HHS Public Access

Author manuscript

Am J Sports Med. Author manuscript; available in PMC 2016 August 01.

Published in final edited form as:

Am J Sports Med. 2015 August ; 43(8): 2018–2026. doi:10.1177/0363546515587714.

Rates and predictors of invalid baseline test performance in high school and collegiate athletes for three computerized neurocognitive tests (CNTs): ANAM, Axon, and ImPACT

Lindsay D. Nelson, PhD[1], **Adam Y. Pfaller, BS**[1], **Lisa E. Rein, MS**[2], and **Michael A. McCrea, PhD**[1]

[1]Department of Neurosurgery, Medical College of Wisconsin, Milwaukee, WI

[2]Department of Biostatistics, Medical College of Wisconsin, Milwaukee, WI

Abstract

Background—Preseason baseline testing is increasingly performed on athletes using computerized neurocognitive tests (CNTs). Adequate effort is critical to establish valid estimates of ability, yet many users do not evaluate performance validity, and the conditions that impact validity are not well understood across the available CNTs.

Purpose—We examined the rates and predictors of invalid baseline performance for three popular CNTs: ANAM (Automated Neuropsychological Assessment Metrics), Axon Sports, and ImPACT (Immediate Post-Concussion Cognitive Assessment and Testing).

Study Design—Cross-sectional study.

Methods—High school and collegiate athletes (N = 2,063) completed two of three CNTs each during pre-season evaluations. All possible pairings were present across the sample, and order of administration was randomized. Examiners gave one-on-one, scripted pre-test instructions emphasizing the importance of good effort. Profile validity was determined by the manufacturers' standard criteria.

Results—The overall percentage of tests flagged as of questionable validity was lowest for ImPACT (2.7%) and higher for ANAM and Axon (10.8% and 11.3%, respectively). The majority of invalid baselines were flagged as such due to failure on only one validity criterion. Several athlete and testing factors (e.g., attention deficit-hyperactivity disorder/ADHD, estimated general intellectual ability, administration order) predicted validity status for one or more CNTs. Considering only first CNT administrations and participants without ADHD and/or learning disability (n = 1,835) brought the rates of invalid baselines to 2.1%, 8.8%, and 7.0%, for ImPACT, ANAM, and Axon, respectively. Invalid profiles on the Medical Symptom Validity Test (MSVT) were rare (1.8% of subjects) and demonstrated poor correspondence to CNT validity outcomes.

Conclusion—These CNTs' validity criteria may not identify the same causes of invalidity or be equally sensitive to effort. The validity indicators may not be equally appropriate for some athletes (e.g., those with neurodevelopmental disorders). The data suggest that athletes do not put forth

Correspondence concerning this article can be addressed to Lindsay Nelson, 8701 West Watertown Plank Road, Milwaukee, WI 53226. linelson@mcw.edu.

widespread low effort or that some validity criteria are more sensitive to invalid performance than others. It is important for examiners to be aware of the conditions that maximize the quality of baseline assessments and to understand what sources of invalid performance are captured by the validity criteria they obtain.

Key Terms

baseline testing; computerized testing; neurocognitive testing; validity; concussion; mild traumatic brain injury; ImPACT; ANAM; Axon Sports

Introduction

Neurocognitive testing is widely recognized as an important component of the examination of concussed athletes, as it allows for the objective detection of subtle cognitive changes common in the acute post-concussive period.[2] Because of their numerous purported benefits over traditional paper-and-pencil tests, computerized neurocognitive tests (CNTs) have gained popularity with many athletic programs in recent years. A number of CNTs are commercially available, including ANAM (Automated Neuropsychological Assessment Metrics), the Axon Sports CNT (derived from CogState), and ImPACT (Immediate Post-Concussion and Cognitive Testing).[6, 9, 19, 25, 37] Among the potential advantages of these programs are: (1) widely accessible Internet-based/electronic platforms, (2) highly standardized test administration and scoring procedures, (3) ready access to numerous alternate test forms (e.g., by automatic selection of items from a larger test bank) for use with repeat testing, and (4) storage of performance data in centralized data repositories for the benefit of users and ongoing test development efforts.[12, 31]

Another major reason why CNTs have become so commonly used is that they make it operationally feasible for athletic programs to perform widespread pre-season baseline testing of their athletes. Although there is some debate regarding the value that baseline testing adds to post-injury assessments,[16, 38, 39] pre-season baseline testing is undoubtedly an increasingly common practice, with 94.7% of athletic trainers who use ImPACT reporting that they baseline test their athletes.[13] Clearly, having baseline performance data on injured athletes allows clinicians to take into account premorbid cognitive skills, yet a number of questions have also been raised regarding how baseline scores are obtained in practice and whether athletes' motivations preclude valid assessments. For example, the group settings in which athletes commonly complete baseline testing negatively affect performance for some athletes,[23, 30] and the availability and ease-of-use of CNTs for baseline testing has probably led to the tools being used by individuals who are inadequately trained in psychometrics and test interpretation.[29]

Yet even under optimal testing conditions and with the most experienced examiners, many athletes may be unmotivated to perform their best at baseline testing given that, if concussed, they will have to achieve comparable CNT scores before being cleared to return to play. Using well-established paper-and-pencil neuropsychological measures, for example, 11% of high school football players in one sample failed formal effort tests.[22] Although invalid baseline rates have been highly variable across CNT studies, some have found over a

quarter (27.9–30.3%) of athletes' baseline CNT profiles to be flagged as of questionable validity.[17, 43] Consequently, it is critical that examiners take steps to maximize the validity of baseline test scores. All of the major CNT publishers readily provide flags of potentially poor effort, yet these indices are not readily used by many examiners (e.g., in the survey of athletic trainers using ImPACT mentioned above, only 51.9% of the athletic trainers who baseline test their athletes indicated that they examine these baselines' validity output).[13] Further, the relative sensitivity of the various CNT's validity criteria and the individual factors that may inadvertently affect validity profiles are not well established.

