).

In the sample subset ($n = 33$) with a history of potential major neurological problem(s) other than TBI, conditions included stroke ($n = 9$), seizures ($n = 3$), multiple sclerosis ($n = 2$), hepatic encephalopathy ($n = 2$), electrical injury ($n = 2$), and Parkinson's disease ($n = 1$), among other diagnoses. Nearly half of the sample ($n = 82$) was referred due to memory or concentration problems of unknown etiology. Almost half of this group carried only a primary psychiatric diagnosis ($n = 43$), most commonly anxiety/depression ($n = 33$). Only one of these patients was referred for psychosis and a minority of these patients ($n = 7$) were referred with PTSD as part of the referral question. In the sample subset referred due to memory or concentration problems of unknown etiology, 6% ($n = 5/82$) had no psychiatric or neurological problems, whereas 34% ($n = 28/82$) had comorbid psychiatric and minor medical diagnoses that could potentially affect cognitive functioning, most commonly depression/anxiety co-occurring with a medical disorder, such hypertension, sleep apnea, transient ischemic attack, or coronary artery disease. Seven percent of those referred due to memory or concentration problems of unknown etiology ($n = 6/82$) carried only a minor medical diagnosis potentially causing cognitive difficulties, the same or similar to those previously mentioned.

In terms of patient demographics, the age of participants ranged from 21 to 77 years old, with a mean age of 49.72 years ($SD = 13.06$). Highest year of education completed by participants ranged from 7 to 20, with a mean level of 12.77 years ($SD = 2.52$). In terms of gender, 163 (93.1%) participants were male and 12 (6.9%) were female. One hundred fifty-one participants (86.3%) were Caucasian, 22 participants (12.6%) were African American, one participant (0.6%) was Hispanic, and one participant's race was unknown (0.6%).

Measures and Procedures

The RAVLT (Strauss, Sherman, & Spreen, 2006) is designed to assess verbal learning and memory. It involves oral presentation of 15 nouns (List A) over 5 learning trials, each of which is followed by an immediate recall trial. An interference list of 15 nouns (List B) is then presented and followed by a free-recall test of that list. Subsequently, short delay free recall of the first list (List A) is tested. After a 20-min delay, the test taker is again asked to recall words from List A, and a recognition trial is administered immediately after this long-delay free-recall trial.

There are various alternative formats in which the recognition trial can be administered. The basic concept is that the test taker must identify the items that were on List A. According to Strauss and colleagues (2006), having test takers identify previously presented words from a list is the most popular format. However, in other administration formats, the examiner presents a paragraph, either orally or in written form, that includes all the items from List A. The test taker must identify those words as having been on the list. In the current study, the list, not the paragraph format, was employed. Specifically, the examiner read a list of 50 words (containing all items from Lists A and B and 20 words that are phonemically or semantically similar to those in Lists A and B) and asked the test taker to indicate whether or not each word was on List A. The sum of the number of hits (saying "yes" to items that were actually on List A) on the latter trial constituted the RAVLT Recog score. As explained by Boone and colleagues (2005), the second RAVLT variable used in the present study, RAVLT NC, is calculated by summing the total of true recognition (i.e., RAVLT Recog minus false positives) and primacy recognition (i.e., number of words recognized from the first third of List A).

The TOMM (Tombaugh, 1996) is a recognition memory task in which 50 line drawings of common objects are presented during two learning trials that are each followed by forced-choice recognition trials. An optional forced-choice recognition trial can be administered following a 15-min delay. In the present study, all three trials were always administered, and patients were considered to have failed the TOMM if they performed below cutoffs specified in the manual on Trial 2 or the Retention Trial.

The MSVT (Green, 2004) is a computerized measure of verbal learning and performance validity. A series of semantically related word pairs are presented twice at the beginning of the test. Following the presentation of word pairs, four trials are administered resulting in five test scores: immediate recognition, delayed recognition, consistency, paired-associates, and free recall. MSVT scores were analyzed according to criteria outlined in the Advanced Interpretation (AI) Program (Green, 2009). The AI Program uses profile analysis based upon a variety of normative databases to categorize test takers into three basic groups: those who pass the MSVT, those who fail the MSVT due to poor effort, and those who fail the MSVT with a Genuine Memory Impairment Profile (GMIP). The main criteria that qualify an individual for the GMIP are that they (1) fail at least one of the MSVT validity indices, (2) score an average of 20 points higher on the easy subtests than the hard subtests, (3) exhibit no scores below chance, and (4) evidence clinical correlates of disability.

