

## A review of the validity of computerized neurocognitive assessment tools in mild traumatic brain injury assessment

Computerized neurocognitive assessment tools (NCATs) offer potential advantages over traditional neuropsychological tests in postconcussion assessments. However, their psychometric properties and clinical utility are still questionable. The body of research regarding the validity and clinical utility of NCATs suggests some support for aspects of validity (e.g., convergent validity) and some ability to distinguish between concussed individuals and controls, though there are still questions regarding the validity of these tests and their clinical utility, especially outside of the acute injury timeframe. In this paper, we provide a comprehensive summary of the existing validity literature for four commonly used and studied NCATs (automated neuropsychological assessment metrics, CNS vital signs, cogstate and immediate post-concussion and cognitive testing) and lay the groundwork for future investigations.

First draft submitted: 28 September 2016; Accepted for publication: 19 December 2016; Published online: 30 January 2017

**Keywords:** computerized cognitive testing • computerized neurocognitive assessment tools • computerized neurocognitive test • literature review • psychometric • neuropsychological assessment • validity

The measurement of cognitive functioning via neuropsychological (NP) testing is an important component of assessment after mild traumatic brain injury (mTBI), also known as concussion. A consensus statement on concussion in sport [1] concluded that such testing provides valuable information when evaluating a person with mTBI. The US military also mandates that service members are administered NP assessment to detect cognitive impairment associated with mTBI [2].

Traditional NP assessments are typically comprised of well-established measures with large normative databases and demonstrate evidence of adequate psychometric properties (i.e., reliability and validity). However, these tests are usually administered in a one-on-one format by a trained professional with paper, pencil and stopwatch, and require interpretation by a neuropsychologist. This can make

them expensive and time intensive, not feasible for assessing large groups (e.g., athletic teams, service members) or using on the sideline or in combat settings. Over the past few decades, alternatives to traditional NP assessment batteries have emerged in the form of computerized neurocognitive assessment tools (NCATs).

NCATs offer several potential logistical advantages over traditional NP tests. They can be much less time consuming and do not require administration by a testing specialist. Scoring is automated, and test performance can be easily generated into a summary report for interpretation or an electronic spreadsheet for statistical analysis. Furthermore, NCATs allow for cognitive assessment to be obtained in geographic areas where traditional NP services are limited. They are easier to use for obtaining baseline tests (e.g., preseason, pre-deployment) that can be used for comparison

Jacques P Arrieux<sup>\*1,2</sup>,  
Wesley R Cole<sup>1,2</sup>  
& Angelica P Ahrens<sup>1,2</sup>

<sup>1</sup>Womack Army Medical Center, Fort Bragg, NC, USA

<sup>2</sup>Defense & Veterans Brain Injury Center (DVBIC), Fort Bragg, NC, USA

\*Author for correspondence:  
[Jacques.p.arrieux.ctr@mail.mil](mailto:Jacques.p.arrieux.ctr@mail.mil)

to assessment after concussion, which can be especially advantageous where examinees may have conditions that prevent comparison to normative reference groups (i.e., abnormal cognitive development, ADHD, among others) [3,4]. Also, the computerized nature of NCATs makes it possible to administer alternative forms of a test with numerous combinations of test stimuli, which mitigate practice effects, and allow for multiple administrations in a short amount of time to track recovery after injury [3]. Moreover, being computerized allows for more accurate measurement of reaction time (RT), possibly making NCATs more sensitive to subtle cognitive effects [5].

Despite potential advantages, NCATs are not without limitations, as discussed by Echemendia *et al.* [4]. Specifically, alternate forms might not be equivalent, computers settings can cause erroneous RT measurement, there are differences between administering to groups (as is often done for baseline administrations) and to individuals (as is often done postinjury), and the tests are marketed to professionals (e.g., athletic trainers) who may have little or no training in cognitive testing. Additionally, one of the most important limitations of NCATs is that the psychometric properties are not fully established. Although NCATs have gained momentum as a tool in the management of mTBI, particularly in military and athletic settings, commonly used NCATs have not undergone the same level of validation as many traditional tests. According to a Joint Position Paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology [3], though NCATs may seem to be analogous to traditional NP tests, there are important differences between them that need to be explored. Specifically, modifications of existing measures warrant investigations of the new tests' psychometric properties, such as validity [6].

This manuscript will summarize and evaluate the existing literature regarding the validity of four NCATs commonly used for both clinical and research purposes: Automated Neuropsychological Assessment Metric (ANAM), CNS-Vital Signs (CNS-VS), Axon/CogState/CogSport (CogState) and Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT; please see [Table 1](#) for a description of each NCAT). The tests that will be covered are commonly used in research and clinical settings. Specifically, CNS-VS has been used in several clinical trials and can be billed to medical insurance, ANAM is required for US military Service Members, CogState is commonly used in Australian athletic settings and ImPACT has been approved by the US FDA to detect cognitive deficits following mTBI. Interestingly, to date, these measures have not generated adequate evidence of validity,

yet they are commonly used for TBI-related assessment in sports and military settings. This summary and review will focus on the comparisons of these tests to traditional NP batteries as well as evaluations of the ability of these tests to provide clinically meaningful information regarding cognitive functioning after concussion. The existing state of the literature will be evaluated based on criteria put forth by Randolph *et al.* in a 2005 [7] review of the literature regarding NP testing after sport-related concussion (SRC). That review and those criteria are discussed below (see [Box 1](#)). This is not a systematic literature review, but is rather meant to serve as a concise summary and reference, with recommendations for future studies and considerations identified.

### Validity

Prior to discussing the literature on validity, it is important to establish what is meant by the term 'validity' (see [Box 2](#)). Validity is the most important aspect of test construction and thus is a key consideration when evaluating the clinical utility of a test. In psychometric research, validity describes whether a test measures what it claims to measure, by meeting the criteria that have been established to determine its accuracy [6]. Various models of test validity have been proposed [9,10], though in general, there are three ways to describe the validity of a test: by its content, by the construct it is purported to measure or by its ability to measure a certain criterion [11,12].

Content validity describes the relevance of the test items to the construct that is to be measured [11,12]. For example, determining if a test of attention is comprised of test items and stimuli that accurately and adequately measure attention, rather than some other construct, such as RT. Content validity is often assessed by the subjective agreement among subject matter experts, such as neuropsychologists, that the test items are relevant and appropriate for the test purpose. Construct validity describes the extent to which the measure represents the basic theoretical construct, such as cognitive functioning. It is primarily evaluated with correlations, regression or factor analysis between a domain of interest and other well-established, 'gold-standard' measures [11-13]. It is typically conceptualized as convergent and discriminant validity. Specifically, tests assessing similar constructs should have higher correlations (i.e., convergent validity) than tests of dissimilar constructs (i.e., discriminant validity). Criterion validity describes the relatedness of the measure to a specified criterion, such as a condition of interest or outcome (e.g., concussion), and is often divided into concurrent and predictive types of validity. Concurrent validity is determined by how well a test accurately identifies

Table 1. Descriptions of computerized neurocognitive assessment tools reviewed.		
Test	Subtests and cognitive construct intended to measure	Classification and summary scores
ANAM4	<ul style="list-style-type: none"> <li>• Simple Reaction Time (SRT): visuomotor processing speed, simple motor speed and attention</li> <li>• Procedural Reaction Time (PRO): processing speed, visuomotor reaction time and attention</li> <li>• Code Substitution Learning (CDS): visual scanning, visual perception, attention, associative learning and processing speed</li> <li>• Code Substitution Delayed (CDD): learning and delayed visual recognition memory</li> <li>• Mathematical Processing (MTH): basic computational skills, concentration and working memory</li> <li>• Matching to Sample (M2S): visual-spatial processing, working memory and visual recognition memory</li> <li>• SRT, Second Administration (SRT2): index of attention (i.e., reaction time [RT] and vigilance)</li> </ul>	<p>Summary scores:</p> <ul style="list-style-type: none"> <li>• Throughput (TP; number of correct responses per minute of available response time)</li> <li>• Standardized subtest TP</li> <li>– standardized composite TP</li> <li>• Composite score (standardized average TP [z-score])</li> <li>• Classification of impairment: ANAM composite <math>z \leq -1.28</math></li> </ul>
CNS-VS	<ul style="list-style-type: none"> <li>• Verbal Memory Test, Immediate (VBM): word recognition and memory, immediate and delayed recall</li> <li>• Visual Memory Test, Immediate (VIM): visual recognition and memory, immediate and delayed recall</li> <li>• Finger Tapping Test (FTT): motor speed, fine motor control</li> <li>• Symbol Digit Coding (SDC): information processing speed, complex attention, visual perceptual speed</li> <li>• Stroop Test (ST): SRT, complex reaction time, inhibition, executive skills, processing speed</li> <li>• Shifting Attention Test (SAT): executive functioning, RT</li> <li>• Continuous Performance Test (CPT): sustained attention, choice reaction time (CRT), impulsivity</li> <li>• Verbal Memory Test, Delayed (VBM): word recognition, memory and delayed recall</li> <li>• Visual Memory Test, Delayed (VIM): visual recognition, memory and delayed recall</li> </ul>	<p>Summary scores:</p> <ul style="list-style-type: none"> <li>• Neurocognitive Index (NCI)</li> <li>• Composite memory</li> <li>• Verbal memory</li> <li>• Visual memory</li> <li>• Psychomotor speed</li> <li>• RT</li> <li>• Complex attention</li> <li>• Cognitive flexibility</li> <li>• Processing speed</li> <li>• Executive function</li> <li>• Simple attention</li> <li>• Motor speed</li> </ul> <p>Composite score: IQ scale and percentile</p> <p>Classification of impairment: N/A</p>
CogState	<ul style="list-style-type: none"> <li>• Detection Task: SRT</li> <li>• Identification Task: processing speed</li> <li>• One Back Task: attention, working memory</li> <li>• One Card Learning Task: learning and recognition memory</li> </ul>	<p>– Summary scores: score for each subtest</p> <p>– Composite score: z-scores</p> <p>– Classification of impairment: –1.64 SD on at least two subtests</p> <p>CogState composite &lt; -1.64</p>
ImPACT	<ul style="list-style-type: none"> <li>• Word Memory, Immediate: verbal recognition memory</li> <li>• Design Memory, Immediate: visual recognition memory</li> <li>• X's and O's: visual working memory and visual processing/visual motor speed</li> <li>• Symbol Match: visual processing speed, learning and memory</li> <li>• Color Match: CRT and impulse control/response inhibition</li> <li>• Four Letters: working memory and visual-motor response speed</li> <li>• Word Memory, Delayed: verbal recognition memory</li> <li>• Design Memory, Delayed: visual recognition memory</li> </ul>	<p>Summary scores:</p> <ul style="list-style-type: none"> <li>• Verbal memory</li> <li>• Visual memory</li> <li>• Visual motor speed</li> <li>• RT</li> <li>• Impulse control</li> </ul> <p>Composite score:</p> <ul style="list-style-type: none"> <li>• Test-specific</li> <li>• Standardized scores and percentiles</li> </ul> <p>Classification of impairment: N/A</p>
<p>Note: Each NCAT is online or desktop based and the approximate administration time, inclusive of testing and acquisition of medical history and demographics, is 30 min, with the exception of CogState (20 min). These summaries generally capture the 'standard' battery for each NCAT; however, the batteries used in the reviewed studies may include some variations and different combinations of subtests.</p> <p>ANAM4: Automated neurocognitive assessment metric, version 4; CNS-VS: CNS-vital sign; ImPACT: Immediate post-concussion assessment and cognitive testing; IQ: Intelligence quotient; N/A: Not applicable; NCAT: Neurocognitive assessment tool.</p> <p>Adapted with permission from [8].</p>		

**Box 1. 'Randolph criteria' for proposed neuropsychological batteries.****Criterion & description**

- Establishing test–retest reliability over time intervals that are practical for this clinical purpose
- Demonstrating, through a prospective controlled study, that the battery is sensitive in detecting the effects of concussion
- Establishing validity for any novel test battery, through standard psychometric procedures employed to determine which neurocognitive abilities a new NP test is measuring
- Deriving reliable change scores, with a probability-based classification algorithm for deciding that a decline of a certain magnitude is attributable to the effects of concussion, rather than random test variance
- Demonstrating that the proposed battery is capable of detecting cognitive impairment once subjective symptoms have resolved

Note: These are criteria set forth by Randolph *et al.* [7] for both traditional and computerized NP batteries. Randolph *et al.* proposed that NP tests should first meet these criteria prior to their consideration as part of routine standard of care for sport-related concussion. NP: Neuropsychological.

a diagnosis or condition of interest when that condition is known (e.g., control vs concussion cohorts), as compared with an existing 'gold standard.' Predictive validity is determined by the test's ability to inform about some type of future outcome. It is important to note that there can be some overlap in these different types of validity, as they are not meant to conceptually represent mutually exclusive subcategories of validity, but rather describe the various ways in which validity can manifest [12].

