

Initial Clinical Validation of an Embedded Performance Validity Measure within the Automated Neuropsychological Metrics (ANAM)

Tresa M. Roebuck-Spencer^{1,*}, Andrea S. Vincent¹, Kirby Gilliland¹, Dan R. Johnson², Douglas B. Cooper³

¹*Cognitive Science Research Center, University of Oklahoma, Norman, OK, USA*

²*Washington and Lee University, Lexington, VA, USA*

³*Defense and Veterans Brain Injury Center, Department of Neurology, San Antonio Military Medical Center (SAMMC), Ft. Sam Houston, TX, USA*

*Corresponding author at: Cognitive Science Research Center, University of Oklahoma, 3200 Marshall Ave., Suite 260, Norman, OK 73072, USA.

Tel.: +1-405-325-7467; fax: +1-405-325-2523.

E-mail address: tresa@ou.edu (T.M. Roebuck-Spencer).

Accepted 23 June 2013

Abstract

The measurement of effort and performance validity is essential for computerized testing where less direct supervision is needed. The clinical validation of an Automated Neuropsychological Metrics-Performance Validity Index (ANAM-PVI) was examined by converting ANAM test scores into a common metric based on their relative infrequency in an outpatient clinic sample with presumed good effort. Optimal ANAM-PVI cut-points were determined using receiver operator characteristic (ROC) curve analyses and an a priori specificity of 90%. Sensitivity/specificity was examined in available validation samples (controls, simulators, and neurorehabilitation patients). ANAM-PVI scores differed between groups with simulators scoring the highest. ROC curve analysis indicated excellent discriminability of ANAM-PVI scores ≥ 5 to detect simulators versus controls (area under the curve = 0.858; odds ratio for detecting suboptimal performance = 15.6), but resulted in a 27% false-positive rate in the clinical sample. When specificity in the clinical sample was set at 90%, sensitivity decreased (68%), but was consistent with other embedded effort measures. Results support the ANAM-PVI as an embedded effort measure and demonstrate the value of sample-specific cut-points in groups with cognitive impairment. Examination of different cut-points indicates that clinicians should choose sample-specific cut-points based on sensitivity and specificity rates that are most appropriate for their patient population with higher cut-points for those expected to have severe cognitive impairment (e.g., dementia or severe acquired brain injury).

Keywords: Computer-based testing; Performance Validity Testing; ANAM; Simulators; Neuropsychology

Introduction

Valid assessment of cognitive functioning relies on the examinee's full motivation and effort to perform as well as possible. Suboptimal performance during neuropsychological testing results in test scores that do not accurately represent a person's actual current level of cognitive functioning. Given that the results of neuropsychological assessment can have significant implications, such as determining qualification for disability benefits after an injury or ability to return to work or other activities, tests of performance validity are becoming a standard part of neuropsychological evaluations, particularly for forensic neuropsychologists (Bush et al., 2005).

Determination of invalid performance is typically based on the presence of unusual patterns of performance, atypically low scores in comparison with groups with known cognitive impairment, or a combination of these two factors (Heilbronner et al., 2009; Larrabee, 2007a; Slick, Sherman, & Iverson, 1999). Standalone tests, such as the Green Word Memory Test or the Test of Memory Malinger (TOMM), are the most common assessment methods for the detection of invalid performance. Although these types of tests have historically shown reasonable to good sensitivity (for review, see Larrabee, 2007b), there are a number of reasons to consider alternative methods for assessing effort. Trade-offs between sensitivity (the ability to correctly detect invalid performance) and specificity (the ability to correctly identify valid performance) are inherent with the choice of test

score cut-points determining whether sensitivity versus specificity is favored. One important consideration with neuropsychological performance validity tests (PVTs) is the need to avoid the potential false-positive detection of invalid performance (resulting in lower specificity) in clinical patients who may perform at abnormal levels on these tests due to actual cognitive impairment.

Current recommendations suggest comprehensive and continuous assessment of effort over the course of a neuropsychological evaluation as individuals may not put forth consistently good effort throughout the evaluation (Boone, 2009). This practice requires that multiple PVTs be administered throughout the battery, which is most effectively accomplished by the utilization of metrics embedded within neuropsychological tests. Examples of such metrics include forced recognition trials on the California Verbal Learning Test, and failure to maintain set on the Wisconsin Card Sorting Test, and so on (Larrabee, 2007b). Embedded PVTs result in decreased vulnerability to coaching and greater potential to assess effort continuously throughout the assessment period without adding time and burden on the test-taker (Suhr & Gunstad, 2007). Although the sensitivity of individual embedded measures to detect invalid performance can be low, there is evidence that sensitivity is greatly improved when multiple measures are considered in combination (Larrabee, 2008). Additionally, there are several embedded PVTs (e.g., equations from the Rey Auditory Verbal Learning test, Digit Symbol test, and Rey–Osterrieth Complex Figure Test) that meet or exceed the sensitivity reported for standalone measures (Boone, Lu, & Wen, 2005; Kim et al., 2010; Reedy et al., 2013).

The current study will examine a newly developed embedded measure of invalid performance available within the Automated Neuropsychological Assessment Metrics (ANAM), the ANAM Performance Validity Index (ANAM-PVI). ANAM is a computer-based library of tests that was originally developed within the Department of Defense for a range of military applications including an initial goal to measure the potential cognitive side effects of countermeasures to neurotoxic agents (Friedl et al., 2007). Over time, ANAM's applications extended to the measurement of cognitive effects of environmental toxins (McDiarmid et al., 2002; Rahill et al., 1996), exposure to extreme environments (Lowe et al., 2007), medications (Wilken, Sullivan, Lewandowski, & Kane, 2007), clinical disorders (Kane, Roebuck-Spencer, Short, Kabat, & Wilken, 2007; Wilken et al., 2003), and sports-related concussion (Bleiberg et al., 2004; Sim, Terryberry-Spohr, & Wilson, 2008). Strengths of ANAM are that it contains multiple alternate versions allowing for longitudinal assessment and that its precise measurement of reaction time (RT) allows for the measurement of subtle changes in cognition.

In 2008, Congress directed that all U.S. military Service Members receive pre- and post-deployment neuropsychological assessment (United States House of Representatives H.R. 4986, 2008). ANAM was the neurocognitive assessment tool selected by the Assistant Secretary of Defense for Health Affairs to meet this charge and is currently being used to document baseline levels of cognitive functioning prior to military deployment and to assist with assessment and clinical management following a concussion or other cognitive insult. Consistent with findings from civilian sports concussion research, ANAM has been shown to be sensitive to the early effects of concussion sustained during military deployment (e.g., within 72 h of injury; Bryan & Hernandez, 2012; Coldren, Russell, Parish, Dretsch, & Kelly, 2012; Kelly, Coldren, Parish, Dretsch, & Russell, 2012; Luethcke, Bryan, Morrow, & Isler, 2011). As with other neuropsychological tests, the assessment of valid responding within ANAM is essential for the accurate assessment of cognition and for clinical decision-making. This is particularly true given ANAM's widespread use for military baseline assessment, which often employs group-testing formats coupled with the decreased need for direct observation from the examiner during computer-based test administration.