In line with current practice recommendations, multiple PVTs were employed to detect non-credible neuropsychological test performance (Bush et al., 2005; Heilbronner et al., 2009). As noted by Boone (2009), performance validity may vary throughout the session and should be assessed repeatedly throughout the neuropsychological examination. In the current study, the TOMM (Tombaugh, 1996) and the MSVT (Green, 2004) were administered to all participants, as explained previously. Participants were placed in the non-credible group if they failed either the TOMM or the MSVT (with anything but a GMIP) and satisfied Slick and colleagues (1999) criteria for definite or probable malingered neurocognitive dysfunction (with the exception of criterion D given

that the presence of somatoform disorder cannot be ruled out). Participants who met the criteria for a GMIP on the MSVT and also passed the TOMM were excluded from the study analyses ($n = 3$). One individual who failed the TOMM, but did not have an incentive to underperform and, thus, did not meet the minimum Slick and colleagues (1999) criteria for malingered neurocognitive dysfunction was also excluded from the study analyses. These individuals were presumed to have legitimately dysfunctional recognition memory, and were believed to represent individuals to whom the traditional interpretation of PVT failure does not apply. With regard to the individual who failed the TOMM, his Trial 2 score was 40, whereas his Retention Trial score was 38. He was referred with symptoms of stroke and had neuroimaging to support such a diagnosis.

Fifty-nine percent ($n = 37/63$) of the non-credible group failed both the TOMM and the MSVT (without a GMIP), whereas 33% ($n = 21/63$) failed only the MSVT (without a GMIP), and 8% ($n = 5/63$) failed only the TOMM. Ninety-seven percent ($n = 61/63$) of individuals in the non-credible group were either currently service connected/pursuing service connection for a medical/psychiatric condition (i.e., receiving monthly compensation or free medical treatment) or were receiving or pursuing another form of disability payment (i.e., social security disability, long-term disability through an employer, etc.). With regard to the two individuals in the non-credible group who were not pursuing or receiving disability or service-connected benefits, one wrote a letter post-testing acknowledging putting forth less than optimal effort for unclear reasons and another emphasized throughout the interview that he could "not work anymore." In the credible group, 83% ($n = 93/112$) were receiving or pursuing service connected or disability compensation.

Statistical Analyses

Except where indicated, statistical analyses were calculated using SPSS, Version 20.0 Initial analyses consisted of conducting two-tailed Pearson correlations to examine the relationships between age, education, and the RAVLT validity scale scores. Two-tailed Pearson correlations were also used to examine the relationship between RAVLT NC and RAVLT Recog scores. Credible versus non-credible groups were compared in terms of age and education using independent samples *t*-tests. Alpha was set at 0.05 for all analyses. Receiver operating characteristic (ROC) curve analyses were used to evaluate the usefulness of RAVLT NC and RAVLT Recog in predicting credible versus non-credible group membership. As part of the ROC analysis, the sensitivity and specificity of RAVLT NC and RAVLT Recog at various cutoffs were examined. Following the ROC analysis, positive and negative predictive values (NPVs) were calculated. As explained by O'Bryant and Lucas (2006), positive predictive value (PPV) refers to the likelihood that a person has condition X (i.e., non-credible neuropsychological test performance) given positive findings on test Y (i.e., scores equal to or less than cutoff on the RAVLT scale) (Glaros & Kline, 1988; McCaffrey, Palav, O'Bryant, & Labarge, 2003). NPV is defined as the likelihood that the person does not have condition X (i.e., is not demonstrating non-credible neuropsychological test performance) given a negative finding on test Y (i.e., scores above the RAVLT validity cutoff score) (Glaros & Kline, 1988; McCaffrey et al., 2003). Both positive and NPVs were calculated using the formulas presented in O'Bryant and Lucas (2006).