**Past literature reviews**

There are several existing literature reviews of NCATs, including those focused on one NCAT, such as ANAM or ImpACT, for example, [14,15], as well as those focused on the broader body of NCAT literature, for example, [7–8,16]. In a comprehensive review of literature on NP testing (traditional and computerized) in SRC published from 1990 to 2004, Randolph *et al.* [7] identified several gaps with regard to the use of both traditional and computerized NP testing after SRC. They proposed five criteria that needed to be satisfied with additional research in order to consider NP testing standard of care after concussion (see Table 1). Until these requirements are satisfied, the authors sug-

gest that professionals should use NP tests, including NCATs, with caution and rely more on self-report measures and medical evaluations. In this literature review, we will place a specific focus on the validity-related 'Randolph criteria' (i.e., criteria two through five) in order to establish whether the existing research, including the research that has emerged since their review, sufficiently demonstrates that NCATs have satisfied those criteria and demonstrate adequate clinical utility.

Reviews since the Randolph *et al.* [7] paper seem to indicate that while the Randolph criteria have been partially addressed, there is still insufficient evidence that NCATs adequately satisfy the criteria. Resch *et al.* [8] conducted a similar literature review as Randolph *et al.* [7], though for research completed between 2005 and 2013, and for NCATs used primarily for SRC. The authors reported that the evidence of validity varies between NCATs, suggesting that more research is necessary in order to elucidate the relationship between NCATs and their traditional NP counterparts. Iverson and Schatz [16] conducted a literature review of NP assessment in SRC research and presented some evidence indicating that NCATs may be superior to their traditional counterparts because

**Box 2. Evidence of validity.****Content-related**

- The relevance of the test items to the construct that is to be measured
- Evaluated by the subjective agreement among subject matter experts, such as neuropsychologists, that the test items are relevant and appropriate for the test purpose

**Construct-related**

- The extent to which the measure represents the basic theoretical construct
- Evaluated with correlations, regression or factor analysis between a domain of interest and other well-established, 'gold-standard' measures

**Criterion-related**

- The relatedness of the measure to a specified criterion, such as a condition of interest or outcome (e.g., mTBI)
- Evaluated by the group differences, accuracy of diagnosis or identification of a specific condition of interest

It is important to note that there can be some overlap in these different types of validity, as they are not meant to conceptually represent mutually exclusive subcategories of validity, but rather describe the various ways in which validity can manifest [12].  
mTBI: Mild traumatic brain injury.

they can be more precise in the detection of cognitive impairment. However, all subsequent reviews, similar to Randolph *et al.*'s [7] conclusions, suggest additional research is needed in order to further validate NCATs against their traditional NP counterparts and within mTBI populations.

### Summary of literature

The sections below provide a review of the literature published to date investigating the validity of the four NCATs: ANAM, CNS-VS, CogState and ImPACT. Their utility as a neurocognitive assessment is presented in two contexts: the extent to which the test measures the same constructs as traditional NP batteries and the extent to which it provides clinically meaningful information about group membership or cognitive impairment. The reader should refer to [Tables 2–5](#) for details on the specifics of the methodology and findings of each of the studies described (as well as the full definitions of NP test-specific acronyms).

### Methods

The search for primary literature involved several search engines (e.g., Google Scholar, PubMed, EBSCOhost and ScienceDirect). Articles were chosen based on their relevance to evaluations of the validity of the four above-mentioned NCATs. Specifically, the selection criteria were based on search terms such as ANAM, CNS-VS, CogState/CogSport/Axon and ImPACT in conjunction with any number of the following terms: validity, validation, construct validity, criterion validity, convergent validity, discriminant validity, diagnosis, group differences, sensitivity/specificity, mTBI and concussion. Studies were primarily included if analyses involved either, first, comparison of performance on NCATs and traditional NP tests or second, comparison of group differences in NCAT performance between healthy controls and individuals who sustained an mTBI. Revisions to this methodology (i.e., extending relevant study populations to those with neuropsychiatric disorders) were permitted as alternative ways of capturing measures of NCAT validity when search findings were insufficient. For example, we included several articles that studied adolescent samples as many studies of SRC combined high school and college athletes. In addition, several studies of non-mTBI samples (e.g., psychiatric disorders) were also included as they often compared NCAT scores to traditional NP tests in a group of healthy controls. Since this was not a rigorous and systematic literature review, the conclusions drawn should be considered with caution by the reader. However, we believe consolidating these findings in a single review that is invaluable for those interested in knowing where the

literature currently stands in regards to the validity and clinical utility of these NCATs.

## Automated neuropsychological assessment metric

### Comparisons to traditional NP tests

Research to date has largely demonstrated that scores on ANAM and traditional NP tests have weak-to-moderate correlations (see [Table 2](#) for more details on the methodology and findings of these studies). Bleiberg *et al.* [17] concluded that ANAM measures similar cognitive constructs as traditional NP batteries in a group of healthy controls, as correlations were generally moderate. ANAM throughput (TP) scores more strongly correlated with traditional NP test scores than RT and accuracy, and ANAM Mathematical Processing (MTH) and Sternberg Memory Procedure (STN) were most closely associated with scores from the Paced Auditory Serial Addition Test in the traditional NP battery. Kabat *et al.* [18] similarly found moderate correlations in a group of veterans, the strongest of which were between the ANAM Code Substitution Learning median RT and Trail Making Test (TMT) B. However, the median RT score is not a commonly used ANAM score for clinical or research purposes. In another study, with uninjured high school athletes, MTH demonstrated the most statistically significant correlations (i.e., moderate to strong) with a traditional NP score (Digit Symbol Coding) [19]. In a comparison of healthy college students' performance on ANAM and Woodcock Johnson, Test of Cognitive Abilities-Third Edition (WJ-III), Jones *et al.* [20] found some evidence of construct validity, as ANAM moderately correlated with many of the WJ-III subtests and clusters, with the strongest correlation between the WJ-III General Intellectual Ability index (GIA) score and the ANAM Logical Reasoning (LGR) TP score. Woodhouse *et al.* [21] additionally observed several statistically significant correlations between the ANAM and Repeatable Battery for the Assessment of Neuropsychological Status (RBANS), administered to a mixed clinical sample referred for assessments of cognitive functioning. These patients were diagnosed postevaluation with a variety of neurologic and psychiatric disorders. Each of the seven ANAM subtests was correlated with RBANS performance. The strongest correlation existed between ANAM MTH TP and the RBANS Total Index score.

Studies using regression analyses investigate the ability of ANAM to predict scores on traditional NP batteries. Results have generally provided evidence for construct validity, as certain ANAM scores can significantly predict performance on traditional NP bat-

Table 2. Summary of Automated Neuropsychological Assessment Metric validity studies.

Study (year)	Study sample	Methods: tests	Methods: statistics	Results	Ref.
Bleiberg <i>et al.</i> (2000)	122 healthy high school and college students	Compared ANAM to WAIS-R, Finger Tapping, TMT A and B, CT total, PASAT, HVLIT and Stroop Color-Word Test	Pearson <i>r</i> correlation coefficients Stepwise regression PCA	Correlations ranged from -0.60 to 0.66 Stepwise regression: MTH and STN predict NP test scores PCA: identified a four-factor solution accounting for 66% of the variance	[17]
Kabat <i>et al.</i> (2001)	191 veterans referred for outpatient NP testing	Compared ANAM to WAIS-R Digit Symbol Coding, CVLT and TMT A and B	Pearson <i>r</i> correlation coefficients Stepwise regression PCA	Correlations ranged from -0.64 to 0.66. CDS best predict TMT B and WAIS-R DS PCA identified two three-factor solutions accounting for 60–62% variance	[18]
Woodard <i>et al.</i> (2002)	Uninjured high school athletes	Compared ANAM to HVLIT, COWAT, WAIS-III DS and SS, BTA	Pearson <i>r</i> correlation coefficients	Correlations ranged from -0.07 to 0.82. MTH most associated with traditional NP test scores MTH was significantly correlated with four of six traditional scores	[19]
Jones <i>et al.</i> (2008)	77 healthy college students	Compared ANAM to WJ-III	Pearson <i>r</i> correlation coefficients Stepwise regression PCA	Correlations ranged from -0.10 to 0.55. Strongest correlation was between LGR TP and GIA scores LGR and MTH best predictors of WJ-III (Efficiency Index and Visual Matching) PCA: MTH identified as a distinct factor, accounting for 19% variance	[20]
Woodhouse <i>et al.</i> (2013)	143 patients referred for outpatient NP assessments. RBANS used to group participants as impaired (n = 30) and not impaired (113)	Compared ANAM to RBANS	RBANS Total Index score ( $\leq$ 15th percentile) Pearson <i>r</i> correlation coefficients Logistic regression	Correlations ranged from 0.01 to 0.52. Strongest correlation was between MTH and RBANS total scores ANAM Composite TP score accounted for 37% variability in RBANS Total Index Score ANAM subtest scores on predicting RBANS impairment: sensitivity = 81%, specificity = 89.1%, PPV = 56.7%, NPV = 87.9%, odds ratio = 34.65	[21]
Bleiberg and Warden (2002)	US Military Academy cadets, 68 with mTBI and 16 healthy controls	Administered ANAM at baseline and then again four-times over 2-week period (first 2 weeks of recovery for mTBI group)	RCI-defined impairment Fisher's exact test	mTBI group, RCI-based decline in 19% (2 scores) and 81% (1 score) Fisher's: MTH identified 21% of mTBI group. Fisher's: mTBI failed to demonstrate practice effect (80%)	[22]

ANAM: Automated neuropsychological assessment metric; ANOVA: Analysis of variance; AUC: Area under the curve; BTA: Brief Test of Attention; CDS: Code substitution learning; COWAT: Controlled Oral Word Association Test; CT: Consonant Trigrams; CVLT: California Verbal Learning Test; DS: Digit Symbol; FFT: Finger Tapping Test; GIA: General intellectual ability index; GSC: Graded Symptom Checklist; HVLIT: Hopkins Verbal Learning Test; LGR: Logical reasoning; M25: Matching to sample; MANOVA: Multiple analysis of variance; mTBI: Mild traumatic brain injury; MTH: Mathematical processing; NP: Neuropsychological; NPV: Negative predictive value; PASAT: Paced Auditory Serial Addition Test; PCA: Principal component analysis; PPV: Positive predictive value; PRO: Procedural reaction time; RBANS: Repeatable Battery For the Assessment of Neuropsychological Status; RCI: Reliable Change Index; ROC: Receiver operating characteristics; RTD: Return to Duty; SOT: Sensory Organization Test; SRT: Simple reaction time; SS: Symbol Search; STN: Sternberg memory procedure; TMT: Trail Making Test; WAIS-III Wechsler Adult Intelligence Scale-Third Edition; WAIS-R: Wechsler Adult Intelligence Scale-Revised; WJ-III: Woodcock Johnson, Test of Cognitive Abilities-Third Edition; WRAT-R: Wide Range Achievement Test-Revised.