Overall rates of invalid baselines on ImPACT have been reported by several authors, and these rates have varied considerably across studies. At the low end, Moser and colleagues[30] reported a failure rate of 0.6% in a sample of high school athletes who were scheduled and accompanied by their parents to an individual testing session in which performance feedback was provided; another sample yielded 27.9%[43] invalid baselines. While it is somewhat difficult to make inferences across studies about the role of certain demographic factors (e.g., age) in rates of invalid baselines (due to evolution of validity criteria over time and differences in testing protocols across studies), the overall invalid percentages in high school, collegiate, and professional (National Football League) samples have been reported to range from 0.4–11.9% (high school),[10, 11, 14, 21, 23, 30, 36] 4.1–27.9% (collegiate),[33, 36, 43] and 2.2–5.4% (professional),[40, 41] respectively. Probably more important than age or level of competition is testing environment (with large groups yielding more invalid baselines),[30] age by group size interactions (with athletes age 10–12 years old more sensitive than those 13–18 to the effects of group size),[23] and risk factors for distractibility such as attention deficit-hyperactivity disorder.[36] Invalid rates for Axon have varied widely (1.0% in an University boxing sample,[28] 4.7% in a high school sample,[24] 6.1% in Norwegian soccer players,[42] and 30.3% in Division I University athletes[22]). To our knowledge, rates of invalid baseline performance have not been reported in an athlete sample for ANAM.

Perhaps with more extensive documentation of the performance and properties of the various CNT's validity criteria will come increased adherence to guidelines mandating practitioners to estimate the validity of each baseline assessment.[3, 29] In this study, we examined the rates and predictors of invalid performance for three popular CNTs—ANAM, Axon, and ImPACT—obtained at pre-season baseline evaluations in the same sample, with the aim of informing researchers and clinicians about their properties such that they may be more readily utilized by clinicians. Given the literature reviewed, we hypothesized that ADHD (and perhaps learning disability) would be associated with higher rates of invalidity for the three CNTs. However, the majority of the measures explored in this study were novel and, consequently, the analyses necessarily exploratory. Similarly, this was the first study to directly compare the rates of invalid baselines for these three CNTs, it was unclear to what extent we would observe differences in overall rates of invalidity across the three CNTs (one prior study on a military sample reported invalid rates for ANAM, Axon, and ImPACT of 3.8%, 7.0%, and 12.0%, respectively, although the group *n*s were small and no direct statistical comparison of these rates reported).[11]

Materials and Methods

Participants

As part of a larger study on the assessment of sport-related concussion, contact and collision sport student-athletes from 9 high schools and 4 colleges in southeastern Wisconsin completed baseline testing between August, 2012 and October, 2014. Informed consent was obtained for 2,154 participants. Participants (N = 2,063) who completed at least one CNT were included in the sample for these analyses. On rare occasions there were problems with the administration of the CNTs or athletes did not complete the entire baseline session, and as a result 24 (1.2%) of these participants only have data for one CNT. Sample demographics are summarized in Table 1. Adult athletes and parents of minor athletes completed informed consent, and minor participants completed assent prior to baseline testing. Participants were compensated $30 for their time and effort in completing baseline assessments. All testing procedures were approved by the Institutional Review Board at the Medical College of Wisconsin.

Baseline Testing Session

The baseline testing protocol consisted of, in order: Contact Information, Demographics/ Health History (gathered by one-on-one interview), Wechsler Test of Adult Reading (WTAR),[44] CNT #1, Standard Assessment of Concussion (SAC),[26] Sport Concussion Assessment Tool – 3rd edition (SCAT3) symptom checklist,[27] CNT #2, Green's Medical Symptom Validity Test (MSVT),[20] Satisfaction With Life Scale (SWLS),[15] Brief Symptom Inventory-18 (BSI-18),[14] and the Balance Error Scoring System (BESS).[21] Tests were individually proctored by a research assistant in quiet settings with computers positioned to minimize distractions. Testing groups sizes ranged from 1–20 athletes. Each athlete was read a standardized script at the beginning of the baseline testing session and before each of the CNTs about the importance of valid baseline tests (see Appendix). Testing sessions lasted approximately 90 minutes.

Each athlete took two of three CNTs: Automatic Neuropsychological Assessment Metrics (ANAM v. 4.3; Vista Life Sciences), Axon Sports (Axon; Axon Sports, Inc.), and Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT, Online version; ImPACT Applications Inc.). CNT pairing groups were assigned to each school with the aim of balancing the demographic distribution across CNTs. The overall distribution of CNT pairings (separately by order) across the sample was: 16.2% ANAM-Axon, 14.3% Axon-ANAM, 17.4% ANAM-ImPACT, 16.2% ImPACT-ANAM, 17.9% Axon-ImPACT; 18.0% ImPACT-Axon. For each subject, order of administration was selected at random by a computer algorithm. The reliability and validity of ANAM,[6, 11] Axon,[11, 25] and ImPACT[8, 9, 11, 33] have been reported elsewhere.