An estimated base rate of the condition in question (in this case, non-credible neuropsychological test performance) is needed to calculate PPV and NPV. For the present study, a base rate of 41% was employed, as it represents the average base rate of performance invalidity found in two independently conducted studies related to the issue of malingered neurocognitive dysfunction in veteran and military samples (Armistead-Jehle, 2010; Belanger, Kretzmer, Yoash-Gantz, Pickett, & Tupler, 2009). Specifically, although the Belanger and colleagues (2009) study of veteran and active duty service members reporting brain trauma was not designed to examine poor effort, the authors reported excluding 23% of participants due to suspicion of poor effort or malingering based on clinical presentation and/or if they failed certain measures of performance validity, which varied by research site and included the word memory test (Green, Allen, & Astner, 1996), the MSVT (Green, 2004), and the California Verbal Learning Test-II Long-Delay Forced-Choice Recognition (Delis, Kramer, Kaplan, & Ober, 2000). A much higher rate of performance invalidity was found in the study conducted by Armistead-Jehle (2010) in which participants consisted of veterans who were referred for evaluation of mild TBI after scoring positive on the Veterans Health Administration TBI screening measures. In the latter study, 58% of the study sample scored below cutoffs on a stand-alone PVT, the MSVT, suggesting performance invalidity.

Results

The Relationships Between Age, Education, and the RAVLT Validity Scale Scores

Means and standard deviations for RAVLT Recog and RAVLT NC are presented in Table 1. Age was not significantly correlated with RAVLT Recog scores ($r = -.08$, $p = .29$), but was minimally and significantly correlated with RAVLT NC scores ($r = -.15$, $p < .05$). Education was significantly and positively correlated with both RAVLT Recog scores ($r = .20$, $p < .01$) and RAVLT NC scores ($r = .27$, $p < .01$). RAVLT NC and RAVLT Recog scores were highly correlated with one another ($r = .70$, $p < .01$). The non-credible group was significantly younger ($M = 46.95$, $SD = 13.05$) than the credible group

Table 1. Means, standard deviations, and ranges on RAVLT validity scales for the total sample

	Credible ($N = 112$) M ± *SD* (range)	Non-credible ($N = 63$) M ± *SD* (range)
RAVLT NC	13.6 ± 5.3 (−1 to 20)	6.8 ± 7.3 (−11 to 20)
RAVLT Recog	12.8 ± 2.5 (2–15)	9.8 ± 3.6 (2–15)
RAVLT false positives	3.6 ± 3.8 (0–18)	6.3 ± 5.2 (0–21)

Notes: RAVLT NC = Rey Auditory Verbal Learning Test Non-Credible Score; RAVLT Rec = RAVLT Recognition Raw Score.

Table 2. Classification accuracy of RAVLT scores in predicting malingered neurocognitive dysfunction

Cutoff ≤	Sensitivity	Specificity	PPV[a]	NPV[a]
RAVLT NC				
5	0.43	0.89	0.73	0.69
4	0.37	0.92	0.76	0.68
3	0.30	0.93	0.75	0.66
2	0.20	0.95	0.74	0.63
1	0.16	0.97	0.79	0.62
−1	0.16	0.99	0.92	0.63
−2	0.14	1.0	1.0	0.63
RAVLT Recog				
10	0.52	0.87	0.74	0.72
9	0.48	0.91	0.78	0.72
8	0.40	0.96	0.87	0.70
7	0.27	0.96	0.82	0.65
6	0.22	0.96	0.79	0.64
5	0.13	0.97	0.75	0.62
4	0.06	0.98	0.68	0.60
3	0.05	0.98	0.63	0.60
2	0.02	0.99	0.58	0.59

Notes: RAVLT NC = Rey Auditory Verbal Learning Test Non-Credible Score; RAVLT Recog = RAVLT Recognition Raw Score; PPV = positive predictive value; NPV = negative predictive value.
[a]A 41% base rate of performance invalidity was used in these calculations.

($M = 51.28$, $SD = 12.86$), $t\,(173) = 2.12$, $p < .05$. The non-credible group also had fewer years of education ($M = 12.13$, $SD = 2.24$) than the credible group ($M = 13.13$, $SD = 2.60$), $t\,(173) = 2.56$, $p < .05$).

Because age and education showed a significant correlation with one or more of the RAVLT validity scale scores and were significantly different between credible versus non-credible groups, post hoc analyses were used to re-examine each of the statistical results discussed subsequently based on separate age and educational groups. Specifically, with reference to age, participants were divided into two groups: (1) participants aged 49 or less and (2) participants 50 years old or greater. With regard to education, participants were divided into groups who had ≤12 years of education and those who had ≥13 years of education.