Table 2. Summary of Automated Neuropsychological Assessment Metric validity studies (cont.).					
Study (year)	Study sample	Methods: tests	Methods: statistics	Results	Ref.
Bleiberg et al. (1997)	Six with mTBI and six healthy controls	Administered ANAM and WAIS-R, WRAT-R, FFT, Stroop Color-Word Interference, PASAT, CVLT and COWAT ANAM administered 30-times in four sessions over a 2-week period	MANOVA	MANOVA: Group differences in three out of five of the ANAM subtests Practice effects in control group, while mTBI group's improvement was shorter and more variable	[23]
Kelly et al. (2012)	71 with acute mTBI and 166 controls, deployed in combat environments	Administered ANAM, traditional NP battery and questionnaires within 72 h of injury	ROC curve, including AUC	Difference at enrollment: SRT + PRO, AUC = 0.73, discriminant ability = 71% sensitivity = 59%, specificity = 82% Change from baseline: SRT + MTH + M2S, AUC = 0.79 Discriminant ability = 75%, sensitivity = 53%, specificity = 98%	[24]
Coldren et al. (2012)	47 with mTBI and 108 healthy controls, deployed in combat environments	Compared performance at predeployment baseline and $\leq 3$ , 5, 10+ days postinjury	Mann-Whitney U	Significant differences on five of six ANAM subtests at $\leq 3$ -day postinjury No differences at 5 and 10+ days intervals.	[25]
Norris et al. (2013)	165 soldiers with mTBI	Correlated performance on ANAM with demographic variables (age, number of blast exposures, % of prior mTBI) at 3- and 5-day postinjury	Spearman's $\rho$ statistics z-scores Kaplan-Meier plot	Spearman's $\rho$ : day 5 ANAM performance was associated with RTD rather than demographics SRT2 at day 3 most correlated with RTD time Kaplan-Meier: lowest 25% took 19 days RTD, highest 25% took 7 days RTD	[26]
Register-Milhalik et al. (2013)	38 healthy college players and 132 college athletes with mTBI	Compared ANAM and SOT and GSC. Control group was tested two-times (average of 4 days apart). mTBI group was tested two-times (preseason baseline and within 5 days following mTBI)	ROI-defined impairment sensitivity and specificity	ANAM (80% CI, average of seven subtests): sensitivity = 0.09, specificity = 0.95 When combined with SOT and GSC, ANAM (80% CI): sensitivity = 0.50, specificity = 0.96	[27]
Nelson et al. (2016)	165 mTBI and 166 healthy controls, athletes	Administered ANAM at baseline and then at 1, 8, 15 and 45 days (postinjury/baseline)	ANOVA Cohen's <i>d</i> effect sizes ROI-defined impairment ROC curve, including AUC	ANOVA: significant for 7 of 9 scores at day 1 ( <i>d</i> ranged from -0.58 to -0.84). Limited significant differences at days 8, 15 and 45. AUC: significant for 7 of 9 scores at 1-day assessment (0.63–0.73). Limited significant differences at days 8, 15 and 45	[28]

ANAM: Automated neuropsychological assessment metric; ANOVA: Analysis of variance; AUC: Area under the curve; BTA: Brief Test of Attention; CDS: Code substitution learning; COWAT: Controlled Oral Word Association Test; CT: Consonant Trigrams; CVLT: California Verbal Learning Test; DS: Digit Symbol; FFT: Finger Tapping Test; GIA: General intellectual ability index; GSC: Graded Symptom Checklist; HVLIT: Hopkins Verbal Learning Test; LGR: Logical reasoning; M2S: Matching to sample; MANOVA: Multiple analysis of variance; mTBI: Mild traumatic brain injury; MTH: Mathematical processing; NP: Neuropsychological; NPV: Negative predictive value; PASAT: Paced Auditory Serial Addition Test; PCA: Principal component analysis; PPV: Positive predictive value; PRO: Procedural reaction time; RBANS: Repeatable Battery For the Assessment of Neuropsychological Status; RCI: Reliable Change Index; ROC: Receiver operating characteristics; RTD: Return to Duty; SOT: Sensory Organization Test; SRT: Simple reaction time; SS: Symbol Search; STN: Sternberg memory procedure; TMT: Trail Making Test; WAIS-III Wechsler Adult Intelligence Scale-Third Edition; WAIS-R: Wechsler Adult Intelligence Scale-Revised; WJ-III: Woodcock-Johnson, Test of Cognitive Abilities-Third Edition; WRAT-R: Wide Range Achievement Test-Revised.

Table 3. Summary of CNS-Vital Signs validity studies.

Study (year)	Study sample	Methods: tests	Methods: statistics	Results	Ref.
Gualtieri and Johnson (2006)	144 with neuropsychiatric disorders and 36 healthy controls	Compared CNS-VS to RAVLT, WMS LM subtests, FTT, the ST, TMT B and the VF test	Pearson <i>r</i> correlation coefficients	Correlations ranged from -0.53 to 0.79 Strongest correlation was between CNS-VS SDC and WAIS DS.	[29]
Lanting et al. (2012)	50 with mTBI	Compared CNS-VS to NAB, RST, WTAR. Assessed at 6–8 weeks following injury	Pearson <i>r</i> correlation coefficients	Correlations ranged from 0.28 to 0.58 Strongest correlation was between CNS-VS Psychomotor Speed score and NAB memory index standard score ( $r = 0.58$ )	[30]
Gualtieri and Hervey (2015)	179 with psychiatric disorders	Compared CNS-VS to WAIS-III	Pearson <i>r</i> correlation coefficients Exploratory and confirmatory factor analyses Stepwise discriminant function analysis Logistic regression	Correlations ranged from 0.33 to 0.59. Strongest correlation was between CNS-VS SAT and FSIQ CNS-VS SAT and VIM scores were the only significant predictors of FSIQ	[31]
Lanting et al. (2012)	50 with mTBI and 31 with orthopedic injury	Administered CNS-VS at 6–8-week postinjury	MANOVA	No significant differences between groups	[32]
Gualtieri and Johnson (2008)	145 controls and 141 examinees with TBI separated into four groups: PCS = 13; mTBI recovered = 15; recovered from STBI = 85, unrecovered STBI = 28	Administered CNS-VS. PCS tested within 3-month postinjury mTBI tested within 12-month postinjury STBI time since injury unspecified	MANOVA Post hoc <i>t</i> -tests comparing the five groups ROC curve, including AUC	MANOVA: 18 of 28 scores were significantly different. Post hoc <i>t</i> -tests demonstrated significant differences on most CNS-VS scores between healthy controls and the four TBI cohorts, with mTBI recovered performing as well as healthy controls Post hoc <i>t</i> -tests demonstrated the STBI groups performed significantly worse than the mTBI groups AUC, group membership between injured and healthy groups: significant for psychomotor speed (0.75), NCI (0.75) and cognitive flexibility (0.71)	[33]
Dresch et al. (2015)	458 active-duty soldiers (deemed fit for duty)	Administered CNS-VS, demographic questionnaires and biomarkers 30 days before and after deployment	Cohen's <i>d</i> effect sizes	Cohen's <i>d</i> = 0.22 on pre–post-deployment comparisons Cohen's <i>d</i> = 0.40 on postdeployment comparisons, with the sample divided into 'no traumatic stress' or 'traumatic stress groups'	[34]

AUC: Area under the curve; CNS-VS: CNS-vital sign; FSIQ: Full Scale Intelligence Quotient; FTT: Finger Tapping Test; LM: Logical memory; MANOVA: Multivariate analysis of variance; mTBI: Mild traumatic brain injury; NAB: Neuropsychological assessment battery; NCI: Neurocognitive Index; PCS: Postconcussive syndrome; RAVLT: Rey Auditory Verbal Learning Test; ROC: Receiver operating characteristics; RST: Reynolds Intellectual Screening Test; SAT: Shifting attention test; SDC: Symbol digit coding; ST: Stroop Test; STBI: Severe TBI; TMT: Trail Making Test; VIM: Visual memory test; WAIS DS: Wechsler adult intelligence scale digit symbol; VF: Verbal fluency; WAIS-III: Wechsler Adult Intelligence Scale-Third Edition; WMS: Wechsler Memory Test; WTAR: Wechsler Test of Adult Reading.



**Table 4. Summary of Axon/CogState/CogSport validity studies.**

Study (year)	Study sample	Methods: tests	Methods: statistics	Results	Ref.
Makdissi et al. (2001)	240 healthy athletes. Six of original cohort sustaining acute mTBI and seven matched controls were retested	Compared CogState's SRT subtest to DSST and TMT B mTBI group tested at baseline and then at 72-h postinjury Control group retested ~34 days following baseline	Pearson <i>r</i> correlation coefficients for each group MANOVA	Correlations: Better for controls than mTBI SRT and DSST: <i>r</i> = -0.48.x SRT and TMT: <i>r</i> = 0.42 MANOVA: mTBI group had significant (36%) decline in performance on CogState SRT RT variability was significantly different with the mTBI group	[35]
Collie et al. (2003)	240 elite athletes and 60 demographically matched controls	Compared CogState to DSST and TMT-B	Pearson <i>r</i> correlation coefficients	Correlations ranged from -0.86 to 0.44 Strongest correlation was between CogState Decision-Making Speed and DSST	[36]
Schatz and Putz (2006)	30 healthy controls	Compared CogSport to TMT and WAIS-R DS. Also administered ImpACT Administered three-times over a 5-day period, at 48-h intervals	Pearson <i>r</i> correlation coefficients	Correlations ranged from -0.28 to 0.54 Strongest correlation was between CRT and TMT A and B	[37]
Maruff et al. (2009)	215 healthy controls 50 participants with cognitive impairment (mTBI, schizophrenia, AIDS Dementia Complex)	Compared CogState to GPB, TMT, SDMT, BVMT, RCFT and WMS-III SS subtest	Pearson <i>r</i> correlation coefficients ANOVA Cohen's <i>d</i> effect sizes Non-OL%	Correlations ranged from 0.49 to 0.83 ANOVA: significant differences and medium to large effect sizes for impairment (0.60 to -1.80) Non-OL%: CogState subtests could identify 53–78% impairment that was unique to mTBI alone	[38]
Collie et al. (2006)	615 Australian Rules football players completed baseline assessments 61 sustained mTBI and 84 healthy controls were retested	Compared CogState to TMT-B, DSST and standardized symptom checklist. Tests administered at baseline retested 11 days later. Symptom assessment used to categorize examinees as symptomatic or asymptomatic	ANOVA z-score change	Control group, no significant changes from baseline Asymptomatic significantly declined on one of seven scores Symptomatic group significantly declined on three of seven scores (z-score change ranged from -0.60 to -0.86)	[39]

AUC: Area under the curve; ANCOVA: Analysis of covariance; ANOVA: Analysis of variance; BVMT: Brief Visual Memory Test; CCR: Correct Classification Rate; CRT: Choice reaction time; DS: Digit Symbol; DSST: Digit Symbol Substitution Test; GPB: Grooved Pegboard; ImpACT: Immediate post-concussion assessment and cognitive testing; MANOVA: Multivariate analysis of variance; mTBI: Mild traumatic brain injury; Non-OL%: Nonoverlap statistics; RCFT: Rey Complex Figure Test; RCI: Reliable Change Index; ROC: Receiver operating characteristics; RT: Reaction time; SDMT: Symbol Digit Modalities Test; SRT: Simple Reaction Time; SS: Spatial span; TMT: Trail Making Test; WAIS-III: Wechsler Adult Intelligence Scale-Third Edition; WAIS-R: Wechsler Adult Intelligence Scale-Revised; WMS-III: Wechsler Memory Scale-Third Edition.

Table 4. Summary of Axon/CogState/CogSport validity studies (cont.).