Computerized Neurocognitive Tests

ANAM—The ANAM test battery was originally developed by the Department of Defense (DoD) for the assessment of processing speed and cognitive efficiency. It has been used in pre- and post-deployment evaluations and has been adapted for the assessment of sport-related concussion.[7] The version used in this study included eight subtests: Simple Reaction

Time (SRT), Code Substitution-Learning (CDS), Procedural Reaction Time (PRO), Mathematical Processing (MTH), Matching to Sample (M2S), Code Substitution-Delayed (CDD), Simple Reaction Time 2 (SR2), and Go/No-Go (GNG).

Axon—The Axon Sports Computerized Cognitive Assessment Tool (CCAT, or CNT to be consistent with the terminology used in this report) was developed by CogState and has also been referred to as CogSport and the CogState Brief Battery. The test is comprised of four tasks: Processing Speed (PS; simple reaction time), Attention (AT; choice reaction time), Learning (LN; visual recognition memory) and Working Memory (WM; one-back), which map onto the CogState tasks of Detection, Identification, Visual (One Card) Learning, and One-Back, respectively.

ImPACT—ImPACT was developed for the baseline and post-injury assessment of concussed athletes and is comprised of six tasks, Word Memory, Design Memory (DM), X's and O's, Symbol Match, Color Match, and Three Letters (TL), which yield the following composite scores: Verbal Memory, Visual Memory, Visual Motor Speed, Reaction Time, and Impulse Control.

Validity Criteria

Test validity was determined by the standard output available from each CNT manufacturer, with the most updated versions of the validity criteria at the time of data analysis used for all participants. The criteria for each CNT are presented in Table 2. ImPACT indicates invalid baseline scores using a mark ("++") on the report with a note indicating that the data may be invalid. ANAM provides profile validity indices in a separate effort report document and divides its criteria into two categories: "Consider Retest" for accuracy scores so low (below or equal to chance) as to suggest poorly understood instructions, and a "Questionable Effort" checkbox that reflects poor performance on an index developed to discriminate between individuals putting forth good effort versus those who were instructed to put forth suboptimal effort.[32] Any profile with one or both of these criteria checked was coded as invalid for our purposes.

For Axon, the primary "Integrity Check" criteria most prominent in the report output are listed in Table 3, although the test also flags profiles as of questionable validity due to rare scores (i.e., test scores more than two standard deviations below the mean). Because of the high degree of overlap between this criterion and the Integrity Check criteria, and the fact that it only independently accounted for a small number of invalid profiles ($n = 4$, or 2.6% of invalid Axon profiles), this was excluded from the table but is discussed in the Results.

Statistical Analysis

Overall percentages of invalid baseline test performance were computed for each CNT along with 95% confidence intervals. For those profiles that were flagged as invalid, we then computed the frequency of failure on each individual validity criterion. Predictors of baseline validity status (valid, invalid) were explored using simple logistic regression for each CNT given the high degree of overlap among some predictors explored (Table 4). Predictors considered included sex, age, sport, attention deficit-hyperactivity disorder

(ADHD), learning disability, number of prior concussions, grade point average (GPA), estimated verbal intellectual ability (WTAR word reading standard score), and order of CNT administration (1st or 2nd CNT taken in the testing session). Variables related to participant demographics and history (GPA, ADHD, learning disability, prior concussions) were gathered via self-report. As a large number of participants completing ImPACT had taken the test before (56.2%), prior exposure to ImPACT was also explored as a predictor of baseline validity for this measure (the percentage of participants who had previously taken ANAM and Axon were too low —0.1% and 1.1%—to include this variable in the analysis of those tests). Because of the exploratory nature of these analyses and the multiple comparisons that were performed for each measure, we applied the false discovery rate control method[4] and indicate in Table 4 which findings remained significant after applying this correction. This approach is a sequential Bonferroni-type procedure that, unlike traditional Bonferroni correction (which controls the familywise error rate), is aimed at controlling the expected proportion of incorrectly rejected null hypotheses ("false discoveries") and, consequently, better preserves statistical power while also providing a reasonable degree of control of type I errors.[4, 5]

As is described in the Results, analyses were also conducted to directly statistically compare the three CNTs on relevant variables (e.g., frequency of invalid baselines) using generalized estimating equations (GEE) with logit link function to account for repeated measures. In addition, multiple logistic regression was conducted to compare the three CNTs for the subset excluding ADHD, LD, and second test order; variable selection was based on univariate GEE analyses using an inclusion criteria of $p < 0.05$. First order interactions among the significant variables were examined and none of them were significant.

Results

Percentage of Invalid Baselines by CNT

Table 3 presents the overall percentage of invalid profiles by CNT. Overall, the percentage of athletes who produced invalid profiles was lowest for ImPACT (2.7%, 95% *CI* = 1.9, 3.7) and highest for ANAM (10.7%, 95% *CI* = 9.9, 12.5) and Axon (11.3%, 95% *CI* = 9.7, 13.1). Statistical comparisons of validity status for each CNT pair (using GEE with logit link function to account for repeated measures) revealed equivalent odds of producing an invalid baseline for ANAM vs. Axon (*OR* = .95, *p* = .686) and higher odds for Axon and ANAM versus ImPACT (ANAM vs. ImPACT *OR* = 4.20, Axon vs. ImPACT *OR* = 4.41, *p*s < .001). Of the 2,039 participants who completed two CNTs, 1,731 (84.9%) produced valid profiles on both CNTs administered, while 284 (13.9%) produced one invalid CNT, and only 24 (1.2%) produced invalid profiles on both CNTs completed. Considering CNT and MSVT validity jointly, only 1 participant failed all three tests, while 34 (1.7%) failed two and 296 (14.5%) failed one test.