ROC, Sensitivity, Specificity, and Predictive Power Analyses

Entire sample. RAVLT NC. As shown by ROC analysis, the area under the curve (AUC) of 0.78 (95% CI = 0.70–0.85) suggests that the predictive information captured by the RAVLT NC score was acceptable (Hosmer & Lemeshow, 2000). Using the RAVLT NC cutoff score of ≤12, suggested by Boone and colleagues (2005), resulted in a reasonable sensitivity of 71%, but a poor specificity of 65%. A more reasonable false-positive rate (0.08) was found when using a much lower cutoff score of ≤4 (Table 2). At the RAVLT NC cutoff of ≤4, the specificity (92%) in predicting non-credible versus credible neuropsychological test performance was similar to that reported by Boone and colleagues (2005). However, sensitivity (37%) when using a RAVLT NC cutoff ≤4 was much lower than that reported by Boone and colleagues, who noted sensitivity of 74% using a cutoff of ≤12. As can be seen in Table 2, using a cutoff of ≤4 on RAVLT NC in the present study resulted in a PPV of 0.76 and an NPV of 0.68.

RAVLT Recog. For the RAVLT Recog score, as shown by the ROC analysis, the AUC of 0.75 (95% CI = 0.67–0.83) suggests that predictive information captured by the scale was acceptable. Using the cutoff of ≤9 was optimal in the current study and resulted in a sensitivity of 48% and specificity of 91%. As can be seen in Table 2, using a cutoff of ≤9 on the RAVLT Recog in the present study resulted in a PPV of 0.78 and an NPV of 0.72.

Sample divided by age. When the sample was split into age groups (≤49 years old and ≥50 years old), the distribution of subjects was reasonable among those ≤49 years old ($N = 72$, $n = 29$ non-credible, $n = 43$ credible) and those ≥50 years old ($N = 103$, $n = 34$ non-credible, $n = 69$ credible).

RAVLT NC. Among persons 50 years old or older, as shown by ROC analysis, the AUC of 0.80 (95% CI = 71–0.89) suggests that the predictive information captured by the RAVLT NC score was excellent. As shown for the entire sample, a cutoff of ≤4 for RAVLT NC among persons 50 years old or older was ideal, resulting in a 41% sensitivity and 90% sensitivity. Among persons ≤49 years old, as shown by the ROC analysis, the AUC was 0.77 (95% CI = 0.66–0.88). However, among persons ≤49 years of age, the ideal cutoff (i.e., one resulting in at least 90% specificity) for RAVLT NC was much higher than that which was ideal for older persons, falling at ≤8 and showing 41% sensitivity and 95% specificity.

RAVLT Recog. Among persons 50 years old or older, as shown by the ROC analysis, the AUC of 0.80 (95% CI = 0.70–0.90) suggests that predictive information captured by the scale was excellent. As for the entire sample, using the cutoff of ≤9 was optimal and resulted in a sensitivity of 56% while retaining a reasonable specificity (91%). Similarly, among persons ≤49 years of age, as shown by the ROC analysis, the AUC of 0.71 (95% CI = 0.59–0.84) suggests that the predictive information captured by RAVLT Recog was acceptable. Similar to the entire sample, the RAVLT Recog cutoff of ≤9 was optimal and resulted in a sensitivity of 38% and a specificity of 91%.

Sample divided by education. When the sample was split into groups based on highest year of education completed (≤12 and ≥13 years), the distribution of subjects was reasonable among those with ≤12 years of education ($N = 120$, $n = 47$ non-credible, $n = 73$ credible). However, among those with ≥13 years of education group, there were few non-credible participants ($N = 55$, $n = 16$ non-credible, $n = 39$ credible). As the modal years of education was 12, re-defining the groups in in terms of those with ≤11 years of education and those with ≥12 years of education did not better distribute the participants, only transferring the majority from the lesser group education level to the higher group education level.