Study (year)	Study sample	Methods: tests	Methods: statistics	Results	Ref.
Louey <i>et al.</i> (2014)	29 athletes with acute mTBI and 260 healthy control athletes	Administered CogState Compared normative and baseline comparison methods to classify cognitive impairment (subtest scores $\leq 1.65$ SD below mean)	RCI-defined impairment ANCOVA Cohen's <i>d</i> effect sizes Chi-squared analyses to determine diagnostic accuracy of impaired performance CCR	ANCOVA: significant differences between groups on all scores Cohen's <i>d</i> ranged from -0.94 to -2.95 Chi-square: baseline data, sensitivity = 96.6%, specificity = 86.9%, CCR = 87.9% Normative data, sensitivity = 69%, specificity = 91.5%, CCR = 89.3%	[40]
Gardner <i>et al.</i> (2012)	51 rugby players with acute mTBI and 41 controls	Compared CogState/CogSport to WAIS-III. Also administered ImpACT	t-tests Effect sizes Logistic regression	Significant differences between groups on four out of five CogState tests Effect sizes ranged from -0.46 to -0.88 Logistic regression: CogState improved classification accuracy 3.5% beyond demographics	[41]
Nelson <i>et al.</i> (2016)	165 athletes with mTBI and 166 healthy controls	Administered CogState at baseline and then at 1, 8, 15 and 45 days (postinjury/baseline)	ANOVA Cohen's <i>d</i> effect sizes RCI-defined impairment ROC curve, including AUC	ANOVA: significant differences for four of five scores at day 1 ( <i>d</i> ranged from -0.51 to -0.72). Limited significant differences at days 8, 15 and 45 AUC: significant for four of five scores at day 1 (0.64–0.69). No significant differences at days 8, 15 and 45	[28]

AUC: Area under the curve; ANCOVA: Analysis of covariance; ANOVA: Analysis of variance; BVMT: Brief Visual Memory Test; CCR: Correct Classification Rate; CRT: Choice reaction time; DS: Digit Symbol; DSST: Digit Symbol Substitution Test; GPB: Grooved Pegboard; ImpACT: Immediate post-concussion assessment and cognitive testing; MANOVA: Multivariate analysis of variance; mTBI: Mild traumatic brain injury; Non-OL%: Nonoverlap statistics; RCFT: Rey Complex Figure Test; RCI: Reliable Change Index; ROC: Receiver operating characteristics; RT: Reaction time; SDMT: Symbol Digit Modalities Test; SRT: Simple Reaction Time; SS: Spatial span; TMT: Trail Making Test; WAIS-III: Wechsler Adult Intelligence Scale-Third Edition; WAIS-R: Wechsler Adult Intelligence Scale-Revised; WMS-III: Wechsler Memory Scale-Third Edition.

**Table 5. Summary of Immediate Postconcussion Assessment and Cognitive Testing validity studies.**

Study (year)	Study sample	Methods: tests	Methods: statistics	Results	Ref.
Iverson <i>et al.</i> (2005)	72 athletes with acute mTBI.	Compared IMPACT to SDMT	Pearson <i>r</i> correlation coefficients EFA	Correlations ranged from -0.60 to 0.70 EFA: revealed a two-factor solution describing, first, speed/RT and second, memory	[42]
Schatz and Putz (2006)	30 healthy controls	Compared IMPACT to TMT and WAIS-R DS. Also administered CogState Administered three-times over a 5-day period, at 48-h intervals	Pearson <i>r</i> correlation coefficients	Correlations ranged from -0.506 to 0.641 Strongest correlation was between CRT and TMT-A (0.54) and TMT-B (0.54)	[37]
Maerlender <i>et al.</i> (2010)	68 healthy controls	Compared IMPACT to CVLT, BVMT-R, DKEFS, CPT, GPB and PASAT	Pearson <i>r</i> correlation coefficients Canonical correlations	Correlations ranged from -0.39 to 0.59 Canonical correlations: two of the five canonical dimensions were significant (0.80 and 0.73), indicating the batteries measured similar constructs	[43]
Maerlender <i>et al.</i> (2013)	68 healthy controls	Conducted additional analyses of Maerlender <i>et al.</i> (2010) data	Point-biserial correlations	Point-biserial correlation: IMPACT did not significantly discriminate between dissimilar measures	[44]
Allen and Gfeller (2011)	100 healthy controls	Compared IMPACT to the NFL NP battery: HVLTR, BVMT-R, TMT, COWAT and WAIS-III	Pearson <i>r</i> correlation coefficients	Correlations ranged from -0.38 to 0.43. Strongest correlation was between WAIS-III DS and IMPACT Visual Motor Speed	[45]
Solomon and Kuhn (2014)	226 NFL draft picks, with and without mTBI history	Compared IMPACT to Wonderlic	Pearson <i>r</i> correlation coefficients	Correlations ranged from -0.26 to 0.49. Strongest correlation was with Visual Motor Speed	[46]
Van Kampen <i>et al.</i> (2006)	122 athletes with acute mTBI and 70 healthy controls	Administered IMPACT at baseline and then at 2-day postinjury for mTBI group, and postseason for controls	RCI-defined impairment Percent change from baseline	93% of the mTBI group performed lower than at their baseline 30% of control group performed lower than baseline	[47]
Broglio <i>et al.</i> (2007)	75 with mTBI high school students	Compared IMPACT to HVLTR, SDMT, DS, COWAT	Chi-square	IMPACT demonstrated better sensitivity to mTBI (62.5%) than traditional battery (43.5%)	[48]
Covassin <i>et al.</i> (2008)	57 athletes with mTBI (36 with no history of mTBI, 21 with history of 2+ mTBI)	Administered IMPACT at baseline and then at 1- and 5-day postinjury	Chi-square and univariate <i>post hoc</i> tests	mTBI group exhibited significantly lower scores on Verbal Memory ( $p = 0.01$ ) and RT ( $p = 0.023$ ) at 5-day postinjury	[49]

AUC: Area under the curve; ANOVA: Analysis of variance; BVMT-R: Brief Visual Memory Test-Revised; COWAT: Controlled Oral Word Association Test; CPT: Continuous Performance Test; CRT: Choice Reaction time; CVLT: California Verbal Learning Test; DFA: Discriminant Functional Analysis; DKEFS: Delis Kaplan Executive Functioning System; DS: Digit Symbol; EFA: Exploratory factor analysis; GPB: Grooved Pegboard; HVLTR: Hopkins Verbal Learning Test-Revised; IMPACT: Immediate postconcussion assessment and cognitive testing; MANOVA: Multivariate analysis of variance; mTBI: Mild traumatic brain injury; NFL: National Football League; NP: Neuropsychological; PASAT: Paced Auditory Serial Addition Test; RCI: Reliable Change Index; ROC: Receiver operating characteristics; RT: Reaction time; SDMT: Symbol Digit Modalities Test; TMT: Trail Making Test; WAIS-III: Wechsler Adult Intelligence Scale-Third Edition; WAIS-R: Wechsler Adult Intelligence Scale-Revised.

Table 5. Summary of Immediate Postconcussion Assessment and Cognitive Testing validity studies (cont.).

Study (year)	Study sample	Methods: tests	Methods: statistics	Results	Ref.
Gardner et al. (2012)	51 rugby players with acute mTBI and 41 controls	Compared IMPACT to WAIS-III. Also administered CogState	ANOVA Cohen's <i>d</i> effect sizes Logistic regression to distinguish between groups	Significant differences between groups on 1/4 of the IMPACT scores. Cohen's <i>d</i> effect size = -0.57 Logistic Regression: IMPACT improved classification accuracy 3.5% beyond demographics	[41]
Schatz et al. (2006)	72 high school athletes with mTBI and 66 healthy controls with no history of mTBI	Administered IMPACT at 3-day postinjury	MANOVA DFA: classification of group status	MANOVA: significant differences between groups on four out of five index scores (partial $\eta^2$ ranged from 0.19 to 0.31) DFA: sensitivity = 81.9, specificity = 89.4	[50]
Schatz and Maerlender (2013)	Analyzed pre-existing data of 21,537 athletes' baseline assessments and 560 post-mTBI datasets	Performed factor analysis on preexisting IMPACT datasets	Factor analysis	Identified two primary cognitive factors: first, memory (verbal and visual memory) and second, speed (visual motor speed and RT)	[51]
Schatz and Sandel (2013)	Analyzed pre-existing data of athletes with mTBI (81 symptomatic and 37 asymptomatic), and matched controls ( <i>n</i> 's = 81 & 37)	Administered IMPACT at baseline and then at 3-day postinjury for mTBI group, and postseason for controls	DFA	DFA: symptomatic vs controls. Symptomatic group: sensitivity = 91.4%, specificity = 69.1% DFA: asymptomatic vs controls: sensitivity = 94.6%, specificity = 97.3%	[52]
Nelson et al. (2016)	165 athletes with mTBI and 166 healthy controls	Administered IMPACT at baseline and then at 1, 8, 15 and 45 days (postinjury/baseline)	ANOVA Cohen's <i>d</i> effect sizes RCI-defined impairment ROC curve, including AUC	ANOVA: significant for four or five scores at day 1 (Cohen's <i>d</i> ranged from -0.70 to -0.80). Minimal significant differences at days 8, 15, or 45 AUC: significant for four of five scores at day 1 (0.70–0.71). Minimal significant differences at days 8, 15 and 45	[28]

AUC: Area under the curve; ANOVA: Analysis of variance; BVMT-R: Brief Visual Memory Test-Revised; COWAT: Controlled Oral Word Association Test; CRT: Choice Reaction time; CVLT: California Verbal Learning Test; DFA: Discriminant Functional Analysis; DKEFS: Delis Kaplan Executive Functioning System; DS: Digit Symbol; EFA: Exploratory factor analysis; GPB: Grooved Pegboard; HVLIT-R: Hopkins Verbal Learning Test-Revised; IMPACT: Immediate postconcussion assessment and cognitive testing; MANOVA: Multivariate analysis of variance; mTBI: Mild traumatic brain injury; NFL: National Football League; NP: Neuropsychological; PASAT: Paced Auditory Serial Addition Test; RCI: Reliable Change Index; ROC: Receiver operating characteristics; RT: Reaction time; SDMT: Symbol Digit Modalities Test; TMT: Trail Making Test; WAIS-III: Wechsler Adult Intelligence Scale-Third Edition; WAIS-R: Wechsler Adult Intelligence Scale-Revised.

teries [17–18,20–21]. MTH, STN and LGR appear to be the ANAM subtests that best predict performance on traditional tests such as Wechsler Adult Intelligence Scale-Revised (WAIS-R), WJ-III, TMT and RBANS. Principal component analysis (PCA) has also been used to further investigate the relationship between scores on ANAM and traditional NP tests. Generally, data from such studies also provide evidence for construct validity, demonstrating that ANAM is assessing underlying cognitive constructs of efficiency, working memory and resistance to interference similarly to traditional NP tests [17–18,20].

### Group differences

Results from several studies to date suggest that ANAM may have some diagnostic utility in mTBI cases, particularly in the acute phase of injury (see [Table 2](#) for accompanying details on the existing literature to supplement this text). Bleiberg and Warden [22] administered baseline ANAM assessments to US Military Academy cadets and made comparisons in performance at four time points over a 2-week period (during the first 2 weeks after injury for those in the mTBI cohort). Using ANAM's Reliable Change Index (RCI) as a cutoff for impairment [53,54], ANAM scores were generally able to differentiate examinees with mTBI from healthy controls, as the mTBI group had two or more RCI-based performance declines, while controls did not. Additionally, significant practice effects were only demonstrated in the control group (53%). The authors suggested that the absence of a practice effect in the mTBI group might be one of the better ways to identify cognitive impairment following concussion.

An earlier study [23] investigating the diagnostic capabilities of ANAM as compared with traditional NP batteries found differences between the mTBI and demographically matched control group on four of the five ANAM subtests. Participants were tested 30-times over the course of 4 days (i.e., six-times on day 1, eight-times per day for the next 3 days) to attempt to replicate previous findings that examinees with mTBI had larger variability on measures of RT and response speed over multiple assessment sessions [55–57]. Their findings also revealed control participants demonstrated less variability and a practice effect over time, while the mTBI group's performance was more variable and actually declined across repeated test sessions.