ANAM is a data-rich instrument and provides many performance-based metrics for analysis and interpretation. Calculation of the ANAM-PVI capitalizes on two specific metrics available for each ANAM test: Accuracy of responding and RT (in ms). Accuracy on ANAM is measured as the percent of total items correct. Similar to effort tests using forced-choice methodology, it is based on a two-alternative, forced choice where a range of values around 50% correct reflects chance responding (depending on the number of test items). Responding significantly below 50% would indicate either reversal of the response keys (i.e., misunderstanding of test directions) or intentional poor responding. Additionally, the difficulty level of ANAM tests is quite low (Vincent, Roebuck-Spencer, Gilliland, & Schlegal, 2012) with performance exceeding 80% in non-impaired individuals on most ANAM tests. Even individuals with cognitive impairments commonly perform significantly above 50% with low performance generally falling in the 80% range (Woodhouse et al., 2013). Thus, unusually low accuracy scores are atypical and may suggest intentionally poor performance.

Researchers have also recognized the importance of examining RT to identify individuals with invalid performance. For instance, a number of studies have demonstrated that individuals simulating cognitive deficits often exhibit slowed RT on computerized tests of performance validity (Bolan, Foster, Schmand, & Bolan, 2002; Rees, Tombaugh, Gansler, & Moczynski, 1998; Slick, Hopp, Strauss, & Thompson, 2005; Tan, Slick, Strauss, & Hultsch, 2002). Further, similar studies show significantly slowed and more variable RTs among simulators compared with controls and clinical groups (Reicker, 2008). It has been suggested that this exaggerated slowing among individuals intentionally feigning cognitive impairment is due to their tendency to “aim too low” so that their RT is well below that expected from a severely injured individual (Van Gorp et al., 1999). Theories proposed to explain this phenomenon (Bolan et al., 2002; Cercy, Schretlen, & Brandt, 1997) suggest that the RTs of individuals intentionally performing poorly reflect additional decision processes above and beyond the automatic, perceptual-motor processes required

to respond to a test item (e.g., conscious attempts to slow down, recalling responses from previous trials, etc.). This additional decision process leads to increased RTs and variability in performance. Increased variability can be quantified using cognitive psychology and mathematical theories that model the impact of decision-based and automatic processes on theoretically expected RT distributions. Therefore, RTs may serve as a useful metric for identifying aberrant patterns of performing and have the added benefit of being more resistant to coaching than measures of accuracy (Dunn, Shear, Howe, & Ris, 2003; Rose, Hall, & Szalda-Petree, 1995).

The derivation of the ANAM-PVI capitalizes on this cognitive theory and combines it with the more traditional approaches. Specifically, the ANAM-PVI incorporates a discrepancy metric derived from the decomposition of an individual's RT distribution into components believed to measure decision-based and automatic processes that can be used to quantify the variability of intra-individual RTs (Luce, 1986; Schmiedek, Oberauer, Wilhelm, Suss, & Wittmann, 2007). The greater the additional decision-based responding above and beyond the actual response time to the test stimulus, the greater the RT Discrepancy score (Johnson, Gilliland, & Vincent, 2009).

In an initial study of an embedded measure of invalid performance for ANAM, Johnson and colleagues (2009) examined the sensitivity/specificity of aforementioned accuracy and RT Discrepancy scores computed for a series of ANAM tests to detect simulated cognitive impairment and also assessed the correspondence of these scores to findings from traditional standalone effort measures. This study randomized a college sample to take ANAM either normally or to simulate poor performance to convince the examiner that they had sustained a brain injury. An additional group was provided with coaching instruction on how to feign impairment without being detected. Measures of accuracy and RT Discrepancy were significantly different across these groups, with simulators showing worse performance on both variables compared with controls. Whereas the coached group was able to modify their accuracy scores to look more like controls, they were not able to modify their RT Discrepancy scores, indicating that these scores were resistant to coaching. Sensitivity of this embedded ANAM performance validity measure to detect simulators and coached participants was 87% and specificity to detect controls was 90%. Classification rates were highly concordant with concurrently administered established measures of performance validity (i.e., Victoria Symptom Validity Test and the Computerized Assessment of Response Bias; Allen, Green, Cox, & Conder, 2006; Slick et al., 2005).

The goal of the current study was to provide the clinical validation of an embedded ANAM-PVI and to establish cut-points that minimize potential false-positive errors in individuals with known cognitive impairment. To achieve this goal, variables validated in the original simulator study were combined into an overall ANAM-PVI score based on methods described by Silverberg, Wertheimer, and Fichtenberg (2007) in their study of a PVI for the Repeatable Battery for the Assessment of Neuropsychological Status. ANAM data collected from an outpatient clinical sample with known neurological diagnoses and documented adequate performance validity were used to establish cut-points for the ANAM-PVI that maximized specificity in this group. Validation of these cut-points in other available samples was then explored.

Method

Participants

Samples used in this study were drawn from several sources. The first sample served as the reference group for establishing cut-points for the ANAM-PVI. This group initially consisted of 66 consecutive referrals drawn from an outpatient brain injury clinic within the Department of Orthopedics and Rehabilitation at a large, military medical center who took ANAM as part of a standard neuropsychological evaluation. Patients were referred for a neuropsychological evaluation to determine the extent of cognitive impairment and to assist with rehabilitation treatment recommendations. Neuropsychological evaluations for administrative purposes, including medical evaluation boards or disability determination, were conducted in a separate department within the medical center. Of this initial sample, four patients were excluded due to missing ANAM data; one patient was excluded due to failure on PVT testing; and one patient was excluded due to the treating clinician's description of the patient having "disengaged" behaviors that invalidated portions of testing battery (despite initially passing PVT testing).

Included patients ($n = 60$) represented a heterogeneous sample of patients with acquired brain injury, with the following range of diagnoses; traumatic brain injury (TBI; $n = 45$), stroke ($n = 2$), subarachnoid hemorrhage/aneurysm ($n = 2$), brain tumor ($n = 4$), anoxic injury ($n = 4$), and electrical injury ($n = 3$). Specific to those with a diagnosis of TBI, the majority of these individuals were described as having sustained a moderate, penetrating, or severe TBI ($n = 29$). The remaining individuals were described as having sustained either a complicated mild TBI with positive findings on neuroimaging ($n = 9$) with only seven individuals being described as having a mild TBI. This is consistent with an existing policy that individuals with mild TBI are not routinely referred for a comprehensive neuropsychological evaluation within this clinic. The small sample of patients with mild TBI ($n = 7$) included in the sample underwent neuropsychological evaluation due to their occupational status (either healthcare providers, senior military leadership, or special forces). Average time between onset of injury/diagnosis and neuropsychological evaluation was 34.8 weeks.