RAVLT NC. Among persons with ≤12 years of education, as shown by ROC analysis, the AUC of 0.75 (95% CI = 0.66–0.84) suggests that the predictive information captured by the RAVLT NC score was acceptable. Similar to the entire sample analysis, where an RAVLT NC cutoff of ≤4 was ideal, a cutoff of ≤3 for RAVLT NC among persons with ≤12 years of education was ideal, resulting in a 26% sensitivity and a 90% specificity. For persons with ≥13 years of education, as shown by the ROC analysis, the AUC of 0.82 (95% CI = 0.68–0.96) suggests that the predictive information captured by the RAVLT NC was excellent. However, among persons with ≥13 years of education, the ideal cutoff (i.e., one resulting in at least 90% specificity) for RAVLT NC was much higher than that which was ideal for persons with less education, falling at ≤8 and showing 56% sensitivity and 92% specificity.

RAVLT Recog. Among persons with ≤12 years of education, as shown by ROC analysis, the AUC of 0.75 (95% CI = 0.66–0.84) suggests that the predictive information captured by the RAVLT Recog score was acceptable. Similar to the entire sample analysis, where an RAVLT Recog cutoff of ≤9 was ideal, a cutoff of ≤8 for RAVLT Recog among persons with ≤12 years of education was ideal, resulting in 36% sensitivity and a 96% specificity. For persons with ≥13 years of education, as shown by the ROC analysis, the AUC of 0.74 (95% CI = 0.57–0.90) suggests that the predictive information captured by the RAVLT Recog was acceptable. However, among persons with ≥13 years of education, the ideal cutoff (i.e., one resulting in at least 90% specificity) for RAVLT Recog was slightly higher, falling at ≤10 and showing 50% sensitivity and 90% specificity.

Discussion

Several studies have suggested that scores on the RAVLT Recog trial hold particular promise in predicting credible versus non-credible neuropsychological test performance (e.g., Binder et al., 2003; Meyers et al., 2001; Suhr & Barrash, 2007). The purpose of the present examination was to compare the RAVLT Recog score with a more recently developed RAVLT combination score, RAVLT NC (Boone et al., 2005), in predicting credible versus non-credible neuropsychological test performance among veterans seen for outpatient neuropsychological evaluation. Due to the larger scope of potentially useful data captured by RAVLT NC than RAVLT Recog, it was predicted that RAVLT NC would show better classification accuracy in predicting performance invalidity. Unexpectedly, study results supported the opposite conclusion.

With regard to suggested cutoffs on the RAVLT Recog, classification analyses suggested that using a cutoff of ≤9 resulted in reasonable sensitivity (48%) and acceptable specificity (91%). A cutoff of ≤9 is identical to the cutoff recommended by Boone and colleagues (2005) and Meyers and colleagues (2001). The finding of concordance between recommended cutoffs on RAVLT Recog is particularly interesting because various studies have used different administration formats for this task, with the current study and that of Meyers and colleagues (2001) using a list format, while others, like the study of Boone and colleagues (2005) having employed a story recognition format.

With regard to suggested cutoffs on RAVLT NC, classification analyses suggested that a cutoff of ≤4 on RAVLT NC minimized false positives (specificity = 0.92) while retaining a reasonable sensitivity of 37%. This cutoff is considerably lower than that suggested by Boone and colleagues (2005) (≤12), which resulted in impressive sensitivity (71%), but poor specificity (65%). However, post hoc analyses showed that, among persons <50 years old and those with post high school education, respectively, an RAVLT NC cutoff falling somewhere in the middle ground between the optimal cutoff for the entire sample in the current study and the Boone and colleagues (2005) study was ideal. Specifically, an RAVLT NC cutoff of ≤8 among persons <50 years old and those with post high school education resulted in sensitivities of 41% and 56%, respectively, while maintaining at least 90% specificity. The elevation in the RAVLT NC cutoff when considering only younger and more highly educated persons in the current study makes conceptual sense. It is well documented that older males, in particular, perform more poorly on auditory verbal learning tests (Geffen, Moar, O'hanlon, Clark, & Geffen, 1990). The participants in the current study were more predominantly male, slightly older, and slightly less educated than those in the Boone and colleagues (2005) sample. Cross-validation of the current findings may help develop differing RAVLT NC cutoffs based on age and also demonstrate whether these cutoffs generalize to the female population.