In their assessment of a mixed clinical sample, Woodhouse *et al.* [21] used logistic regression to determine the classification accuracy of ANAM to predict RBANS scores that were  $\leq 15$ th percentile (i.e., 'impaired'). All seven subtests generated significant differences among the groups of healthy and impaired individuals. This model indicates that ANAM TP scores can predict

impairment status with high sensitivity and specificity, suggesting that ANAM is capable of classifying impairment similarly to the RBANS. However, the RBANS is typically used in the assessment of age-related cognitive decline, and therefore may not be the most suitable assessment for postconcussion evaluation. Kelly *et al.* [24] found that, in baseline and intergroup comparisons among concussed and healthy soldiers deployed to combat environments, the best area under the curve (AUC), which is an indicator of discriminant ability (i.e., differentiate between those with mTBI and controls), came from Simple Reaction Time (SRT) TP scores. The data suggest that this distinction may be the most accurate within 72 h of injury. Some other score combinations improved ANAM's discriminant ability (e.g., SRT + Procedural Reaction Time [PRO] for normative comparisons, SRT + MTH + Matching to Sample [M2S] for baseline comparisons), though not drastically so. These results possibly provide support for using only RT- and PRO-based tasks in a potentially shorter ANAM battery. Coldren *et al.* [25] also sought to evaluate the diagnostic capability of ANAM in the combat environment and found that, in comparison to controls and baseline scores, the mTBI group demonstrated lower scores with statistically significant differences on five of the six ANAM subtests scores at  $\leq 3$  days postinjury. However, only minimal differences were found at the 5 and 10+ days intervals.

Norris *et al.* [26] found that ANAM assessments at 3- and 5-day postinjury may demonstrate prognostic utility. In their study, those soldiers who performed at or lower than 25% needed 19 days to recover and be cleared to return to duty (RTD), while those who performed in the top 25% were able to RTD in just 7 days after injury, with the largest effect sizes seen for the SRT2 subtest. Results from another study suggest that the use of ANAM as a diagnostic tool for concussion may be limited. In a sample of college football athletes, few examinees with concussion were consistently classified as impaired across ANAM and a traditional NP battery [27]. In this study, ANAM had high specificity but low sensitivity; however, when combined with the Sensory Organization Test and symptom measures, the sensitivity improved, albeit only slightly. These results indicating low sensitivity raise questions about the isolated use of ANAM or any other concussion test (Sensory Organization Test or Graded Symptom Checklist) for clinical decision-making.

Finally and most recently, Nelson *et al.* [28] prospectively compared three NCATs, including ANAM, in groups of concussed and healthy athletes at 1, 8, 15 and 45 days following injury, with similar intervals for matched controls. At 1-day postinjury, AUC was fair and scores from six of the seven subtests as well as

the composite score were significantly different from the control group. At days 8 and 15 postinjury, only one subtest (M2S) showed significant differences, and there were minimal significant differences at 45-day postinjury. The authors concluded that ANAM has limited clinical utility after 8 days following mTBI.

### Summary

When evaluating the existing literature on ANAM according to the Randolph criteria [7], it does not appear that these criteria have been sufficiently addressed. Correlations with traditional NP tests are generally moderate at best, though often weaker. Moreover, the stronger correlations are not consistently between tests that purport to measure the same cognitive construct. The scores that seemed to be the most robust from the ANAM were MTH and those that are primarily RT based, often most strongly correlated with traditional tests of motor and processing speed. Therefore, construct validity as measured by correlation analyses is questionable at best. However, there are indications from regression analyses and PCA that similar cognitive constructs are being measured by ANAM and traditional NP tests, though perhaps in a slightly different manner.

The results from the existing literature also suggest that ANAM has questionable sensitivity to the effects of concussion, especially if testing is completed more than a week from injury. While the mTBI groups often displayed worse performance, more variability in performance or a lack of practice effects as compared with controls, the diagnostic utility of these differences is currently unconvincing. Though specificity was often high and approaching clinically meaningful levels, sensitivity was generally lower than desired. However, there are indications that identifying cognitive impairment rather than mTBI status may be more meaningful and yield better diagnostic accuracy. This approach was recommended by Iverson and Schatz [16] and may be the best approach to addressing Randolph *et al.*'s [7] second criterion evaluating the sensitivity to the effects of concussion.

## CNS-Vital Signs

### Comparisons to traditional NP tests

There is not a large body of published literature regarding the validity of CNS-VS. The correlational studies suggest some degree of relatedness between CNS-VS and traditional NP tests, although no consistently clear patterns have been determined (see Table 3 for details on the methodology and results of these investigations). Gualtieri and Johnson [29] found significant correlations between CNS-VS and a traditional NP test battery in groups of healthy controls and patients with

various neuropsychiatric disorders, including postconcussion syndrome (PCS) and severe brain injury. CNS-VS Symbol Digit Coding and WAIS Digit Symbol Coding subtest scores were identified as the strongest correlated scores, providing some evidence of convergent validity. Another study [30] evaluating scores in a sample of examinees 6–8 weeks removed from concussion generated significant, though modest, correlations between CNS-VS and the traditional NP tests, with the strongest correlation between CNS-VS Psychomotor Speed Standard Score and the Neuropsychological Assessment Battery Memory Index Standard Score. Gualtieri and Hervey [31] found that overall, correlations among the CNS-VS and traditional battery were weak to moderate (-0.33 to 0.59) in a sample of psychiatric patients. WAIS-III and -IV Full Scale Intelligence Quotient (FSIQ) and CNS-VS Shifting Attention Test were the most strongly correlated. The authors also conducted multiple regression analyses to further explore the relationship between CNS-VS and traditional NP tests, demonstrating that only two of the CNS-VS scores (Shifting Attention Test and Visual Memory Test) were significant predictors of FSIQ.

### Group differences

To date, there are three published studies looking at the diagnostic utility of CNS-VS with mTBI (see Table 3 for study summaries to supplement this text). Lanting *et al.* [32] administered CNS-VS to patients 6–8 weeks after sustaining either an mTBI or orthopedic injury. Though the mTBI group did have a higher proportion of scores at least one standard deviation below the mean, effect sizes were small and multivariate analysis of variance demonstrated no statistically significant differences between the two groups. Gualtieri and Johnson [33] compared healthy controls to a TBI cohort divided into four subgroups: those with PCS, those recovered from mTBI, those recovered from severe TBI and those who had not recovered from a severe TBI. Multivariate analysis of variance demonstrated statistically significant differences between the five groups in 18 of the 28 scores investigated. *Post hoc t*-tests clarified significant differences from healthy controls existed on all scores in both severe TBI groups and on five of six CNS-VS scores in the PCS group. There were no differences between the controls and those in the mTBI-recovered group. Receiver operating characteristic (ROC) curve analyses revealed which index scores better identified differences between the groups. A greater AUC identified those tests that could best distinguish between groups as the CNS-VS Psychomotor Speed index score, which had the greatest AUC (0.752), followed by the Neurocognitive Index (NCI; 0.747) and Cognitive Flexibility Score (0.708). Although these

results indicate that the NCI may have some diagnostic capabilities, the authors question the clinical use of the NCI score, as it is currently only utilized in research settings and is not common in traditional NP assessment. Lastly, another study compared CNS-VS scores before and after deployment in active-duty service members. Though there were significant differences between examinees on pre- and postdeployment measures, there were no significant differences on CNS-VS performance [34].

### Summary

Unfortunately, the existing research suggests somewhat mixed, though largely unfavorable, results for the validity of the CNS-VS battery. Specifically, correlation analyses show no clear pattern of convergent or discriminant validity, and generally CNS-VS is at best moderately correlated with traditional NP tests. Group comparisons suggest no clear or clinically meaningful differences between groups with mTBI and control groups. However, there is a paucity of research for CNS-VS, and further investigation is required to address the Randolph criteria. Additional studies taking different approaches may yield different and more promising results. For example, more clinically meaningful differences may be evident when comparing those with acute mTBI to control groups, as the existing literature was based solely on assessments administered long after the acute postinjury timeframe. Additionally, Gualtieri and Johnson [33] found significant differences between still symptomatic groups and controls, suggesting identification of cognitive impairment or still symptomatic individuals may be more clinically meaningful in identifying someone as recently concussed or not.

### CogState

#### Comparisons to traditional NP tests

Studies comparing CogState to traditional NP tests have typically focused on traditional tests of processing speed and executive functioning, generally finding some evidence for construct validity (see [Table 4](#) for more detail associated with the findings). Makdissi *et al.* [35] reported statistically significant, though moderate at best, correlations between the CogState SRT subtest and the traditional NP tests (i.e., Digit Symbol Substitution Test and TMT-B) in samples of healthy controls and patients with acute mTBI. They found increases in variability and latency of responses in the dataset from these injured players. In a similar study [36], correlations were weak between the CogState battery's accuracy scores and the DSST and TMT scores; however, when CogState speed scores were used, there were several strong correlations, most

notably between the DSST and the decision-making and working memory speed scores (-0.86 and -0.72, respectively). Schatz and Putz [37] reported moderate correlations between CogSport and a traditional NP battery, with SRT being the strongest correlated score with WAIS-R Digit Symbol Coding. In a study where healthy controls' performance on CogState was compared with a larger battery of traditional NP tests, Maruff *et al.* [38] reported moderate-to-strong correlations between the various CogState domains (processing speed, attention, working memory and learning) and the traditional NP tests measuring similar constructs, suggesting support for the construct validity of CogState.

#### Group differences

There have been four studies published to date looking at the difference in performance on CogState between healthy controls and mTBI patients, with mixed results (see [Table 4](#) for accompanying details on the methodology and results). In one of the earlier studies [35], traditional NP scores did not significantly change from baseline in either the concussed or nonconcussed samples of football players, though the concussed group did demonstrate a significant (36%) decline in performance on the CogState SRT task. *Post hoc t*-tests demonstrated that the control and concussed groups' SRT variability were not statistically different at baseline, but the concussed groups had significantly more RT variability at follow-up. Similarly, a prospective study of cognitive functioning following concussion in football players found that the symptomatic group of patients who sustained a concussion demonstrated a significant decline in CogState performance and no change in traditional NP test scores, while the controls and asymptomatic concussion groups mostly improved in their performance in both CogState and the traditional NP battery [39]. Maruff *et al.* [38] found evidence of criterion validity by administering CogState to three groups of examinees with cognitive impairment (mTBI, schizophrenia and AIDS Dementia Complex), and three groups of demographically matched controls. Of interest to this review, the mTBI group was significantly different from the control group, with large effect sizes observed on the OCL/learning task. In addition, the authors used a measurement called the nonoverlap statistic (non-OL%) to identify the percentage of each group's data distributions that do not overlap. Using this metric, they found that each of the CogState subtests was able to identify between 53 and 78% of the impairment unique to the mTBI group ( $p < 0.0001$ ). Utilizing both baseline and normative reference groups, Louey *et al.* [40] have also provided evidence that CogState can be used to detect

concussion-related cognitive impairment. The authors found that the baseline method demonstrated a higher sensitivity and comparable specificity to the normative method, and even after taking into account baseline performances, the concussed group showed performance declines. The baseline and normative methods could be used to correctly classify individuals as cognitively impaired up to an accuracy of 87.9 and 89.3%, respectively.

However, in two of the studies to date, comparisons among healthy and impaired individuals provided less convincing support of CogState's criterion validity, as CogState either only moderately improved classification between groups or, similar to ANAM, could do so only at earlier postinjury time points. Gardner *et al.* [41] administered CogState alongside ImpACT and WAIS-III to rugby players with or without acute mTBI. They observed statistically significant differences between the groups on four of the five CogState subtest scores. However, logistic regression demonstrated that CogState scores only minimally improved classification accuracy above what demographics predicted when added to the regression model. In their prospective study (previously mentioned in the ANAM review), Nelson *et al.* [28] found that at 1-day postinjury, all of CogState's subtests were significantly different from the control group. There were two subtests (Attention Speed and Learning Speed) that also demonstrated statistically significant differences at 8-day postinjury, though with small effect sizes. The ROC analyses revealed that only the CogState subtests administered at 1-day postinjury demonstrated significant AUC, suggesting that at the later time points, CogState subtests likely do not have diagnostic utility.