Table 1. Outpatient sample demographics

Total (<i>n</i>)	60
Gender (men:women)	56:4
Ethnicity (White:Other:Unknown)	35:21:4
Age (mean [<i>SD</i>], range)	29.7 (8.6), 19–57
Years of Education (<i>n</i> = 57)	13.3 (2.0), 12–20
Time Since Injury (weeks; <i>n</i> = 54)	34.8 (47.0), 1.6–263
PTSD Checklist-Military Version	31.9 (14.2), 17–64
Wechsler Adult Intelligence Scale-Third Edition; Full-Scale IQ	102.6 (16.7), 76–155
Wisconsin Card Sorting Test-Perseverative Errors T-Score (<i>n</i> = 47)	56.1 (34.2), 1–99
Trail Making Test, Part A (<i>T</i> -score; <i>n</i> = 55)	45.8 (13.2), 16–70
Trail Making Test, Part B (<i>T</i> -score; <i>n</i> = 55)	46.3 (14.0), 16–71
Wechsler Memory Scale-Third Edition; Logical Memory Subtest I/II (<i>n</i> = 54)	10.3 (3.2)/10.2 (3.2) 3–17/1–15
Wechsler Memory Scale-Third Edition; Visual Reproduction Subtest I/II (<i>n</i> = 52)	10.7 (3.5)/10.2 (3.5) 2–17/1–17

Notes: PTSD = Post-Traumatic Stress Disorder; IQ = Intelligence Quotient.

The majority of patients were men (93%) with an average age of 29.77 years ($SD = 8.6$; range, 19–57). Sample characteristics and cognitive performance are further described in Table 1. As these were clinical evaluations, not all patients received the same PVT. Most patients received Green's Medical Symptom Validity Test (40%) or the TOMM (46%). A minority of cases was given the Victoria Symptom Validity Test (4%) or the embedded Forced Choice Memory test from the California Verbal Learning Test-Second Edition (7%). Two patients were administered more than one standalone performance validity measure and were only kept in the study sample if they passed all administered effort measures. The remaining four patients received abbreviated clinical batteries with three of them passing other embedded performance validity measures (e.g., Reliable Digit Span < 7; Finger Tapping dominant and non-dominant combined average taps < 63). All included patients passed available performance validity measures. The fourth patient was determined to have adequate levels of effort based on average or better scores on all tests administered and clinical judgment at the time of testing. No patient was known to be involved in litigation and none was evaluated for the determination of disability or compensation status.

Additional samples were used to explore resulting sensitivity and specificity based on the cut-points derived from the outpatient sample. The first of these samples was the control group from the initial simulation study (Johnson et al., 2009). This sample included 27 healthy college students (37% men) between the ages of 17 and 26 ($Mn = 20.2$, $SD = 2.0$). Racial/ethnicity composition was 78% Caucasian, 11% African American, 7.4% American Indian/Alaska Native, and 3.7% Asian. This sample was used to determine the specificity of the new ANAM-PVI scores and cut-points in a healthy non-clinical sample.

The second validation sample included the simulator group from the initial simulation study (Johnson et al., 2009). This sample included 28 healthy college students (46% men) instructed to simulate cognitive impairment. These participants were between the ages of 18 and 23 ($Mn = 20.3$, $SD = 1.6$). Racial/ethnicity composition was 75% Caucasian, 10.7% African American, 10.7% American Indian/Alaska Native, and 3.6% Latino/Hispanic. Based on a simulator script used by Willison and Tombaugh (2006), individuals in this group were instructed to pretend that they had experienced a mild brain injury with initial symptoms that resolved back to normal. They were also told to pretend that they were involved in a legal proceeding to determine a financial settlement for their previously acquired brain injury and that they would receive a larger settlement if they could demonstrate that they are still suffering symptoms from this brain injury. Finally, they were told that major exaggerations are easy to detect, therefore, their job is to “convince us by your performance on these tasks that you are brain injured, but do so in a believable way.” This sample was used to determine sensitivity of the new ANAM-PVI cut-points to simulated cognitive impairment.

Ancillary analyses included a group of patients ($n = 17$) recently discharged from inpatient rehabilitation for moderate to severe TBI or stroke who took the ANAM battery as part of a separate research study. This sample included 12 men/5 women between the ages of 17 and 67 ($Mn = 36.5$, $SD = 16$). Racial/ethnicity composition was 65% Caucasian, 29% African American, and 6% Latino/Hispanic. The majority of these patients had moderate to severe TBI ($n = 15$). The remaining patients had a diagnosis of stroke ($n = 2$). All required inpatient rehabilitation for their injuries and none were in a period of post-traumatic amnesia/confusion at the time of testing. Average time since injury was 10 months (range 1–49 months). All patients were known to have severe cognitive impairment in at least one cognitive domain based on neuropsychological testing conducted concurrent with ANAM testing. The average RBANS Total Index Score was 74.6 (range 50–108). Although one individual performed in the average range on the RBANS, this individual demonstrated severe executive dysfunction from a bilateral frontal lobe injury. This sample was included to examine the ability of the new ANAM-PVI cut-points to be insensitive to the effects of severe cognitive impairment.

Measure

Automated Neuropsychological Assessment Metrics (v4). The ANAM4 Core Battery is designed to aid in the assessment of general cognitive function following suspected brain injury or other cognitive insult. ANAM has a long history of use in medication trials, assessment of cognitive effects of extreme conditions, assessment of neurological disorders, military research, and sports concussion (for review, see McCaffrey & Kane, 2007). ANAM has been shown to have good construct validity with traditional tests of attention, processing speed, and working memory (Bleiberg, Kane, Reeves, Garmoe, & Halpern, 2000; Kabat, Kane, Jefferson, & DiPino, 2001). Test–retest reliability for individual tests has been shown to vary between 0.41 and 0.74 in healthy military samples (Vincent, Roebuck-Spencer, Lopez, et al., 2012) and 0.47 and 0.90 in a general community sample (Cognitive Science Research Center [CSRC], 2012). Tests with higher cognitive processing demands show better reliability, and tests of simple RT (SRT) show lower reliability, most likely due to restriction of range. The battery takes approximately 20–25 min to complete via personal computer. Brief descriptions of each test are provided in Table 2 in the sequence of administration. Detailed descriptions of ANAM tests can be found elsewhere (Vincent et al., 2008; Vincent, Roebuck-Spencer, Gilliland, et al., 2012; Vincent, Roebuck-Spencer, Lopez, et al., 2012).