Although there are a myriad of reasons that RAVLT NC cutoff scores may have been lower in the present study compared with the study conducted by Boone and colleagues (2005), one obvious explanation may be that the recognition formats varied between the studies. As explained previously, Boone and colleagues (2005) employed a story recognition format in their study, whereas the current study employed a list-recognition format. The paragraph administered in the recognition trial in the Boone and colleagues (2005) study only included one noun from List B, the distractor list. Aside from the 15 List A target nouns and the one List B distractor noun in the 70-word paragraph, there were only 14 other nouns. Most of these other nouns could be considered to be at least loosely semantically related to items in List A or B. In contrast, the list-recognition format employed in the current study contained all 15 List B distractor nouns and 20 additional nouns that were phonemically or semantically similar to those in Lists A and B. Thus, there is a greater likelihood of false-positive errors using the list format as opposed to the paragraph recognition format.

Looking closely at the results of the current study and the Boone and colleagues (2005) study, the non-credible participants in the current study made an average of six false-positive errors (Table 1), while the non-credible group in the Boone and colleagues (2005) study made an average of only two false-positive errors. A similar pattern was demonstrated for credible patients, where our credible group made an average of four false-positive errors (Table 1) and the credible clinic patients in the Boone and colleagues (2005) study made an average of only 1 false-positive error. Because false-positive errors are subtracted from the RAVLT NC score, they lower the total RAVLT NC score. Thus, due to the greater likelihood of making false-positive errors using the list-recognition format than the story recognition format, the current findings regarding RAVLT NC scores will only generalize to RAVLT administrations employing the list-recognition format.

When administering the list-recognition format to a heterogeneous sample, these data suggest that, in terms of positive predictive value (PPV), using an RAVLT NC cutoff of ≤4, a clinician would have a 76% probability of being correct in suspecting a patient of invalid neuropsychological test performance. In terms of NPV, given an above cutoff RAVLT NC score, the same clinician would have a 68% probability of being correct in not suspecting a patient of invalid performance based on their RAVLT NC score. A clinician's chances of being correct in suspecting invalid performance using RAVLT Recog, rather than RAVLT NC, are slightly better. Specifically, the use of an RAVLT Recog cutoff of ≤9 resulted in a 78% chance of being correct in suspecting a patient of invalid test performance. In terms of being correct in not suspecting a patient of invalid performance based on their RAVLT Recog score, the chance of being correct was still somewhat low, falling at 72%.

The finding of less than ideal PPVs for the individual RAVLT variables is, perhaps, not surprising. Some researchers have suggested that the classification accuracy of embedded symptom validity measures in general is so poor that these measures should not be used in the absence of free standing measures symptom validity measures (Miele, Gunner, Lynch, & McCaffrey, 2012). Other researchers have found that requiring failure on any one of several scoring methods for the same embedded measure, rather than utilizing just one scoring method for the embedded measure, improves sensitivity while maintaining a similar specificity in predicting invalid neuropsychological test performance (Axelrod, Myers, & Davis, 2014). Employing such a method in the current study actually resulted in less than ideal specificity, but slightly improved sensitivity. Specifically, considering failure on either RAVLT NC (≤4) or RAVLT Recog (≤9) as an indication of RAVLT failure in the current study resulted in a sensitivity of 52% and a specificity of 86% in predicting credible versus non-credible neuropsychological test performance. Using a base rate of 41% for non-credible performance, the latter values result in a 72% PPV and a 72%NPV, which is not better than using the RAVLT Recog score alone. Still other researchers specifically recommend employing multiple separate embedded symptom validity tests and requiring failure on at least two of these to predict performance invalidity (Victor, Boone, Serpa, Buehler, & Ziegler, 2009).

As participants in the current study also completed a larger neuropsychological battery including the Response Bias Scale (RBS) of the MMPI-2 (Gervais, Ben-Porath, Wygant, & Green, 2007) and Reliable Digit Span of the Wechsler Adult Intelligence Scale-Fourth Edition (Wechsler, 2008), it was possible to examine the latter strategy. Specifically, in the current study, post hoc analyses showed that requiring failure on two of three embedded performance/symptom validity tests (i.e.,

RAVLT Recog ≤9 or RBS ≥17 or Reliable Digit Span ≤7) resulted in low, but acceptable, sensitivity and unmatched specificity. Specifically, in predicting non-credible neuropsychological test performance, the sensitivity of using failure on two of the three embedded validity measures fell at 29%, while the specificity was impressive, falling at 97%. Using the latter strategy in this sample, along with a base rate of performance invalidity of 41%, a clinician would have a 87% chance of being correct in suspecting a patient of non-credible neuropsychological test performance based on failure of two of the three embedded symptom/performance validity tests and a 66% probability of being correct in not suspecting a patient of non-credible neuropsychological test performance based on their performance on the three embedded symptom/PVTs.