### Summary

Though correlations between CogState and traditional NP tests have been wide ranging, there is some evidence for convergent validity, with a general pattern of tests supposedly measuring similar cognitive constructs being more strongly correlated than dissimilar measures. Investigations of the clinical utility of CogState with concussion have had mixed results. Some tests have demonstrated the ability for CogState to distinguish between those with concussion and controls, even outside of the 7-day window (e.g., 1 month). In fact, one study suggested CogState may have had more clinical utility in postconcussion assessments than traditional NP tests [39]. Also, CogState was able to correctly classify over 88% of individuals as concussed or not by comparing scores to both normative databases and baseline performance [40].

Similar to research with other NCATs, the clinical utility of CogState may be increased by identifying

individuals who are symptomatic after injury, rather than just comparing those with concussion (and possibly asymptomatic) to healthy controls. However, other studies have found that CogState's ability to detect postconcussive symptoms may be limited outside of the acute stage of injury (e.g., beyond 7 days), and even in the acute stage it may not provide much information beyond demographics. Overall, though the Randolph criteria have been largely addressed by the existing research, there is inconclusive support for meeting those criteria and additional research with CogState is warranted. The study samples and traditional NP test batteries used in the existing studies have been fairly narrow. Additionally, the wide range of correlations between CogState and traditional NP tests warrants regression and PCA to determine the extent to which CogState is measuring similar cognitive constructs to traditional NP tests. Also, additional studies investigating the clinical utility of CogState to detect cognitive impairment, both in and beyond the acute injury phase, are necessary.

### ImpACT

#### Comparisons to traditional NP tests

Alsalaheen *et al.* [15] conducted a comprehensive and systematic review of the validity of ImpACT, and the intention is not to repeat their work. The reader is encouraged to consult their review for a comprehensive summary of ImpACT literature to date. Alsalaheen *et al.* [15] concluded that there is strong evidence for convergent validity of ImpACT though weak or inconclusive evidence for discriminant validity, criterion validity or diagnostic accuracy and utility. This would suggest inconclusive support for meeting the Randolph criteria [7]. Below, we highlight several studies investigating convergent and criterion validity, as well as diagnostic utility in mTBI cases (see Table 5 for study summaries to accompany the findings described below).

Iverson *et al.* [42] compared ImpACT results with those of a paper and pencil test commonly used as a measure of attention and processing speed (Symbol Digit Modalities Test [SDMT]) in a cohort of young athletes. The strongest correlations with SDMT were ImpACT's Processing Speed and RT composite scores. Exploratory factor analysis (EFA) uncovered a two-factor solution of speed/RT and memory, suggesting ImpACT is measuring similar cognitive constructs as SDMT. Schatz and Putz [37] found moderate correlations among ImpACT and a traditional NP battery in a group of healthy controls, with the strongest correlation being ImpACT Choice RT score and Trails A. Similarly, Maerlender *et al.* [43] found that ImpACT was moderately correlated with tests of similar cog-



nitive domains. Canonical correlation analyses indicated that two of the five canonical dimensions were statistically significant, with coefficients of 0.801 and 0.729, confirming that the two batteries generally measure similar cognitive constructs. However, a follow-up study by the same authors [44] re-analyzed the 2010 dataset to specifically evaluate the discriminant validity of ImpACT as compared with traditional NP tests. The results indicated that while the traditional battery demonstrated evidence of discriminant validity (i.e., all domains' p-values > 0.05 except RT), ImpACT did not discriminate between measures of different cognitive skills. Specifically, three of the four domain scores were strongly correlated with expectedly different traditional NP measures.

Allen and Gfeller [45] compared performance measures of ImpACT to those of the NFL NP battery, which consists of the Hopkins Verbal Learning Test-Revised, Brief Visual Memory Test-Revised, TMT, Controlled Oral Word Association Test and three subtests from the WAIS-III, in a sample of healthy controls. Correlations were moderate at best, with the strongest correlation between WAIS-III Coding and ImpACT's Visual Motor Speed Composite. Solomon and Kuhn [46] examined the relationship between performance on the Wonderlic and ImpACT in 226 NFL draft picks with and without a history of concussion. Concussion history did not have a significant effect on performance on either of the tests. Correlations between the batteries were weak to moderate, with Visual Motor Speed being the most strongly correlated with Wonderlic performance.

### Group differences

Studies to date comparing ImpACT to a variety of traditional NP tests and among many different patient populations have corroborated that ImpACT may be useful as a diagnostic tool postconcussion, and perhaps even the most sensitive of the four NCATs described in this review (see [Table 5](#) for supporting summaries of each study). Van Kampen *et al.* [47] compared the ImpACT performance of college athletes with acute concussion to matched controls, also utilizing preseason baseline assessment scores. RCI scores defining abnormal performance indicated that 83% of the participants in the mTBI group performed abnormally lower than their baseline. When cognitive data were combined with symptom questionnaires, 93% were categorized as abnormal. However, 30% of the control group also generated abnormal ImpACT test data or self-reported symptoms. Broglio *et al.* [48] reported that groups of students with and without acute mTBI differed on all indices except Impulse Control. Furthermore, the ImpACT battery demonstrated better sensitivity to mTBI (79.2%) than a traditional NP battery (43.5%).

Similarly, in a study of recently concussed college athletes, Covassin *et al.* [49] found that there were significant differences in Verbal Memory and RT based on whether participants had a history of prior concussion (i.e., those with a prior concussion performed worse) at both 1- and 5-day postinjury. Schatz *et al.* [50] also observed that ImpACT classified a group of recently concussed high school athletes with a sensitivity/specificity of 81.9/89.4. Schatz and Maerlender [51] performed factor analyses using existing ImpACT datasets, which included 21,537 baselines and 560 postinjury assessments. They identified two primary cognitive factors, memory (comprised of Verbal and Visual Memory domains) and speed (comprised of Visual Motor Speed and RT domains), that accurately classified individuals as concussed or not concussed, with a sensitivity/specificity of 89/70.

However, there has not been universal evidence that ImpACT adequately differentiates between healthy controls and recently concussed individuals. As previously mentioned, Gardner *et al.* [41] administered ImpACT, CogSport and WAIS-III to professional rugby players with acute mTBI and to matched controls. They found statistically significant differences between the groups on only one of the four ImpACT composite scores (Visual Motor Speed). Logistic regression demonstrated that ImpACT scores were unable to distinguish between the injured and control groups beyond demographic variables, as ImpACT scores only added 3.5% improvement in accuracy to the overall classification model. ROC curve analyses demonstrated modest sensitivity and specificity for the ImpACT composite score.

There is additional support for the clinical utility of ImpACT from studies investigating the test's ability to distinguish between symptomatic and asymptomatic mTBI patients. Schatz and Sandel [52] administered ImpACT to groups of high school and college athletes with acute mTBI (symptomatic and asymptomatic) within 72 h of injury. The data were compared with demographically matched controls with pre- and post-season assessments. ImpACT data demonstrated the ability to detect differences between the groups (sensitivity/specificity of 91.4/69.1 and 94.6/97.3 for the symptomatic and asymptomatic groups, respectively). In the prospective NCAT comparison by Nelson *et al.* previously described [28], all ImpACT composite scores were significantly different at 1 day following injury. However, there was only one score that was significantly different, and with a small effect size, after this timeframe (day 8, Verbal Memory,  $d = -0.40$ ). Likewise, AUC for the composite scores were fair at the 1-day postinjury assessment [0.70,0.71] though poor (<0.69) for the other timeframes. However, of the three

NCATs evaluated, ImpACT demonstrated the highest percentage of test scores that significantly declined from baseline to 1-day postinjury according to the RCI criteria (67.8% for both symptomatic and asymptomatic concussed populations), although the test also had a slightly higher false-positive rate than ANAM and CogState in the same 24-h period (29.6% compared with 25.0 and 22.0%, respectively). When examinees were dichotomized as symptomatic or not, ImpACT also demonstrated the largest percentage of patients with a significant decline from baseline performance (53.8% at 1-day postinjury).

### Summary

ImpACT is the most widely studied of the NCATs, and as such the Randolph criteria [7] have been thoroughly addressed through the existing body of research. Though the Randolph criteria have been satisfied to a degree, as Alsalaheen *et al.* [15] concluded that there are mixed results regarding the overall validity of ImpACT. Specifically, there appears to be solid evidence that ImpACT has adequate relatedness with traditional NP tests, especially those purported to measure similar cognitive constructs. More advanced statistical approaches suggest there is also evidence that ImpACT is measuring similar cognitive constructs to traditional NP testing. However, there is not a clear pattern of weaker relationships between tests of dissimilar cognitive constructs, calling into question the discriminant validity of ImpACT. ImpACT's tests of RT and processing speed, especially Visual Motor Speed, seem to have the most robust relationships with traditional NP tests. And with regard to identifying postconcussion issues, ImpACT does show the ability to distinguish between concussed and noninjured individuals during the early stages postinjury. And though sensitivity is generally better than specificity, there were some studies that found comparable sensitivity and specificity, both of which approached desired levels for clinical decision-making. However, after the early postinjury stages, and certainly outside of 7 days, the clinical utility of ImpACT for postconcussion assessments appears limited. Improved clinical utility may be demonstrated if identification of symptomatic individuals postinjury is the focus, rather than identifying individuals as concussed or not.

### Discussion

The goal of this review was to provide a summary of literature regarding the validity on four commonly used and studied NCATs: ANAM, CNS-VS, CogState and ImpACT. The literature was viewed through the lens of Randolph *et al.*'s criteria presented in their 2005 [7] literature review of NP testing after SRC (Box 1). NCATs are becoming the standard of care for

mTBI screening in athletic and military deployment settings given the improvement in efficiency and feasibility of test administration over their traditional NP counterparts. However, it is clear from the above summary of the literature to date that there has yet to be definitive evidence in support of the validity of any of the four NCATs, per Randolph's validity-related criteria (i.e., criteria two through five).

Currently, the body of literature suggests mixed results regarding NCATs' validity. Specifically, there is evidence that NCATs measure similar cognitive constructs as traditional NP tests (i.e., Randolph's 3rd criterion). And there is some support that NCATs, or at least components of each NCAT, can distinguish between individuals with acute concussion and healthy controls, or between still symptomatic individuals and individuals who are symptom free (i.e., Randolph's 2nd criterion and 5th criterion). However, there is little to no evidence for discriminant validity as compared with traditional NP tests, and inconsistent evidence for the clinical utility of NCATs for identifying concussion-related problems, especially beyond the first 7-day postinjury and when the tests are used in isolation. We did not review the literature regarding Randolph's 1st criterion, related to test–retest reliability, as this was beyond the scope of the paper. With regard to Randolph's 4th criterion, establishing RCIs and probability-based algorithms for clinical use is dependent on well-established test–retest reliability and well-defined constructs of the tests. That is, we need to know what the test is measuring, how it is measuring it and how consistently it does so before we can calculate them. As such, additional research will be needed before any of the NCATs fully satisfy the criteria for validity and ultimately for clinical utility.

Although there is not consistent evidence regarding the validity and clinical utility of NCATs, and the criteria presented by Randolph *et al.* [7] have not been sufficiently addressed, there is evidence suggesting that NCATs are of potential benefit in postconcussion assessments. It may be that the tests are fundamentally different than traditional NP tests, and therefore using traditional NP tests as a point of comparison, or using traditional psychometric approaches to defining validity creates a logical fallacy of false analogy or an 'apples to oranges' comparison. That is, perhaps NCATs should not be faulted for not being a good proxy for traditional NP tests, but rather should be investigated as an altogether different assessment tool. Therefore, we explore future directions for this field of research through the lens of the Randolph criteria.

Studies should seek to address Randolph's 2nd and 5th criteria by designing studies that "establish the ability to identify cognitive impairment after concussion and distinguish between individuals who are symp-

omatic and those who are asymptomatic post-injury.” This approach shifts away from a group-based approach (e.g., mTBI vs controls) that has dominated the literature to date, focusing on cognitive impairment and symptom-driven approaches, while allowing for a wider range of methodology in future studies. There were several studies identified in this review that demonstrated NCATs consistently found more clinically meaningful differences between symptomatic versus asymptomatic groups, and that asymptomatic individuals often performed like healthy controls [28,39,52]. Similar future research may prove more valuable in elucidating the clinical utility of these tests.