Table 3 also lists the rates of failure on each validity criterion as well as the number of validity criteria marked as failed for each CNT. Across ANAM, Axon, and ImPACT, the overwhelming majority (79–90%) of participants who produced an invalid baseline did so because of failure on only one of the test's core validity criteria. As mentioned earlier, Axon also flags profiles as of questionable validity due to test scores that are rare normatively (> 2

SD below the mean). Failure on this criterion overlapped significantly with the criteria listed in Table 3 (only 4, or 2.6% of invalid Axon profiles, were flagged as invalid solely due to this additional criterion). Similarly, only 2.1% of ANAM baselines were flagged as invalid solely due to the Questionable Effort criterion.

ANAM profiles were most likely to be flagged as invalid due to failure on the mathematical processing (23.6% of invalid profiles), matching to sample (35.0%), and code substitution-delayed (38.6%) tests. For Axon, the most common criterion failed was that for learning accuracy (83.0% of invalid profiles), while for ImPACT, the most common criteria failed were those for design memory learning (46.2% of invalid profiles) and three letters total correct (43.6%).

Predictors of Invalid CNT Profiles

Table 4 presents the extent to which various personal and test administration variables predicted the validity status of each CNT profile. Lower GPA predicted invalid baselines for all tests. Similarly, lower WTAR standard score and a history of ADHD[1] predicted invalid baseline performance for ANAM and Axon, and learning disability additionally predicted invalid test profiles for ANAM. In fact, 25.9% of athletes with ADHD (vs. 9.9% without ADHD) produced invalid baselines on Axon. Similarly, 20.0% of participants with ADHD (versus 10.0% without), and 27.3% of participants with a learning disability (versus 10.0% without), achieved invalid baselines on ANAM). Further, participants who took Axon as their 2nd CNT had a higher odds (OR = 1.91, 95% CI = 1.35–2.72) of achieving an invalid profile than those who took Axon first. For ImPACT, lower age (and playing at the high school level) were modestly associated with higher odds of invalidity, but these predictors became nonsignificant after adjustment for multiple comparisons. Select sports showed differential patterns of validity for Axon and ImPACT, but these associations were no longer significant after accounting for group differences in other variables (ADHD, GPA, WTAR). Gender and prior history of concussion (all CNTs) were not predictive of validity status on any CNT and, similarly, familiarity with ImPACT was not predictive of validity status.

Given the influence of test order, ADHD, and LD on some of the CNTs, the overall rate of invalid baselines was recomputed for each CNT excluding 2nd test order and participants with ADHD and/or LD (n = 1,835) to yield a fairer comparison of expected baseline rates under more typical testing conditions in the majority of athletes. This yielded an invalid baseline frequency of 8.8% for ANAM, 7.0% for Axon, and 2.1% for ImPACT. Direct comparison of validity status, excluding ADHD, LD, and second test order and while additionally adjusting for GPA and WTAR score (as these factors predicted validity status) yielded similar overall findings to unadjusted estimates, with ANAM and Axon showing equivalent odds of producing invalid baselines (OR = .92, p = .522) that were both higher than those for ImPACT (ANAM vs. ImPACT OR = 4.36, Axon vs. ImPACT OR = 4.73, ps < .001).

[1]43% of participants in this sample with a history of ADHD reported currently prescribed a stimulant medication

Comparison of CNT Validity to MSVT

The majority (98%) of the sample also completed the MSVT as an additional measure of effort. Failure of one or more effort indices on the MSVT was rare (n = 36, or 1.8%, of all participants). Exploratory analyses were conducted to examine the overlap in validity profiles between the CNTs and MSVT. Agreement between profile validity for each CNT and the MSVT was poor. For example, all 35 subjects who produced invalid ImPACT profiles passed the MSVT, while all 26 participants who failed the MSVT but completed ImPACT produced valid ImPACT profiles (Cohen's kappa = −.02). Agreement was similarly low when comparing profile validity for the MSVT and ANAM (kappa = .01) and the MSVT and Axon (kappa = .08).

Discussions

Establishing valid estimates of athletes' premorbid cognitive abilities is critical to maximizing the utility of pre-season baseline testing for concussion management programs. Although computerized neurocognitive tests (CNTs) facilitate the estimation of performance validity using embedded measures that are readily available to examiners, prior work suggests that practitioners underutilize these indices. Here, we reported the rates and predictors of invalid baseline test performance for three popular CNTs—ANAM, Axon, and ImPACT—gathered within the same sample of athletes, with the aim of informing users of these tests about the performance and properties of each CNT's standard validity indices.