A clear limitation of this study is that the participants were primarily middle-aged Caucasian males, all of whom were receiving neuropsychological services at a Veterans Affairs Medical Center. It is possible that the results of this study might not generalize to other clinical settings where the base rate of performance invalidity may be lower. Since 83% of the credible group in the current study had or were pursuing service connected or disability compensation, it is likely that an unknown percentage of individuals who were not consistently performing to true ability were retained in the credible sample. Published research suggests that the PVTs used for group assignment have imperfect sensitivity, with the TOMM showing only 50% sensitivity in detecting malingering among persons with chronic pain and TBI (Greve, Ord, Curtis, Bianchini, & Brennan, 2008). Retention of individuals who are malingering in the credible group precludes the possibility of determining true specificity rates and cut-scores.

Another limitation of this study is that the categorization of patients into credible versus non-credible groups involves clinical judgment. A diagnosis of malingering is not a decision that should be based on test results alone, but must be made in consideration of other psychometric, behavioral, and collateral data (see Slick et al., 1999). Slick and colleagues (1999) criteria for malingered neurocognitive dysfunction were employed in the present study. However, it is notable that the application of these criteria remains somewhat subjective in nature. It is possible that other clinicians may or may not have perceived or uncovered a motive to feign impairment in the patients examined herein and, thus, group placement may have varied from the current study, thereby affecting the final study results. Future research may address the applicability of these study findings and methods to various patient populations and settings. Given that four of nine participants in the credible group who performed below RAVLT NC cutoff (i.e., false-positive cases) had a history of likely moderate-to-severe TBI, it may be helpful to consider future research examining RAVLT validity indices in a sample with greater representation across the spectrum of TBI severity.

References

Armistead-Jehle, P. (2010). Symptom validity test performance in U.S. veterans referred for evaluation of mild TBI. *Applied Neuropsychology, 17 (1)*, 52–59.

Axelrod, B. N., Meyers, J. E., & Davis, J. J. (2014). Finger Tapping Test performance as a measure of performance validity. *The Clinical Neuropsychologist, 28 (5)*, 876–888.

Barrash, J., Suhr, J., & Manzel, K. (2004). Detecting poor effort and malingering with an expanded version of the Auditory Verbal Learning Test (AVLTX): Validation with clinical samples. *Journal of Clinical and Experimental Neuropsychology, 26*, 125–140.

Belanger, H. D., Kretzmer, T., Yoash-Gantz, R., Pickett, T., & Tupler, L. A. (2009). Cognitive sequelae of blast-related versus other mechanisms of brain trauma. *Journal of the International Neuropsychological Society, 15 (1)*, 1–8.

Bernard, L. C. (1991). The detection of faked deficits on the Rey Auditory Verbal Learning Test: The effects of serial position. *Archives of Clinical Neuropsychology, 6*, 81–88.

Bernard, L. C., Houston, W., & Natoli, L. (1993). Malingering on neuropsychological memory tests: Potential objective indicators. *Journal of Clinical Psychology, 49*, 45–53.

Binder, L. M., Kelly, M. P., Villanueva, M. R., & Winslow, M. M. (2003). Motivation and neuropsychological test performance following mild head injury. *Journal of Clinical and Experimental Neuropsychology, 25*, 420–430.

Boone, K. B. (2009). The need for continuous and comprehensive sampling of effort/response bias during neuropsychological examinations. *The Clinical Neuropsychologist, 23*, 729–741.

Boone, K. B., Lu, P., & Wen, J. (2005). Comparison of various RAVLT scores in the detection of noncredible memory performance. *Archives of Clinical Neuropsychology, 20*, 301–319.

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 and planning committee. *Archives of Clinical Neuropsychology, 20 (4)*, 419–426.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). California verbal learning test *(2nd ed.)*. San Antonio, TX: PsychCorp.

Geffen, G., Moar, K. J., O'hanlon, A. P., Clark, C. R., & Geffen, L. B. (1990). Performance measures of 16- to 86-year-old males and females on the auditory verbal learning test. *Clinical Neuropsychologist, 4 (1)*, 45–63.