This impairment and symptom-based approach is also consistent with the recommendation of Iverson and Schatz [16] to specifically investigate cognitive impairment rather than mTBI status. They go further and describe new approaches to identifying cognitive impairment, such as taking a base rate approach and categorizing performance based on the total number of low scores across a battery. Determining clinically meaningful definitions of cognitive impairment, and then establishing the NCATs’ sensitivity and specificity in classifying individuals with concussion as cognitively impaired, will be key to further establishing the validity, and ultimately, the clinical utility of NCATs, especially with regard to informing return to play and RTD decisions.

Randolph’s 3rd criterion may be addressed by studies seeking to “determine what cognitive constructs NCATs are measuring, and if those constructs, and the manner in which they are measured, are clinically meaningful.” This direction is suggested in light of the evidence that the standard statistical approaches of assessing validity have yielded at best moderate convergent validity, poor discriminant validity and inconsistent evidence that NCAT scores predict traditional NP scores. However, alternative statistical approaches, such as PCA and EFA, have suggested NCATs are measuring similar cognitive constructs, though perhaps in different ways. Specifically, it may be that the names given to NCAT subtests and index scores may not accurately reflect the actual cognitive construct being measured. Therefore, statistically guided comparisons, rather than those guided by nomenclature, could yield better evidence for convergent and discriminant validity.

We also recommend a shift away from ‘standard psychometric procedures’ since this often relies on comparisons to a gold standard, such as traditional NP tests. However, there is mixed evidence for the utility of traditional NP tests for use in postconcussion assessments, especially outside of the acute injury phase [24]. NCATs are typically presented as potential proxies for traditional NP tests, and as such, validity is often evaluated by direct comparisons between supposedly compa-

table tests. However, adapting pencil and paper tests to a technological interface can fundamentally change the test. Some have suggested that an NCAT’s ability to precisely measure RT may be an advantage over traditional NP tests in detecting subtle cognitive declines after concussion [5]. In fact, RT and processing speed scores are often the most robust in studies predicting concussion status or cognitive impairment. Also, several studies have identified alternative scores or interpretative methods that may provide more clinical utility than the standard scores currently provided. For example, RT variability and lack of practice effects may be more sensitive to concussion-related effects [22–23,35,58–60]. Thus, the potential technological advantages provided by NCATs warrant closer investigation. A caveat, however, as others have identified sources of error that are introduced into test scores due to the use of technology. This includes a participant’s familiarity with using a computer [61] to hardware and software configurations [4,62–63]. The literature is limited in identifying how technology can affect the measurement of performance, and this will be important to clarify in future studies.

There are several other considerations with regard to the manner in which NCATs assess cognitive functioning, and the subsequent impact on clinical utility. First, though comparisons to baseline assessments have routinely been used with NCATs, and can be helpful in the context of cognitive changes in examinees with pre-existing unique cognitive abilities (i.e., upper or lower 20th percentile, ADHD or LD), the use of baseline testing does not appear to be necessary for determining cognitive deficits following concussion [64,65]. Research should focus on the ability for baseline assessments, normative comparisons or some combination of the two to accurately and adequately identify symptomatic individuals. Also, the use of group versus individual settings during test administration should be considered, as there is mixed evidence regarding the potential impact a group versus individual test setting has on test scores [66,67]. The different administration settings could potentially impact the findings as NCATs are often administered in group settings either preseason in athletics or predeployment in the military, and then individually postinjury. Additionally, the environment in which NCATs are often desired to be administered, such as athletic sidelines or combat zones, is an important consideration as much of the research takes place in highly controlled settings. The clinical utility of NCATs in such austere environments warrants further investigation [68,69].

## Conclusion & future perspective

Though the body of literature regarding the validity of the four NCATs discussed in this review has been

steadily growing, there appears to be insufficient evidence suggesting that these tools are adequate proxies for traditional NP tests and have limited clinical utility in postconcussion assessments. However, by investigating NCATs with the same methodology used to investigate traditional NP tests, these tests may have been set up for failure. Using the 2005 Randolph criteria, we have provided additional and alternative ways forward for investigating the validity and utility of NCATs that are better suited for the intended clinical use and design of these tests. Future efforts are encouraged to focus on cognitive impairment (e.g., symptomatic vs asymptomatic) rather than group status (e.g., concussed vs controls), the ability to inform return to play and RTD decisions, and utilization of alternative and novel statistical approaches (e.g., RT variability, base rate analyses to identify impairment, etc.). Additional prospective comparisons of multiple NCATs in differing study samples, similar to the one conducted by

Nelson *et al.* [28] are also warranted. NCATs have the potential to fundamentally change the nature of care following mTBI. However, until their clinical utility can be further established and clarified, they should be used with caution and at most as screening tools in combination with multifaceted assessments.

#### Disclaimer

The views expressed herein are those of the author(s) and do not reflect the official policy of the Department of the Army, Department of Defense or the US Government.

#### Financial & competing interests disclosure

This material is published by permission of the Defense and Veterans Brain Injury Center, operated by General Dynamics Information Technology for the US Defense Health Agency under Contract No. W91Y7Z-13-C-0015. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial

### Executive summary

#### Background

- The assessment of cognitive functioning has been identified as an important aspect in the management of concussion.
- Due to logistical advantages over traditional neuropsychological (NP) tests, computerized Neurocognitive Assessment Tools (NCATs) have gained popularity in athletic and military settings.
- However, the psychometric properties, especially validity, and clinical utility of NCATs have yet to be consistently established.

#### Neurocognitive assessment tools

- Automated Neuropsychological Assessment Metric is commonly used to assess cognitive functioning in US Military Service Members.
- CNS-Vital Signs is commonly used in psychiatric and neurological clinical trials.
- CogState/Axon/CogSport is commonly used in Australian athletics.
- Immediate Post-Acute Concussion Test (ImPACT) is the most widely used NCAT in US athletics. It has the US FDA approval for postconcussion assessments.

#### Existing evidence for validity & clinical utility

- Automated Neuropsychological Assessment Metric related best with traditional NP tests of processing speed, with evidence of moderate sensitivity/specificity for concussion or postconcussive symptoms during the acute injury period.
- CNS-Vital Signs had the least amount of validity-related research, with findings revealing at best moderate correlations with traditional NP tests and no clear evidence for clinically meaningful differences between concussed and controls, though data were from the postacute injury timeframe.
- CogState demonstrated some evidence of validity with several moderate to strong correlations to traditional NP measures and the ability to detect concussion-related cognitive decline during the acute injury period. However, research has had a narrow focus on primarily reaction-based scores and with Australian athletes.
- ImPACT is the best studied of the NCATs, with research indicating mixed results regarding validity. It does appear ImPACT is measuring similar cognitive constructs as traditional NP tests, with some evidence for detecting concussion-related cognitive decline during the acute injury period at levels approaching those desired for clinical decision-making.

#### Future perspective

- Additional investigation of the validity and clinical utility of NCATs is warranted, with future efforts encouraged to focus on cognitive impairment (e.g., symptomatic vs asymptomatic) rather than group status (e.g., concussed vs controls), the ability to inform return to play and return to duty decisions and novel statistical approaches (e.g., reaction time variability and base rate analyses to identify impairment).
- Additional prospective comparisons of multiple NCATs in differing study samples, similar to the one conducted by Nelson *et al.* (2016) are also warranted.

conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

**Open access**

This work is licensed under the Creative Commons Attribution 4.0 License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

**References**

Papers of special note have been highlighted as:

• of interest; •• of considerable interest

- 1 McCrory P, Meeuwisse WH, Aubry M *et al.* Consensus statement on concussion in sport: the 4th International Conference on Concussion in Sport held in Zurich, November 2012. *Br. J. Sports Med.* 47(5), 250–258 (2013).
- 2 US House of Representatives. National Defense Authorization Act for Fiscal Year 2008. Section 1618. Comprehensive plan on prevention, diagnosis, mitigation, treatment, and rehabilitation of, and research on, traumatic brain injury, post-traumatic stress disorder, and other mental health conditions in members of the armed forces. (2008). [www.gpo.gov](http://www.gpo.gov)
- 3 Bauer RM, Iverson GL, Cernich AN, Binder LM, Ruff RM, Naugle RI. Computerized neuropsychological assessment devices: joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *Arch. Clin. Neuropsychol.* 27(3), 362–373 (2012).
- 4 Echemendia RJ, Iverson GL, McCrea M *et al.* Advances in neuropsychological assessment of sport-related concussion. *Br. J. Sports Med.* 47(5), 294–298 (2013).
- 5 Roebuck-Spencer T, Sun W, Cernich AN, Farmer K, Bleiberg J. Assessing change with the Automated Neuropsychological Assessment Metrics (ANAM): issues and challenges. *Arch. Clin. Neuropsychol.* 22(Suppl. 1), S79–S87 (2007).
- 6 Nunnally JC, Bernstein IH. Validity. In: *Psychometric Theory (3rd Edition)*. McGraw-Hill, NY, USA, 83–113 (1994).
- 7 Randolph C, McCrea M, Barr WB. Is neuropsychological testing useful in the management of sport-related concussion? *J. Athl. Train.* 40(3), 139–152 (2005).
- **Reviewed the literature on neuropsychological testing the management of concussion. Presented criteria for evaluating clinical utility.**
- 8 Resch JE, McCrea MA, Cullum CM. Computerized neurocognitive testing in the management of sport-related concussion: an update. *Neuropsychol. Rev.* 23(4), 335–349 (2013).
- **Literature review of Neurocognitive Assessment Tools used in sport-related concussion.**
- 9 Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56(2), 81–105 (1959).
- 10 Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol. Bull.* 52(4), 281–302 (1955).
- 11 Mitrushina M, Boone KB, Razani J, D’Elia LF. Statistical and psychometric issues. In: *Handbook of Normative Data for Neuropsychological Assessment (2nd Edition)*. Oxford University Press, NY, USA, 41–56 (2005).
- 12 Strauss E, Sherman EMS, Spreen O. Psychometrics in neuropsychological assessment. In: *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary (3rd Edition)*. Oxford University Press, NY, USA, 18–31 (2006).
- 13 Groth-Marnat G. Validity in clinical practice. Introduction. In: *The Handbook of Psychological Assessment (4th Edition)*. John Wiley & Sons, NJ, USA, 17–30 (2006).
- 14 Cernich A, Reeves D, Sun W, Bleiberg J. Automated neuropsychological assessment metrics sports medicine battery. *Arch. Clin. Neuropsychol.* 22(Suppl. 1), S101–S114 (2007).
- **Reviewed the validity of Automated Neuropsychological Assessment Metric in military settings.**
- 15 Alsalaheen B, Stockdale K, Pechumer D, Broglio SP. Validity of the Immediate Post-Concussion Assessment and Cognitive Testing. *Sports Med.* 46(10), 1487–1501 (2016).
- **Systematically reviewed of the validity Immediate Post-Concussion Assessment and Cognitive Testing.**
- 16 Iverson GL, Schatz P. Advanced topics in neuropsychological assessment following sport-related concussion. *Brain Inj.* 29(2), 263–275 (2015).
- 17 Bleiberg J, Kane RL, Reeves DL, Garmoe WS, Halpern E. Factor analysis of computerized and traditional tests used in mild brain injury research. *Clin. Neuropsychol.* 14(3), 287–294 (2000).
- 18 Kabat MH, Kane RL, Jefferson AL, DiPino RK. Construct validity of selected Automated Neuropsychological Assessment Metrics (ANAM) battery measures. *Clin. Neuropsychol.* 15(4), 498–507 (2001).
- 19 Woodard J, Marker C, Tabanico F *et al.* A validation study of the automated neuropsychological assessment metrics (ANAM) in non-concussed high school players. *J. Int. Neuropsychol. Soc.* 8(2), 175 (2002).
- 20 Jones WP, Loe SA, Krach SK, Rager RY, Jones HM. Automated Neuropsychological Assessment Metrics (ANAM) and Woodcock-Johnson III Tests of cognitive ability: a concurrent validity study. *Clin. Neuropsychol.* 22(2), 305–320 (2008).
- 21 Woodhouse J, Heyanka DJ, Scott J *et al.* Efficacy of the ANAM General Neuropsychological Screening Battery (ANAM GNS) for detecting neurocognitive impairment in a mixed clinical sample. *Clin. Neuropsychol.* 27(3), 376–385 (2013).
- 22 Bleiberg J, Warden DL. Computerized neuropsychological concussion surveillance instruments: using the Reliable Change Index (RCI) as a basis for clinical decision-making. Presented at: *American Academy of Neurology 54th Annual Meeting*. Denver, CO, USA, 13–20 April 2002.
- 23 Bleiberg J, Garmoe WS, Halpern EL, Reeves DL, Nadler JD. Consistency of within-day and across-day performance