The overall rate at which these profiles were flagged as invalid varied by test, with the percentage of invalid ANAM and Axon profiles higher than ImPACT. This finding may have several implications which are not readily teased apart using these data. On the one hand, differences in rates of invalid baselines may be due to variably stringent validity criteria across CNTs. Probably more accurate is that the validity criteria for each test are differentially sensitive to differing sources of invalidity. Test scores may be invalid for a variety of reasons, including technological issues during test administration, reading/comprehension problems, fatigue, environmental distractions, and low motivation (which itself reflects a continuum and includes both individuals who are not highly motivated to try their best as well as those, probably rarer, athletes who intentionally underperform or "sandbag" to a high degree). Further, legitimate pre-morbid cognitive difficulties or neurodevelopmental disorders (e.g., ADHD, learning disability) may cause some well-intentioned athletes giving full effort to fail performance validity measures, particularly when those measures are highly stringent or sensitive to the cognitive difficulties associated with these disorders.

As outright sandbagging (i.e., putting forth low effort or intentionally working to produce deflated scores) is probably relatively rare,[11, 23] lenient criteria that only require test accuracy to exceed chance levels would probably most commonly capture issues related to technical issues, poor comprehension of test instructions (including left-right confusion on tests that are susceptible to this), and guessing. Each CNT used in this study contain some criteria that appear to be capable of capturing these types of issues. However, it may be that more stringent or sophisticated criteria are needed to capture more subtle sources of invalid performance such as lower levels of suboptimal effort or periodic distraction.

ANAM explicitly provides an empirically derived index (labeled Questionable Effort) aimed at identifying individuals who are knowingly putting forth poor effort. Axon, while not explicitly labeling its criteria so clearly, also contains requirements more stringent than those requiring chance performance, including criteria that appear to be aimed at capturing inconsistent and unusual patterns of performance (which may indicate issues with engagement in or effort on the test). This is consistent with the language available in the Axon manual which suggests that the criteria used to establish profile validity are aimed at detecting low effort in addition to factors that would have more obvious/catastrophic influence on performance. The ImPACT manual is less clear with regard to the sources of invalidity targeted by its validity criteria, but its list of common sources of invalid performance (failure to read directions, learning or attentional disorders, excessive fatigue, horseplay, and left-right confusion[2])[1] might suggest that its criteria are aimed at identifying more obvious sources of poor performance and could explain the lower rate of invalid baselines observed in our sample. However, we also cannot rule out the possibility that our participants' effort was truly better on this CNT versus the others. Further, prior work suggests that intentional underperforming ("sandbagging") is relatively difficult to accomplish without detection on ImPACT.[18, 34]

Most invalid baselines were flagged as such because of failure on only one of several validity criteria, supporting the idea that athletes do not broadly sandbag their evaluations. Also consistent with this idea was that failure on the MSVT was quite rare (1.8% of the sample), and across the three tests containing validity measures taken by each athlete measures (two CNTs and the MSVT), only 1 athlete (.05%) produced invalid profiles on all three tests. Instead of sandbagging, some athletes may (1) display continual, mildly suboptimal performance (picked up only by the more stringent criteria) or (2) put forth suboptimal performance (including low effort or unintentional loss of focus) selectively. In the current sample, each CNT had a subset of validity criteria that were systematically more likely to be failed than others. Our review of these data suggests that the more difficult tests tended to be flagged more often and is consistent with both possibilities (1) and (2) above in that the criteria (e.g., ANAM MTH, M2S, and CDD; Axon LN, and ImPACT DML, TL) for these tests may be more sensitive to low effort (alternatively, test-takers may be more likely to be overwhelmed by and give up on these more challenging or confusing tests). That the failure rate of Axon's Integrity Check for Learning Accuracy was substantially higher than Working Memory accuracy (despite both criteria only requiring accuracy over 53%), for example, could suggest that some athletes find the Learning task too difficult, confusing, or overwhelming to put forth adequate effort and meet this minimal performance threshold. This finding underscores the importance of providing encouragement to examinees to do their best even on challenging tasks and illustrates the notion that the degree to which tests scores are valid may vary from task-to-task (or, moment-to-moment) even within a single CNT administration.

[2]Although left-right confusion is listed in the Online ImPACT manual cited, this is apparently not considered a prominent issue on the online version of test (which allows for right and left responses from each hand on a keyboard) and was more prevalent for the prior desktop version of ImPACT (which required right and left mouse clicks for responses to one reaction time task). **36**. Schatz P, Moser RS, Solomon GS, Ott SD, Karpf R. Prevalence of invalid computerized baseline neurocognitive test results in high school and collegiate athletes.

Finally, our evaluation of predictors of invalid baseline performance revealed some important insights. First, the validity indicators may not be appropriate or equally meaningful for some populations (e.g., those with ADHD, especially for ANAM and Axon where a large minority of participants with ADHD produced invalid profiles). We cannot rule out that these participants failed validity criteria at a higher rate because of lower motivation/effort. One way to tease this out may be to see how frequently individuals with ADHD (who failed their first baseline test) produce valid profiles given another testing opportunity. As most athletes that produced invalid baselines did not complete repeat baseline evaluations, we cannot comment on the effect this may have had. If the ANAM and Axon validity criteria are indeed more stringent (sensitive), then detecting unintended factors (i.e., bona fide cognitive impairment) would be an expected natural consequence of more stringent criteria for validity.