Variables

Accuracy and RT Discrepancy scores from four commonly used ANAM tests (Matching to Sample [M2S], SRT, Procedural RT [PRO], and Code Substitution Learning [CDS]) were used to derive the ANAM-PVI. These tests were empirically chosen based on the observed sensitivity and specificity of each for detecting insufficient effort in the original simulator study (Johnson et al., 2009). Raw data (i.e., trial by trial data) were examined for each individual subject to create the RT discrepancy score. Individual trials with response times less than 130 ms are automatically filtered out by the ANAM software. Previous research suggests that valid response times must be at least 100 ms (typically 100–150) to account for time needed for physiological processes such as stimulus perception and motor response (Luce, 1986). These short response times are typically rare and in this sample occurred only on the SRT test. Participants, on average, demonstrated one anticipatory response out of 40 trials ($M = 1.18$, range = 0–5). RT discrepancy scores were computed for each individual by calculating the difference (in ms) between the RTs representing the 90th and 10th percentiles as described in Johnson and colleagues (2009). This difference score serves to quantify the magnitude of the discrepancy between the decision-based and automatic processes governing response times, where larger differences indicate larger discrepancies between these two processes. Although the RT discrepancy score is essentially a measure of variability, it is rooted in a component-process cognitive theory (Hohle, 1965; Luce, 1986; Madden et al., 1999; Schmiedek et al., 2007). Response accuracy was calculated as the percentage correct.

Following procedures described by Silverberg and colleagues (2007), Accuracy and RT Discrepancy scores from each of the four selected ANAM tests were converted to a common metric based on the relative infrequency of these scores in the outpatient

Table 2. ANAM4 core test descriptions

Test name	Description
Sleepiness Scale	Self-assessment of the user's level of sleepiness; modification of the Stanford Sleepiness Scale (Hoddes, Zarcone, Smythe, Phillips, & Dement, 1973)
Mood Scale	Self-assessment of the user's mood state in seven categories: Vigor, Happiness, Depression, Anger, Fatigue, Anxiety, and Restlessness
*SRT	Measures simple motor reaction time by having the user respond as quickly as possible to a target stimulus
*Code Substitution – Learning	Measures visual scanning, processing speed, attention, and learning by asking the user to compare a single symbol-digit pairing with a set of defined symbol-digit pairs presented at the top of the screen. The user is instructed to learn the symbol-digit pairing for a memory test to follow later in the battery
*Procedural Reaction Time	Measures attention and processing speed by having the user respond as quickly as possible to different sets of stimuli based on simple rules (e.g., press left mouse button if you see a 2 or 3 and right mouse button if you see a 4 or 5)
Mathematical Processing	Measures attention, basic computational skills, and working memory by asking the user to solve a single-digit arithmetic problem (e.g., “5–2 + 3 =”) involving two operations
*Matching to Sample	Measures visual spatial discrimination and working memory by presenting the user with a visual pattern for a specified period of time and then, following a brief delay, asking the user to select the previously seen pattern from two choices
Code Substitution-Delayed Memory	Measures visual recognition memory by asking the user to compare a single displayed symbol-digit pair with the previously learned symbol-digit pairs presented earlier in the test battery (i.e., during the Code Substitution-Learning Test)
SRT (R)	Identical to the earlier administered SRT test and designed to measure fatigue

Notes: Tests marked with an asterisk are those used to calculate the ANAM Performance Validity Index. ANAM = Automated Neuropsychological Assessment Metrics; SRT = Simple Reaction Time.

derivation sample. This was achieved by assigning weighted scores of 6, 5, 4, 3, 2, 1, and 0 for RT Discrepancy and Accuracy scores falling in the percentile ranges of 0, 0.1–1.9, 2–4.9, 5–8.9, 9–15.9, 16–24.9, and ≥ 25 , respectively. This resulted in eight weighted scores (two scores for each of four tests). Higher weighted scores indicate greater infrequency of these scores in a sample known to have good effort. These weighted scores were then summed to create the ANAM-PVI, with resulting values ranging between 0 and 48. Higher ANAM-PVI scores indicate greater overall infrequency and a higher likelihood of atypical performance compared with individuals providing good effort.

Data Analysis Plan

ANAM-PVI scores were calculated for the derivation sample (outpatients) and validation samples (college controls and simulators). A one-way analysis of variance was conducted to evaluate differences between the ANAM-PVI scores for each of these samples. Follow-up tests were conducted to evaluate pairwise differences among the means using the Games–Howell post hoc test due to the unequal variances among the groups. Effect sizes were calculated using Cohen's *d*. Receiver operator characteristic (ROC) curve analysis was conducted to calculate the area under the ROC curve (AUC) which represents the discriminability of the ANAM-PVI in the derivation and simulator samples. Resulting cut-points were evaluated and chosen such that the number of false positives were minimized in the derivation sample (i.e., specificity $> 90\%$). This cut-point was then examined in the additional validation samples. The resulting AUC, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (+LR), and odds ratio (OR) were calculated for each of the samples. Finally, an ancillary analysis examined specificity of ANAM-PVI cut-points in a neurorehabilitation sample known to have severe and global levels of cognitive impairment.

Results

ANAM-PVI scores differed across groups, $F(2, 112) = 40.8, p < .0001, \text{adj. } R^2 = .41$. Follow-up tests revealed that the outpatient ANAM-PVI scores were higher than the control group ($p = .001; d = 0.50$) and significantly lower than the simulator group (both $p < .0001; d = 1.75$) with a much larger effect size seen with the latter group comparison. Further, ANAM-PVI scores in the simulator group were significantly higher than the control group ($p < .0001; d = 2.25$; Fig. 1).

The ROC curve analysis indicated excellent discriminability of the ANAM-PVI in the outpatient and simulator samples, $\text{AUC} = 0.858 (\text{SE} = 0.046), p < .0001$. The OR for the ANAM-PVI in detecting invalid performance was 15.6. The optimal cut-point derived from the ROC curve analysis (using Youden's Index) was an ANAM-PVI score ≥ 5 (classification accuracy = 78.4%; Fig. 2). However, an ANAM-PVI score of ≥ 10 is required to achieve the minimum a priori required specificity of 90% in the outpatient sample, resulting in a lowered sensitivity of 68% to detect simulators and an overall classification accuracy of 83.0%.

An ancillary analysis, examined the specificity of the ANAM-PVI to correctly identify patients with known severe and global cognitive impairment. The data-optimized ANAM-PVI cut-point of 5 had extremely low specificity for severe cognitive impairment (17%). Likewise, specificity was still only 47% with an ANAM-PVI of 10. Qualitative examination of the data revealed that an ANAM-PVI cut-point ≥ 14 allowed for the highest specificity in the severely impaired group (70%) while still maintaining a sensitivity to detect simulators of greater than 50%.

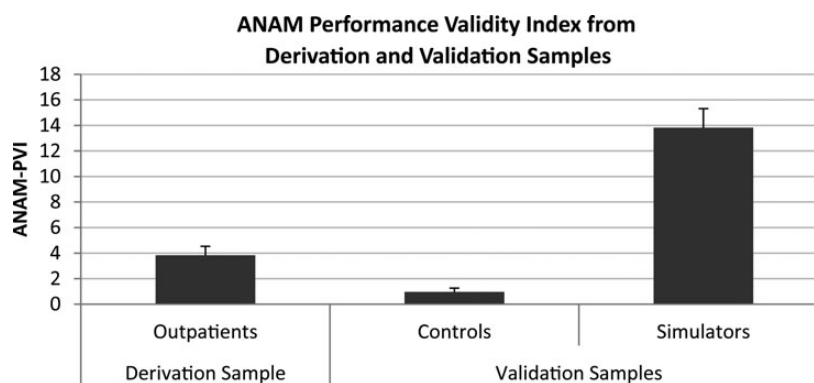


Fig. 1. ANAM-PVI scores from derivation and validation samples. ANAM-PVI = Automated Neuropsychological Assessment Metrics Performance Validity Index. *Validation samples differ from outpatient group at $p < .05$.