Gervais, R. O., Ben-Porath, Y. S., Wygant, D. B., & Green, P. (2007). Development and validation of a Response Bias Scale (RBS) for the MMPI-2. *Assessment, 14*, 196–208.

Glaros, A. G., & Kline, R. B. (1988). Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. *Journal of Clinical Psychology, 44*, 1013–1023.

Green, P. (2004). *Green's medical symptom validity test (MSVT) for windows: User's manual*. Edmonton, Canada: Green's Publishing.

Green, P. (2009). *The advanced interpretation program for the WMT, MSVT, NV-MSVT, and MCI*. Edmonton, Canada: Green's Publishing.

Green, P., Allen, L., & Astner, K. (1996). *Manual for the computerized word memory test*. Durham, NC: CogniSystem.

Haines, M. E., & Norris, M. P. (2001). Comparing student and patient simulated malingerers' performance on standard neuropsychological measures to detect feigned cognitive deficits. *The Clinical Neuropsychologist, 15*, 171–182.

Greve, K. W., Ord, J., Curtis, K. L., Bianchini, K. J., & Brennan, A. (2008). Detecting malingering in traumatic brain injury and chronic pain: A comparison of three forced-choice symptom validity tests. *The Clinical Neuropsychologist, 22*, 896–918.

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., & Millis, S. R., & Conference Participants. (2009). American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 23*, 1093–1129.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Malec, J. F., Brown, A. W., Leibson, C. L., Flaada, J. T., & Mandrekar, J. N. (2007). The Mayo classification system for traumatic brain injury severity. *Journal of Neurotrauma, 24*, 1417–1424.

McCaffrey, R. J., Palav, A., O'Bryant, S. E., & Labarge, A. S. (2003). Practitioner's guide to symptom base rates in clinical neuropsychology. New York: Plenum.

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

Miele, A. S., Gunner, J. H., Lynch, J. K., & McCaffrey, R. J. (2012). Are embedded validity indices equivalent to free-standing symptom validity tests? *Archives of Clinical Neuropsychology, 27 (1)*, 10–22.

O'Bryant, S. E., & Lucas, J. A. (2006). Estimating the predictive value of the Test of Memory Malingering: An illustrative example for clinicians. *The Clinical Neuropsychologist, 20*, 533–540.

Schmidt, M. (1996). *Rey auditory and verbal learning test: A handbook*. Las Angeles: Western Psychological Services.

Slick, D. J., Sherman, E. M. S., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist, 13 (4)*, 545–561.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York: Oxford University Press.

Suhr, J., Tranel, D., Wefel, J., & Barrash, J. (1997). Memory performance after head injury: Contributions of malingering, litigation status, psychological factors, and medication use. *Journal of Clinical and Experimental Neuropsychology, 19*, 500–514.

Suhr, J. A. (2002). Malingering, coaching, and the serial position effect. *Archives of Clinical Neuropsychology, 17*, 69–77.

Suhr, J. A., & Barrash, J. (2007). Performance on standard attention, memory, and psychomotor speed tasks as indicators of malingering. In G. J. Larrabee (Ed.), *Assessment of malingered neuropsychological deficits* (pp. 131–170). New York: Oxford University Press.

Tombaugh, T. N. (1996). *TOMM: Test of memory malingering*. North Tonawanda, NY: Multi-Health Systems.

Victor, T. L., Boone, K. B., Serpa, G., Buehler, J., & Ziegler, E. A. (2009). Interpreting the meaning of multiple symptom validity test failure. *The Clinical Neuropsychologist, 23 (2)*, 297–313.

Wechsler, D. A. (2008). *Wechsler adult intelligence scale-IV*. San Antonio, TX: Psychological Corporation.

Whitney, K. A. (2013). Predicting Test of Memory Malingering (TOMM) and Medical Symptom validity Test (MSVT) Failure within a Veterans Affairs Medical Center: Use of the Response Bias Scale (RBS) and Henry-Heilbronner Index (HHI). *Archives of Clinical Neuropsychology, 28 (3)*, 222–235.

Young, J. C., Kearns, L. A., & Roper, B. L. (2011). Validation of the MMPI-2 Response Bias Scale and Henry – Heilbronner Index in a U.S. veteran population. *Archives of Clinical Neuropsychology, 26*, 194–204.