Although we did not record data on testing group size to a degree needed for analysis of this variable, given prior findings on the relevance of this variable in CNT validity,[23, 30] it would be valuable for future studies to record this and explore interactions among testing conditions (e.g., group size) and other individual difference variables (e.g., ADHD) in order to clarify the athletes who may be more or less vulnerable to the impact of such testing conditions. Until then, it would be wise to follow standard recommendations to administer baseline tests individually or in very small, distraction-free groups especially for those athletes who are most at risk of producing invalid tests. Of course, given the exploratory nature of our analyses, it will also be important to replicate our findings on new samples and to explore the extent to which procedures used in this study (e.g., monetary incentive for participation) influence athlete motivation and baseline test performance. Given the poor agreement between validity status for the CNTs and MSVT and possibility that these tests' validity criteria tap different aspects of ability and effort, it would also be valuable to perform more systematic manipulation of examinee instructions and other testing factors in order to better identify what is being tapped by the various validity criteria applied across these tests as a means to further refine these measures and facilitate the interpretation of validity output.

Our results demonstrate that the vast majority of athletes are capable of producing valid baseline tests given the proper testing conditions. This is consistent with other published reports, including prior work finding that most athletes (87%) who produce an invalid baseline obtain a valid profile given a reassessment.[35] Similarly, although we did not require select athletes (in particular those with ADHD, learning disability, and/or low WTAR scores) to repeat invalid baselines, 73.5% of the athletes in our sample who did repeat testing after an invalid baseline achieved a valid profile on the second attempt (with roughly equal percentages across ANAM, Axon, and ImPACT: 75.4%, 71.7%, and 73.3%). Overall, our rate of invalid test performance was low relative to some published estimates, especially when considering only the first test administrations (which would better match a typical athlete evaluation). Given the published literature on this topic and common assumptions in neuropsychological practice, we might presume that several procedural factors facilitated this, including (1) individually-proctored tests,[23, 30] (2) scripted test instructions emphasizing the importance of good effort, (3) controlled testing conditions (i.e. tests administered in quiet, supervised settings), and (4) payment for study participation. Our data

also provide evidence that performance is optimized when (5) athletes are not cognitively fatigued. While we recognize that financial payment is an incentive unique to the experimental setting, because it was a constant across all examinations, we believe that this component of our protocol did not negate our ability to draw comparisons between the CNTs. These findings reinforce, as others have articulated,[3, 29] that pre-season evaluations should be administered under conditions that will maximize athletes' interest in and performance on baseline tests, which could feasibly include all factors listed above (with the exception of #4). Although incorporating some of these factors (e.g., individually-proctored tests) into routine practice would negate some of the purported advantages of CNTs, it remains important for test users to be aware that decisions around how baseline tests are administered has important implications for their potential value in later post-injury assessments.

Acknowledgement

This work was supported by the U.S. Army Medical Research and Materiel Command under award number W81XWH-12-1-0004. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army. This publication was also supported by the Clinical and Translational Science Institute grant 1UL1-RR031973 (−01) and by the National Center for Advancing Translational Sciences, National Institutes of Health grant 8UL1TR000055. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. We thank Kwang Woo Ahn, PhD, for his input on the statistical analyses.

References

1. Technical manual: Online ImPACT 2007–2012. ImPACT Applications, Inc.; Available at: https://www.impacttest.com/pdf/ImPACTTechnicalManual.pdf.
2. Aubry M, Cantu R, Dvorak J, et al. Summary and agreement statement of the 1st International Symposium on Concussion in Sport, Vienna 2001. Clin J Sport Med. 2002; 12(1):6–11. PMID: 11854582. [PubMed: 11854582]
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. 2012; 27(3):362–373. PMID: 22382386. [PubMed: 22382386]
4. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995; 57:289–300. PMID.
5. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics. 2001; 29:1165–1188. PMID.
6. Bleiberg J, Cernich AN, Cameron K, et al. Duration of cognitive impairment after sports concussion. Neurosurg. 2004; 54(5):1073–1078. PMID: 15113460.
7. 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. 2000; 14(3):287–294. PMID: 11262703. [PubMed: 11262703]
8. Broglio SP, Ferrara MS, Macciocchi SN, Baumgartner TA, Elliott R. Test-retest reliability of computerized concussion assessment programs. J Athl Train. 2007; 42(4):509–514. PMID: 18174939. [PubMed: 18174939]
9. Broglio SP, Macciocchi SN, Ferrara MS. Sensitivity of the concussion assessment battery. Neurosurg. 2007; 60(6):1050–1057. PMID: 17538379.
10. Brooks BL, Mrazik M, Barlow KM, McKay CD, Meeuwisse WH, Emery CA. Absence of differences between male and female adolescents with prior sport concussion. J Head Trauma Rehabil. 2014; 29(3):257–264. PMID: 24413074. [PubMed: 24413074]