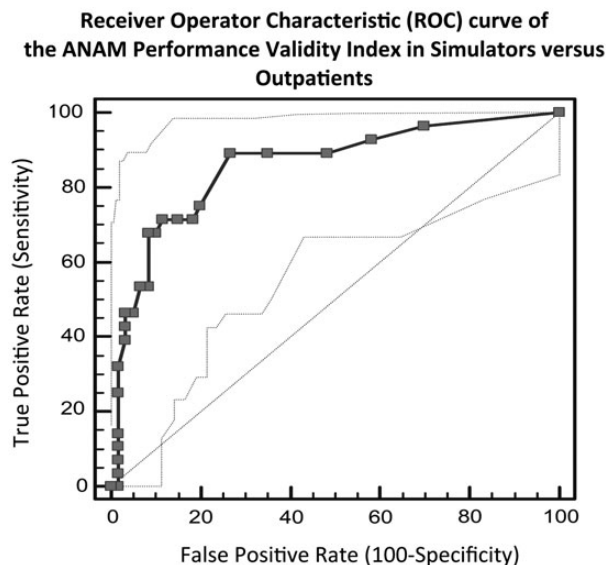


Fig. 2. ROC curve for simulators versus outpatients.

Table 3 displays the actual number of subjects within each group at or above a given ANAM-PVI score. Sensitivity refers to the cumulative proportion of simulators at or above a given ANAM-PVI score, and specificity refers to the cumulative proportion of effortful patients below a given ANAM-PVI score. As expected, there is a trade-off between specificity for severe impairment and sensitivity to detect poor effort in simulators for any cut-score selected.

Sensitivity, specificity, PPV, NPV, +LR, and OR are presented in Table 4 for each of the samples using the cut-scores described above. As expected, specificity rates in the derivation sample were lower than in the control group, but higher than the neurorehabilitation group, suggesting that sample-specific cut-points should be considered based on patient population being tested.

Discussion

This study presents the first clinical data on an embedded performance validity measure within the ANAM4-Core Battery. The ANAM-PVI combines measures of accuracy and response speed across four of the most commonly used ANAM tests: SRT, CDS, PRO, and M2S. The ANAM-PVI was created by weighting performance according to its relative infrequency within a clinical sample with known cognitive impairment and adequate performance validity (Silverberg et al., 2007), with higher scores indicating more atypical performance. Sensitivity and specificity of ANAM-PVI scores across samples with varying levels of valid performance and cognitive impairment were examined.

ANAM-PVI scores differed across samples with the highest (i.e., worst, least valid) scores observed in a simulator group (ANAM-PVI = 13.82) and the lowest scores in a healthy college sample with known adequate effort (ANAM-PVI = 0.96). An outpatient brain injury clinic sample with known adequate performance validity demonstrated significantly higher ANAM-PVI scores than the controls (ANAM-PVI = 3.83) but still remained well below that of the simulators. These findings provide good support for a wide range of performance on the ANAM-PVI with higher values seen simulator groups.

The primary purpose of this paper was to examine clinically derived ANAM-PVI cut-points using an outpatient sample more representative of a “typical” neuropsychology practice with the ultimate goal of avoiding false-positive identification of invalid performance in individuals with true cognitive impairment. This outpatient sample was comprised of individuals with a wide range of cognitive functioning, with Full-Scale IQs (FSIQs) ranging from the mildly impaired to the very superior range (FSIQ range 76–155; Mn = 102.6). Potential false positives in groups with more severe levels of cognitive impairment are always of concern. Thus, a sample of patients with moderate to severe acquired brain injury who were recently discharged from inpatient rehabilitation was also examined in an ancillary analysis. The mean ANAM-PVI within this group was higher than that of the control and outpatient groups (ANAM-PVI = 10.9) but was still significantly lower than that seen in the simulator group. However, suboptimal effort cannot be ruled out in this group because independent performance validity measures were not administered.

ROC curve analysis revealed excellent ability of the ANAM-PVI to discriminate outpatients from simulators with an OR of 15.6 at an empirically derived ANAM-PVI cut point of ≥ 5 . This means that the odds for meeting criteria for poor effort

Table 3. Sensitivity and specificity for a range of ANAM-PVI scores

ANAM-PVI	Simulators (<i>n</i> = 28)		Outpatient (<i>n</i> = 60)		Neurorehabilitation (<i>n</i> = 17)		Controls (<i>n</i> = 27)	
	<i>n</i>	Sensitivity	<i>n</i>	Specificity	<i>n</i>	Specificity	<i>n</i>	Specificity
0	28	1.0	60	0.000	17	0.000	27	0.000
1	27	0.964	42	0.300	17	0.000	11	0.593
2	26	0.929	35	0.417	16	0.059	6	0.778
3	25	.893	29	0.517	15	0.118	5	0.815
4	25	0.893	21	0.65	15	0.177	3	0.889
5 ^a	25	0.893	16	0.733	14	0.177	1	0.963
6	21	0.75	12	0.80	14	0.294	0	1.0
7	20	0.714	11	0.817	12	0.294		
8	20	0.714	9	0.85	11	0.353		
9	20	0.714	7	0.883	11	0.471		
10 ^b	19	0.679	6	0.9	9	0.471	0	1.0
11	19	0.679	5	0.917	9	0.529		
12	15	0.536	5	0.917	8	0.529		
13	15	0.536	4	0.933	7	0.588		
14 ^c	15	0.536	4	0.933	5	0.706	0	1.0
15	13	0.464	3	0.95	3	0.824		
16	13	0.464	2	0.967	3	0.882		
17	12	0.429	2	0.967	3	0.882		
18	12	0.429	2	0.967	2	0.882		
19	11	0.393	12	0.967	1	0.941		
20	9	0.321	1	0.983	1	0.941		
21	7	0.25	1	0.983	1	0.941		
22	4	0.143	1	0.983	1	0.941		
23	3	0.107	1	0.983	1	0.941		
24	3	0.107	1	0.983	1	0.941		
25	3	0.107	1	0.983	1	0.941		
26	2	0.071	1	0.983	1	0.941		
27	2	0.071	1	0.983	1	0.941		
28	1	0.036	1	0.983	1	0.941		
29	0	0.000	1	0.983	1	0.941		
30			1	0.983	0	1.0		
31			1	0.983				
32			1	0.983				
33			0	1.0				

Note: ANAM-PVI = Automated Neuropsychological Assessment Metrics Performance Validity Index.

^aOptimal cut-point as determined by ROC analysis.