- after mild brain injury. *Neuropsychiatry Neuropsychol. Behav. Neurol.* 10(4), 247–253 (1997).
- 24 Kelly MP, Coldren RL, Parish RV, Dretsch MN, Russell ML. Assessment of acute concussion in the combat environment. *Arch. Clin. Neuropsychol.* 27(4), 375–388 (2012).
- 25 Coldren RL, Russell ML, Parish RV, Dretsch M, Kelly MP. The ANAM lacks utility as a diagnostic or screening tool for concussion more than 10 days following injury. *Mil. Med.* 177(2), 179–183 (2012).
- 26 Norris JN, Carr W, Herzig T, Labrie DW, Sams R. ANAM4 TBI reaction time-based tests have prognostic utility for acute concussion. *Mil. Med.* 178(7), 767–774 (2013).
- 27 Register-Mihalik JK, Guskiewicz KM, Mihalik JP, Schmidt JD, Kerr ZY, McCrema MA. Reliable change, sensitivity, and specificity of a multidimensional concussion assessment battery: implications for caution in clinical practice. *J. Head Trauma Rehabil.* 28(4), 274–283 (2013).
- 28 Nelson LD, LaRoche AA, Pfaller AY *et al.* Prospective, head-to-head study of three computerized Neurocognitive Assessment Tools (CNTs): reliability and validity for the assessment of sport-related concussion. *J. Int. Neuropsychol. Soc.* 22(1), 24–37 (2016).
- **Prospective comparison of three Neurocognitive Assessment Tools in mild traumatic brain injury in sport.**
- 29 Gualtieri CT, Johnson LG. Reliability and validity of a computerized neurocognitive test battery, CNS-Vital Signs. *Arch. Clin. Neuropsychol.* 21(7), 623–643 (2006).
- **Analyzed the psychometric properties of CNS-Vital Signs.**
- 30 Lanting SC, Iverson GL, Lange RT. Concurrent validity of CNS vital signs in patients with mild traumatic brain injury. Presented at: *The American Congress of Rehabilitation Medicine Conference*. Vancouver, BC, Canada, 11 October 2012.
- 31 Gualtieri CT, Hervey AS. The structure and meaning of a computerized neurocognitive test battery. *Front. Psychol. Behav. Sci.* 4(2), 11–21 (2015).
- 32 Lanting SC, Iverson GL, Lange RT. Comparing patients with mild traumatic brain injury to trauma controls on CNS Vital Signs. Presented at: *The American Congress of Rehabilitation Medicine Conference*. Vancouver, BC, Canada, 11 October 2012.
- 33 Gualtieri CT, Johnson LG. A computerized test battery sensitive to mild and severe brain injury. *Medscape J. Med.* 10(4), 90 (2008).
- 34 Dretsch MN, Silverberg ND, Iverson GL. Multiple past concussions are associated with ongoing post-concussive symptoms but not cognitive impairment in active-duty army soldiers. *J. Neurotrauma* 32(17), 1301–1306 (2015).
- 35 Makdissi M, Collie A, Maruff P *et al.* Computerised cognitive assessment of concussed Australian Rules footballers. *Br. J. Sports Med.* 35(5), 354–60 (2001).
- 36 Collie A, Maruff P, Makdissi M, McCrory P, McStephen M, Darby D. CogSport: reliability and correlation with conventional cognitive tests used in postconcussion medical evaluations. *Clin. J. Sport Med.* 13(1), 28–32 (2003).
- 37 Schatz P, Putz BO. Cross-validation of measures used for computer-based assessment of concussion. *Appl. Neuropsychol.* 13(3), 151–159 (2006).
- 38 Maruff P, Thomas E, Cysique L *et al.* Validity of the CogState brief battery: relationship to standardized tests and sensitivity to cognitive impairment in mild traumatic brain injury, schizophrenia, and AIDS dementia complex. *Arch. Clin. Neuropsychol.* 24(2), 165–178 (2009).
- 39 Collie A, Makdissi M, Maruff P, Bennell K, McCrory P. Cognition in the days following concussion: comparison of symptomatic versus asymptomatic athletes. *J. Neurol. Neurosurg. Psychiatry* 77(2), 241–245 (2006).
- **Compared Immediate Post-Concussion Assessment and Cognitive Testing and CogState on ability to predict group membership.**
- 40 Louey AG, Cromer JA, Schembri AJ *et al.* Detecting cognitive impairment after concussion: sensitivity of change from baseline and normative data methods using the CogSport/Axon cognitive test battery. *Arch. Clin. Neuropsychol.* 29(5), 432–41 (2014).
- 41 Gardner A, Shores EA, Batchelor J, Honan CA. Diagnostic efficiency of ImPACT and CogSport in concussed rugby union players who have not undergone baseline neurocognitive testing. *Appl. Neuropsychol. Adult* 19(2), 90–97 (2012).
- 42 Iverson GL, Lovell MR, Collins MW. Validity of ImPACT for measuring processing speed following sports-related concussion. *J. Clin. Exp. Neuropsychol.* 27(6), 683–689 (2005).
- 43 Maerlender A, Flashman L, Kessler A *et al.* Examination of the construct validity of ImPACT computerized test, traditional, and experimental neuropsychological measures. *Clin. Neuropsychol.* 24(8), 1309–1325 (2010).
- **Combined traditional neuropsychological battery scores into composite scores. Compared battery composite scores with Canonical correlations.**
- 44 Maerlender A, Flashman L, Kessler A *et al.* Discriminant construct validity of ImPACT: a companion study. *Clin. Neuropsychol.* 27(2), 290–299 (2013).
- 45 Allen BJ, Gfeller JD. The Immediate Post-Concussion Assessment and Cognitive Testing battery and traditional neuropsychological measures: a construct and concurrent validity study. *Brain Inj.* 25(2), 179–191 (2011).
- 46 Solomon GS, Kuhn A. Relationship between concussion history and neurocognitive test performance in National Football League draft picks. *Am. J. Sports Med.* 42(4), 934–939 (2014).
- 47 Van Kampen DA, Lovell MR, Pardini JE, Collins MW, Fu FH. The “value added” of neurocognitive testing after sports-related concussion. *Am. J. Sports Med.* 34(10), 1630–1635 (2006).
- 48 Broglio SP, Macciocchi SN, Ferrara MS. Sensitivity of the concussion assessment battery. *Neurosurgery* 60(6), 1050–1058 (2007).
- 49 Covassin T, Stearne D, Elbin R. Concussion history and postconcussion neurocognitive performance and symptoms in collegiate athletes. *J. Athl. Train.* 43(2), 119–124 (2008).
- 50 Schatz P, Pardini JE, Lovell MR, Collins MW, Podell K. Sensitivity and specificity of the ImPACT test battery for concussion in athletes. *Arch. Clin. Neuropsychol.* 21(1), 91–99 (2006).

- 51 Schatz P, Maerlender A. A two-factor theory for concussion assessment using ImPACT: memory and speed. *Arch. Clin. Neuropsychol.* 28(8), 791–797 (2013).
- 52 Schatz P, Sandel N. Sensitivity and specificity of the online version of ImPACT in high school and collegiate athletes. *Am. J. Sports Med.* 41(2), 321–326 (2013).
- 53 Chelune G, Naugle R, Luders H, Sedlak J, Awad I. Individual change after epilepsy surgery: practice effects and base-rate information. *Neuropsychology* 7(1), 41–52 (1993).
- 54 Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *J. Consult. Clin. Psychol.* 59(1), 12–19 (1991).
- 55 Bleiberg J, Garmoe W, Cederquist J, Reeves D, Lux W. Effects of dexedrine on performance consistency following brain injury: a double-blind placebo crossover case study. *Neuropsychiatry Neuropsychol. Behav. Neurol.* 6(4), 245–248 (1993).
- 56 Segalowitz SJ, Dywan J, Unsal A. Attentional factors in response time variability after traumatic brain injury: an ERP study. *J. Int. Neuropsychol. Soc.* 3(2), 95–107 (1997).
- 57 Van Zomeran AH, Brouwer WH. Head injury and concepts of attention. In: *Neurobehavioral Recovery from Head Injury*. Levin HS, Grafman J, Eisenberg HM (Eds). Oxford University Press, NY, USA, 398–415 (1989).
- 58 Rabinowitz AR, Arnett PA. Intraindividual cognitive variability before and after sports-related concussion. *Neuropsychology* 27(4), 481–490 (2013).
- 59 Wilson BA, Watson PC, Baddeley AD, Emslie H, Evans JJ. Improvement or simply practice? The effects of twenty repeated assessments on people with and without brain injury. *J. Int. Neuropsychol. Soc.* 6(4), 469–479 (2000).
- 60 Roach EB, Bleiberg J, Lathan CE, Wolpert L, Tsao J W, Roach RC. AltitudeOmics: decreased reaction time after high altitude cognitive testing is a sensitive metric of hypoxic impairment. *Neuroreport* 25(11), 814–818 (2014).
- 61 Iverson GL, Brooks BL, Ashton VL, Johnson LG, Gualtieri CT. Does familiarity with computers affect computerized neuropsychological test performance? *J. Clin. Exp. Neuropsychol.* 31(5), 594–604 (2009).
- 62 Cernich AN, Brennana DM, Barker LM, Bleiberg J. Sources of error in computerized neuropsychological assessment. *Arch. Clin. Neuropsychol.* 22(Suppl. 1), S39–S48 (2007).
- 63 Hancock PA, Sawyer BD, Stafford S. The effects of display size on performance. *Ergonomics* 58(3), 337–354 (2015).
- 64 Schmidt JD, Register-Mihalik JK, Mihalik JP, Kerr ZY, Guskiewicz KM. Identifying impairments after concussion: normative data versus individualized baselines. *Med. Sci. Sports Exerc.* 44(9), 1621–1628 (2012).
- 65 Echemendia RJ, Bruce JM, Bailey CM, Sanders JF, Arnett P, Vargas G. The utility of post-concussion neuropsychological data in identifying cognitive change following sports-related MTBI in the absence of baseline data. *Clin. Neuropsychol.* 26(7), 1077–1091 (2012).
- 66 Moser RS, Schatz P, Neidzowski K, Ott SD. Group versus individual administration affects baseline neurocognitive test performance. *Am. J. Sports Med.* 39(11), 2325–2330 (2011).
- 67 Cromer JA, Harel BT, Yu K *et al.* Comparison of cognitive performance on the Cogstate brief battery when taken in-clinic, in-group, and unsupervised. *Clin. Neuropsychol.* 29(4), 542–558 (2015).
- 68 Lowe M, Harris W, Kane RL, Banderet L, Levinson D, Reeves D. Neuropsychological assessment in extreme environments. *Arch. Clin. Neuropsychol.* 22(Suppl. 1), S89–S99 (2007).
- 69 Haran FJ, Dretsch MN, Bleiberg J. Performance on the Defense Automated Neurobehavioral Assessment across controlled environmental conditions. *Appl. Neuropsychol. Adult.* 23(6), 1–7 (2016).