11. Cole WR, Arrieux JP, Schwab K, Ivins BJ, Qashu FM, Lewis SC. Test-retest reliability of four computerized neurocognitive assessment tools in an active duty military population. Arch Clin Neuropsychol. 2013; 28(7):732–742. PMID: 23819991. [PubMed: 23819991]
12. Collie A, Darby D, Maruff P. Computerised cognitive assessment of athletes with sports related head injury. Br J Sports Med. 2001; 35(5):297–302. PMID: 11579059. [PubMed: 11579059]
13. Covassin T, Elbin RJ 3rd, Stiller-Ostrowski JL, Kontos AP. Immediate post-concussion assessment and cognitive testing (ImPACT) practices of sports medicine professionals. J Athl Train. 2009; 44(6):639–644. PMID: 19911091. [PubMed: 19911091]
14. Derogatis, LR. Brief Symptom Inventory 18 (BSI-18): Administration, Scoring, and Procedures Manual. Bloomington, MN: Pearson; 2001.
15. Diener E, Emmons RA, Larsen RJ, Griffin S. The Satisfaction With Life Scale. J Pers Assess. 1985; 49(1):71–75. PMID: 16367493. [PubMed: 16367493]
16. 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. 2012; 26(7):1077–1091. PMID: 23003560. [PubMed: 23003560]
17. Eckner JT, Kutcher JS, Richardson JK. Between-seasons test-retest reliability of clinically measured reaction time in National Collegiate Athletic Association Division I athletes. J Athl Train. 2011; 46(4):409–414. PMID: 21944073. [PubMed: 21944073]
18. Erdal K. Neuropsychological testing for sports-related concussion: how athletes can sandbag their baseline testing without detection. Arch Clin Neuropsychol. 2012; 27(5):473–479. PMID: 22684033. [PubMed: 22684033]
19. Erlanger D, Feldman D, Kutner K, et al. Development and validation of a web-based neuropsychological test protocol for sports-related return-to-play decision-making. Arch Clin Neuropsychol. 2003; 18(3):293–316. PMID: 14591461. [PubMed: 14591461]
20. Green's Medical Symptom Validity Test for Windows [computer program]. Version. Edmonton, Alberta, Canada: Green's Publishing, Inc; 2003.
21. Guskiewicz KM. Postural stability assessment following concussion: one piece of the puzzle. Clin J Sport Med. 2001; 11(3):182–189. PMID: 11495323. [PubMed: 11495323]
22. Hunt TN, Ferrara MS, Miller LS, Macciocchi S. The effect of effort on baseline neuropsychological test scores in high school football athletes. Arch Clin Neuropsychol. 2007; 22(5):615–621. PMID: 17507199. [PubMed: 17507199]
23. Lichtenstein JD, Moser RS, Schatz P. Age and test setting affect the prevalence of invalid baseline scores on neurocognitive tests. Am J Sports Med. 2014; 42(2):479–484. PMID: 24243771. [PubMed: 24243771]
24. MacDonald J, Duerson D. Reliability of a computerized neurocognitive test in baseline concussion testing of high school athletes. Clin J Sport Med. 2014 PMID: 25061807.
25. 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. 2009; 24(2):165–178. PMID: 19395350. [PubMed: 19395350]
26. McCrea M, Kelly JP, Kluge J, Ackley B, Randolph C. Standardized assessment of concussion in football players. Neurology. 1997; 48(3):586–588. PMID: 9065531. [PubMed: 9065531]
27. McCrory P, Meeuwisse W, Aubry M, et al. Consensus statement on concussion in sport--the 4th International Conference on Concussion in Sport held in Zurich, November 2012. Clin J Sport Med. 2013; 23(2):89–117. PMID: 23478784. [PubMed: 23478784]
28. Moriarity JM, Pietrzak RH, Kutcher JS, Clausen MH, McAward K, Darby DG. Unrecognised ringside concussive injury in amateur boxers. Br J Sports Med. 2012; 46(14):1011–1015. PMID: 22547563. [PubMed: 22547563]
29. Moser RS, Schatz P, Lichtenstein JD. The importance of proper administration and interpretation of neuropsychological baseline and postconcussion computerized testing. Appl Neuropsychol Child. 2014; 0:1–8. PMID: 24236894.

30. Moser RS, Schatz P, Neidzwski K, Ott SD. Group versus individual administration affects baseline neurocognitive test performance. Am J Sports Med. 2011; 39(11):2325–2330. PMID: 21828367. [PubMed: 21828367]
31. Rahman-Filipiak AAM, Woodward JL. Administration and environment considerations in computer-based sports-concussion assessment. Neuropsychol Rev. 2014; 23:314–334. PMID: 24306286. [PubMed: 24306286]
32. Roebuck-Spencer TM, Vincent AS, Gilliland K, Johnson DR, Cooper DB. Initial clinical validation of an embedded performance validity measure within the automated neuropsychological metrics (ANAM). Arch Clin Neuropsychol. 2013; 28(7):700–710. PMID: 23887185. [PubMed: 23887185]
33. Schatz P. Long-term test-retest reliability of baseline cognitive assessments using ImPACT. Am J Sports Med. 2010; 38(1):47–53. PMID: 19789333. [PubMed: 19789333]
34. Schatz P, Glatts C. "Sandbagging" baseline test performance on ImPACT, without detection, is more difficult than it appears. Arch Clin Neuropsychol. 2013; 28(3):236–244. PMID: 23403552. [PubMed: 23403552]
35. Schatz P, Kelley T, Ott SD, et al. Utility of repeated assessment after invalid baseline neurocognitive test performance. J Athl Train. 2014; 49(5):659–664. PMID: 25162778. [PubMed: 25162778]
36. Schatz P, Moser RS, Solomon GS, Ott SD, Karpf R. Prevalence of invalid computerized baseline neurocognitive test results in high school and collegiate athletes. J Athl Train. 2012; 47(3):289–296. PMID: 22892410. [PubMed: 22892410]
37. 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. 2006; 21(1):91–99. PMID: 16143492. [PubMed: 16143492]
38. Schatz P, Robertshaw S. Comparing post-concussive neurocognitive test data to normative data presents risks for under-classifying "above average" athletes. Arch Clin Neuropsychol. 2014 PMID: 25178629.
39. Schmidt JD, Register-Mihalik JK, Mihalik JP, Kerr ZY, Guskiewicz KM. Identifying Impairments after concussion: normative data versus individualized baselines. Med Sci Sports Exerc. 2012; 44(9):1621–1628. PMID: 22525765. [PubMed: 22525765]
40. Solomon GS, Haase RF. Biopsychosocial characteristics and neurocognitive test performance in National Football League players: An initial assessment. Arch Clin Neuropsychol. 2008; 23:563–577. PMID: 18614333. [PubMed: 18614333]
41. Solomon GS, Haase RF, Kuhn A. The relationship among neurocognitive performances and biopsychosocial characteristics of elite National Football League draft picks: An exploratory investigation. Arch Clin Neuropsychol. 2013; 28:9–20. PMID: 23220623. [PubMed: 23220623]
42. Straume-Naesheim TM, Andersen TE, Bahr R. Reproducibility of computer based neuropsychological testing among Norwegian elite football players. Br J Sports Med. 2005; (39 Suppl 1):i64–i69. iPMID: 16046358. [PubMed: 16046358]
43. Szabo AJ, Alosco ML, Fedor A, Gunstad J. Invalid performance and the ImPACT in national collegiate athletic association division I football players. J Athl Train. 2013; 48(6):851–855. PMID: 24151810. [PubMed: 24151810]
44. Wechsler, D. Wechsler Test of Adult Reading: WTAR. San Antonio, TX: The Psychological Corporation; 2001.