^bCut-point chosen to maintain a minimum of 90% specificity in the outpatient sample.

^cMaximum cut-point recommended to avoid false-positive errors in significantly impaired populations.

Table 4. Results of ROC curve analyses at ANAM-PVI cut-points of ≥ 10 and ≥ 14 in comparison to simulators

	Specificity	Positive predictive value	Negative predictive value	Area under the curve (95% CI)	Positive likelihood ratio	Odds ratio
ANAM-PVI cut-point = 10; sensitivity = 0.68						
Outpatients (<i>n</i> = 60)	0.90	0.76	0.86	0.86 (0.77–0.92)	6.8	23.3
Controls (<i>n</i> = 27)	1.0	0.97 ^a	0.67 ^a	0.95 (0.86–0.99)	27 ^a	55.2 ^a
Neurorehabilitation (<i>n</i> = 17)	0.47	0.68	0.47	0.61 (0.45–0.75)	1.28	1.9
All groups (<i>n</i> = 104)	0.86	0.56	0.91	0.84 (0.77–0.90)	4.85	30.3
ANAM-PVI cut-point = 14; sensitivity = 0.54						
Outpatients (<i>n</i> = 60)	0.93	0.79	0.81	0.86 (0.77–0.92)	8.04	15.6
Controls (<i>n</i> = 27)	1.0	0.97 ^a	0.67 ^a	0.95 (0.86–0.99)	27 ^a	55.2 ^a
Neurorehabilitation (<i>n</i> = 17)	0.71	0.75	0.48	0.61 (0.45–0.75)	1.82	2.8
All groups (<i>n</i> = 104)	0.91	0.63	0.88	0.84 (0.77–0.90)	6.19	11.9

Note: ANAM-PVI = Automated Neuropsychological Assessment Metrics Performance Validity Index.

^aDue to false-positive count of zero in control sample, approximations of positive predictive value, negative predictive value, positive likelihood ratio, and odds ratio were calculated by adding 0.5 to all counts.

among subjects with poor effort is 15.6 higher than the odds for meeting criteria among subjects demonstrating good effort. At this cut-point, sensitivity to detect simulators was very good at 89% (PPV = 0.79, NPV = 0.81) with even better specificity for a healthy college sample at 96% (PPV = 0.97, NPV = 0.67). It should be noted that PPV and NPV presented here are specific to these samples with a base rate of 31.8 for poor performance validity between outpatients and simulators and thus would not be expected to correspond to other situations with differing base rates.

Although the ANAM-PVI cut-point of ≥ 5 may be appropriate in cognitively-healthy samples, it has the potential for higher than desired false-positive errors in groups with known cognitive impairment. At this cut-point, specificity was only 73% in the outpatient clinic sample. In order to avoid false positives and improve detection of good effort in the presence of true cognitive impairment, data were re-examined using an a priori specificity level of at least 90% for the clinical sample. There is strong support for setting cut-points that maximize specificity, even at the expense of sensitivity, which can be affected by a variety of factors including transparency of the test detection method, individual strategies to feign impairment, and potential coaching (Greve & Bianchini, 2004). As Greve and Bianchini (2004) argue, setting cut-offs to maximize sensitivity will result in unacceptably large numbers of false-positive errors (low specificity) and will ultimately lessen the value of your indicator. To assist the reader, Table 3 provides specificity levels across multiple cut-points with associated sensitivity levels to illustrate the trade-offs between sensitivity and specificity across groups of interest.

Restriction to this a priori specificity rate of at least 90% within the outpatient sample resulted in a more conservative ANAM-PVI cut-point of 10. At this cut-point, sensitivity to detect simulators decreased to 68%, which is generally consistent with or better than the on average 50% sensitivity reported for that of a wide range of other embedded effort measures (Larrabee, 2008). However, because this cut-point continued to result in a large number of potential false-positive errors in a sample of patients with known severe cognitive impairment (specificity = 47%), clinicians should use a cut-point of 19 to achieve at least 90% specificity for similar groups, with sensitivity further falling to just under 40%.

The primary limitations of this study were small criterion group sample sizes and the use of simulators to determine the sensitivity of the ANAM-PVI to detect poor effort, given that performance patterns in simulators may differ in unpredictable ways from those of individuals with real-world incentives or secondary gain to perform poorly on testing. Simulators, particularly college samples as used in this study, may lack the external motivation to feign poor performance in a convincing way, may differ academically and intellectually from the average head injury survivor, may be less deceived by less transparent methods of some PVTs, and may utilize more sophisticated means of deception in their performance (Bianchini, Mathias, & Greve, 2001). A second limitation is that the samples included within this study were drawn retrospectively which precluded the ability to control which and how many standalone concurrent performance validity measures were administered. Further, retrospective sampling did not allow for consistent sampling of potentially important variables such as injury severity or level of cognitive impairment and their potential impact on the ANAM-PVI. Future studies should explore the sensitivity/specificity of the ANAM-PVI and its concordance with standalone and embedded PVTs validated on samples with high incentives to feign cognitive impairment or with known base rates of poor effort. Additionally, the derivation sample was drawn from a primarily male military outpatient population which may limit generalizability to civilian samples. Further, the simulator group was composed of primarily young adults. The potential impact of sex differences and age on the ANAM-PVI should be studied more closely in future studies. A third limitation is that effort was not specifically tested and confirmed in the rehabilitation sample making it difficult to determine whether lower specificity in this group was attributable to the effect of greater cognitive compromise on the ANAM-PVI or the possibility of decreased effort in this group. Future studies should prospectively recruit new samples and co-administer the ANAM-PVI with multiple concurrent well-validated PVTs to cross-validate these findings and should further explore potential differential performance patterns across injury severity and level of cognitive compromise. Finally, the potential effect of cognitive fatigue when the ANAM-PVI is administered at the end of a time-intensive comprehensive neuropsychological battery should be examined.

In conclusion, the current study provides the first step toward the clinical validation of an embedded effort measure available within the ANAM4-Core battery. These initial data support the potential of the ANAM-PVI to discriminate between individuals with valid versus invalid performance on the ANAM battery. An empirically derived cut-point of ≥ 5 , maximizing sensitivity/specificity resulted in optimal sensitivity to detect simulators but resulted in a 27% chance of false positives in a clinical sample. When specificity was constrained to be 90% within a clinical sample, sensitivity was lower, as expected, but was consistent with that reported for other embedded effort measures (Larrabee, 2008), which highlights the importance for clinicians to choose sample-specific cut-points most appropriate for the population they are working with (e.g., use a lower, more stringent cut-point for examinees without medically documented complicated head injury or in groups with suspected motive to exaggerate symptoms). In contrast, clinicians working with patients expected to have severe levels of cognitive impairment may opt to use a higher cut-point to maximize specificity and avoid false-positive errors. Cut-points should be determined with the following information in mind: (a) the type of patient being assessed, (b) the purpose of the evaluation, and (c) the trade-off between false-negative and false-positive errors.

Funding

This work was supported in part by the U.S. Army Medical Research Acquisition Activity, 1054 Patchel Street, Fort Detrick, MD 21702-5012, Project W81XWH-09-1-0707.

Conflict of Interest

The University of Oklahoma (OU) holds the exclusive license for the Automated Neuropsychological Assessment Metrics (ANAM). The Cognitive Science Research Center (formerly C-SHOP) at OU is responsible for research and development of ANAM. Vista Life Sciences holds the exclusive license for ANAM commercialization, distribution, and sales. KG has standard university royalty agreements for the sale of ANAM. No other authors of this manuscript received funds or salary support from ANAM sales.

References

- Allen, L. M., Green, P., Cox, D. R., & Conder, R. L. (2006). *CARB: Computerized Assessment of Response Bias: A manual for computerized administration, reporting, and interpretation of CARB running under CogShell™ assessment environment*. Durham, NC: CogniSyst.
- Bianchini, K. J., Mathias, C. W., & Greve, K. W. (2001). Symptom Validity Testing: A critical review. *The Clinical Neuropsychologist*, *15* (1), 19–45.
- Bleiberg, J., Cernich, A. N., Cameron, K. L., Sun, W., Peck, K., Uhorchak, J., et al. (2004). Duration of cognitive impairment following sports concussion. *Neurosurgery*, *54* (4), 1–6.
- Bleiberg, J., Kane, R. L., Reeves, D. L., Garmoe, W. S., & Halpern, E. (2000). Factor analysis of computerized and traditional tests used in mild brain injury research. *The Clinical Neuropsychologist*, *14* (3), 287–294.
- Bolan, B., Foster, J. K., Schmand, B., & Bolan, S. (2002). A comparison of three tests to detect feigned amnesia: The effects of feedback and the measurement of response latency [Comparative Study]. *Journal of Clinical and Experimental Neuropsychology*, *24* (2), 154–167. doi:10.1076/jcen.24.2.154.1000.
- Boone, K. B. (2009). The need for continuous and comprehensive sampling of effort/response bias during neuropsychological examinations. *Clinical Neuropsychology*, *23* (4), 729–741. doi:904732228.
- Boone, K. B., Lu, P., & Wen, J. (2005). Comparison of various RAVLT scores in the detection of noncredible memory performance [Controlled Clinical Trial]. *Archives of Clinical Neuropsychology*, *20* (3), 301–319. doi:10.1016/j.acn.2004.08.001.
- Bryan, C., & Hernandez, A. M. (2012). Magnitudes of decline on automated neuropsychological assessment metrics subtest scores relative to predeployment baseline performance among service members evaluated for traumatic brain injury in Iraq. *Journal of Head Trauma Rehabilitation*, *27* (1), 45–54. doi:10.1097/HTR.0b013e318238f146.
- 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 [Review]. *Archives of Clinical Neuropsychology*, *20* (4), 419–426. doi:10.1016/j.acn.2005.02.002.
- Cercy, S. P., Schretlen, D., & Brandt, J. (1997). Simulated amnesia and the pseudo-memory phenomena. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 85–107). New York: Guilford Press.
- Cognitive Science Research Center (CSRC). (2012). *ANAM4 Core: Administration manual*. Norman, OK: University of Oklahoma.
- Coldren, R. L., Russell, M. L., Parish, R. V., Dretsch, M., & Kelly, M. P. (2012). The ANAM lacks utility as a diagnostic or screening tool for concussion more than 10 days following injury. *Military Medicine*, *177* (2), 179–183.
- Dunn, T. M., Shear, P. K., Howe, S., & Ris, M. D. (2003). Detecting neuropsychological malingering: Effects of coaching information. *Archives of Clinical Neuropsychology*, *18*, 121–134.
- Friedl, K. E., Grate, S. J., Proctor, S. P., Ness, J. W., Lukey, B. J., & Kane, R. L. (2007). Army research needs for automated neuropsychological tests: Monitoring soldier health and performance status. *Archives of Clinical Neuropsychology*, *22* (SI), S7–S14.
- Greve, K. W., & Bianchini, K. J. (2004). Setting empirical cut-offs on psychometric indicator of negative response bias: A methodological commentary with recommendations. *Archives of Clinical Neuropsychology*, *19*, 533–541. doi:10.1016/j.acn.2003.08.002.
- Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., & Conference, P. (2009). American Academy of Clinical Neuropsychology Consensus Conference Statement on the neuropsychological assessment of effort, response bias, and malingering [Consensus Development Conference]. *Clinical Neuropsychology*, *23* (7), 1093–1129. doi:10.1080/13854040903155063.
- Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., & Dement, W. C. (1973). Quantification of sleepiness: A new approach. *Psychophysiology*, *10* (4), 431–436.
- Hohle, R. H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology*, *69*, 382–386.
- Johnson, D. R., Gilliland, K., & Vincent, A. S. (2009). Automated Neuropsychological Assessment Metrics (ANAM)'s measure of invalid scores: A simulator study. Paper presented at the National Academy of Neuropsychology, New Orleans, LA.
- Kabat, M. H., Kane, R. L., Jefferson, A. L., & DiPino, R. K. (2001). Construct validity of selected Automated Neuropsychological Assessment Metrics (ANAM) battery measures. *The Clinical Neuropsychologist*, *15* (4), 498–507.
- Kane, R. L., Roebuck-Spencer, T., Short, P., Kabat, M., & Wilken, J. (2007). Identifying and monitoring cognitive deficits in clinical populations using Automated Neuropsychological Assessment Metrics (ANAM) tests. *Archives of Clinical Neuropsychology*, *22* (Suppl. 1), S115–S126. doi:10.1016/j.acn.2006.10.006.
- Kelly, M. P., Coldren, R. L., Parish, R. V., Dretsch, M. N., & Russell, M. L. (2012). Assessment of acute concussion in the combat environment. *Archives of Clinical Neuropsychology*, *27* (4), 375–388.
- Kim, N., Boone, K. B., Victor, T., Lu, P., Keatinge, C., & Mitchell, C. (2010). Sensitivity and specificity of a digit symbol recognition trial in the identification of response bias. *Archives of Clinical Neuropsychology*, *25* (5), 420–428. doi:10.1093/arclin/acq040.

- Larrabee, G. J. (2007a). Identification of malingering by pattern analysis on neuropsychological tests. In G. J. Larrabee (Ed.), *Assessment of malingered neuropsychological deficits* (pp. 80–99). New York: Oxford University Press.
- Larrabee, G. J. (Ed.). (2007b). *Assessment of malingered neuropsychological deficits*. New York: Oxford University Press.
- Larrabee, G. J. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *Clinical Neuropsychology*, 22 (4), 666–679. doi:782115341.
- Lowe, M., Harris, W., Kane, R. L., Banderet, L., Levinson, D., & Reeves, D. (2007). Neuropsychological assessment in extreme environments. *Archives of Clinical Neuropsychology*, 22 (Suppl. 1), S89–S99.
- Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.
- Luethcke, C. A., Bryan, C. J., Morrow, C. E., & Isler, W. C. (2011). Comparison of concussive symptoms, cognitive performance, and psychological symptoms between acute blast-versus nonblast-induced mild traumatic brain injury. *Journal of the International Neuropsychological Society*, 17 (1), 36–45. doi:10.1017/S1355617710001207.
- Madden, D. J., Gottlob, L. R., Denny, L. L., Turkington, T. G., Provenzale, J. M., Hawk, T. C., et al. (1999). Aging and recognition memory: Changes in regional cerebral blood flow associated with components of reaction time distributions. *Journal of Cognitive Neuroscience*, 11 (5), 511–520.
- McCaffrey, R. J., & Kane, R. L. (Eds.). (2007). DoD contributions to computerized neurocognitive assessment: The ANAM test system [Special Issue]. *Archives of Clinical Neuropsychology*, 22S1.
- McDiarmid, M. A., Hooper, F. J., Squibb, K., McPhaul, K., Engelhardt, S. M., Kane, R., et al. (2002). Health effects and biological monitoring results of Gulf War veterans exposed to depleted uranium. *Military Medicine*, 167 (2 Suppl.), 123–124.
- Rahill, A. A., Weiss, B., Morrow, P. E., Frampton, M. W., Cox, C., Gibb, R., et al. (1996). Human performance during exposure to toluene. *Aviation, Space, and Environmental Medicine*, 67 (7), 640–647.
- Reedy, S. D., Boone, K. B., Cottingham, M. E., Glaser, D. F., Lu, P. H., Victor, T. L., et al. (2013). Cross validation of the Lu and colleagues (2003) Rey-Osterrieth Complex Figure Test effort equation in a large known-group sample. *Archives of Clinical Neuropsychology*, 28 (1), 30–37. doi:10.1093/arclin/acs106.
- Rees, L. M., Tombaugh, T. N., Gansler, D. A., & Moczynski, N. P. (1998). Five validation experiments on the Test of Memory Malingering (TOMM). *Psychological Assessment*, 10, 10–20.
- Reicker, L. I. (2008). The ability of reaction time tests to detect simulation: An investigation of contextual effects and criterion scores [Controlled Clinical Trial]. *Archives of Clinical Neuropsychology*, 23 (4), 419–431. doi:10.1016/j.acn.2008.02.003.
- Rose, F. E., Hall, S., & Szalda-Petree, A. D. (1995). Portland Digit Recognition Test-Computerized: Measuring response latency improves the detection of malingering. *Clinical Neuropsychologist*, 9, 124–134.
- Schmiedek, F., Oberauer, K., Wilhelm, O., Suss, H. M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence [Research Support, Non-U.S. Gov't]. *Journal of Experimental Psychology: General*, 136 (3), 414–429. doi:10.1037/0096-3445.136.3.414.
- Silverberg, N. D., Wertheimer, J. C., & Fichtenberg, N. L. (2007). An effort index for the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS). *Clinical Neuropsychology*, 21 (5), 841–854. doi:769512323.
- Sim, A., Terryberry-Spohr, L., & Wilson, K. R. (2008). Prolonged recovery of memory functioning after mild traumatic brain injury in adolescent athletes. *Journal of Neurosurgery*, 108 (3), 511–516. doi:10.3171/JNS/2008/108/3/0511.
- Slick, D. J., Hopp, G., Strauss, E., & Thompson, G. B. (2005). *VSVT: Victoria Symptom Validity Test professional manual*. Lutz, FL: Psychological Assessment Resources.
- Slick, D. J., Sherman, E. M., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist*, 13 (4), 545–561. doi:10.1076/1385-4046(199911)13:04;1-Y;FT545.
- Suhr, J. A., & Gunstad, J. (2007). Coaching and malingering: A review. In G. J. Larrabee (Ed.), *Assessment of malingered neuropsychological deficits* (pp. 287–311). New York: Oxford University Press.
- Tan, J. E., Slick, D. J., Strauss, E., & Hultsch, D. F. (2002). How'd they do it? Malingering strategies on symptom validity tests. *The Clinical Neuropsychologist*, 16 (4), 495–505. doi:10.1076/clin.16.4.495.13909.
- United States House of Representatives H.R. 4986. (2008). National Defense Authorization Act for Fiscal Year 2008, Sec. 1618, Comprehensive plan on prevention, diagnosis, mitigation, treatment, rehabilitation of, and research on, traumatic brain injury, post traumatic stress disorder, and other mental health conditions in members of the armed forces. Retrieved January 6, 2010, from <http://thomas.loc.gov/cgi-bin/query/C?R110:/temp/~r110itMIII>.
- Van Gorp, W. G., Humphrey, L. A., Kalechstein, A., Brumm, V. L., McMullen, W. J., et al. (1999). How well to standard clinical neuropsychological tests identify malingering? A preliminary analysis. *Journal of Clinical and Experimental Neuropsychology*, 21, 245–250.
- Vincent, A. S., Bleiberg, J., Yan, S., Ivins, B., Reeves, D. L., Schwab, K., et al. (2008). Reference data from the automated Neuropsychological Assessment Metrics for use in traumatic brain injury in an active duty military sample. *Military Medicine*, 173 (9), 836–852.
- Vincent, A. S., Roebuck-Spencer, T. M., Gilliland, K., & Schlegal, R. (2012). Automated Neuropsychological Assessment Metrics (v4) Traumatic Brain Injury Battery: Military normative data. *Military Medicine*, 177, 256–269.
- Vincent, A. S., Roebuck-Spencer, T. M., Lopez, M., Twillie, D., Logan, B., Schlegal, R., et al. (2012). The effects of deployment on cognitive functioning. *Military Medicine*, 177, 248–269.
- Wilken, J. A., Kane, R., Sullivan, C. L., Wallin, M., Usiskin, J. B., Quig, M. E., et al. (2003). The utility of computerized neuropsychological assessment of cognitive dysfunction in patients with relapsing-remitting multiple sclerosis. *Multiple Sclerosis*, 9 (2), 119–127.
- Wilken, J. A., Sullivan, C. L., Lewandowski, A., & Kane, R. L. (2007). The use of ANAM to assess the side-effect profiles and efficacy of medication. *Archives of Clinical Neuropsychology*, 22 (Suppl. 1), S127–S133. doi:10.1016/j.acn.2006.10.007.
- Willison, J., & Tombaugh, T. N. (2006). Detecting simulation of attention deficits using reaction time tests. *Archives of Clinical Neuropsychology*, 21 (1), 41–52. doi:10.1016/j.acn.2005.07.005.
- Woodhouse, J., Heyanka, D. J., Scott, J., Vincent, A. S., Roebuck-Spencer, T. M., Domboski, K., et al. (2013). Efficacy of the ANAM4 General Neuropsychological Screening battery (ANAM4 GNS) for detecting neurocognitive impairment in a mixed clinical sample. *The Clinical Neuropsychologist*, 27 (3), 376–385.