APPENDIX

Introduction speech

Hello and welcome to baseline testing. We are conducting baseline testing so that if you do happen to get a concussion while playing your sport, we will test you after your concussion and compare the two test scores.

Today, we'll be testing your memory, concentration, and processing speed. We'll also be asking you questions about how you are feeling physically and mentally.

We ask you to give your best effort on these tests because it is imperative to the purpose of this study. If we do not have valid test results, we will not be able to accomplish the goals of this study. All of these tests have built in checks to let us know if you are giving your best effort.

If you do happen to have a concussion while playing your sport, we will not be involved in determining if and when you are ready to return to play – that will be up to your medical staff at your school.

You will be paid for your participation today but only upon the completion of a valid baseline test. If your baseline tests are found to be invalid, we will contact you and your parents to discuss further.

The tests are not very exciting but please give your best effort!

Read to each subject at the start of EVERY CNT

This is a cognitive test that assesses your memory and concentration and takes about 30 minutes.

Try to be as fast and accurate as possible. Do not disturb others while being tested and do not talk while taking the test. Some of the subtests are very similar, so be sure to pay attention and give your best effort, even if you think you have done the test before

Read all instructions carefully and be sure to pay attention to how you are supposed to respond

Once again – please be as fast and accurate as possible. Also, this is one of those tests that have a built in check to know if you are giving your best effort, so please try your best.

What is known about the subject

Obtaining valid pre-season baseline performance from athletes is critical to estimate abilities for use in later post-concussion evaluations, and published computerized neurocognitive tests (CNTs) facilitate the estimation of performance validity through embedded measures. However, clinicians often overlook the importance of estimating profile validity, and the sensitivity and predictors of performance validity are not well documented for the range of available computerized neurocognitive tests.

What this study adds to existing knowledge

As one of the first studies to obtain baseline data on three popular CNTs within a single large sample alongside a wealth of other athlete information, these data significantly advance what is known about the individual and test-related factors that are associated with the validity of baseline assessments and provides important data for test users regarding the properties of the validity scales for these three CNTs.

Table 1

Sample characteristics (N = 2,063)

	***M* (*SD*) or %**
Sex (male)	76.8%
Age	17.8 (1.9)
Level of competition (college)	60.0%
Race	
Caucasian	83.4%
African American	12.1%
Asian	1.2%
Native Hawaiian/Pacific Islander	0.6%
American Indian/Alaska Native	0.5%
Unknown	2.1%
ADHD	7.8%
Learning disability	3.6%
GPA	3.28 (.52)
WTAR SS	101.43 (12.62)
Number of prior concussions	.64 (.94)
Sport	
Football	48.3%
Soccer	32.6%
Lacrosse	7.4%
Wrestling	3.7%
Ice hockey	3.7%
Wrestling	2.5%
Field hockey	1.8%

Note. ADHD = attention deficit-hyperactivity disorder; GPA = grade point average; WTAR SS = Wechsler Adult Test of Reading Standard Score

Table 2

Validity criteria for ANAM, Axon, and ImPACT

ANAM	Axon (CogSport)	ImPACT
Consider Retest	Integrity Checks	Impulse Control Composite (ICC) > 30
Simple Reaction Time (SRT) accuracy < 56%	Processing Speed (PS) accuracy > 80%	Word Memory Learning (WML) < 69%
Code Substitution-Learning (CDS) accuracy < 56%	Attention (AT) accuracy > 80%	Design Memory Learning (DML) < 50%
Procedural Reaction Time (PRO) accuracy < 56%	Learning (LN) accuracy > 53%	Xs and Os (XO) total incorrect > 30
Mathematical Processing (MTH) accuracy < 56%	Working Memory (WM) accuracy > 53%	Three Letters (TL) total correct < 8
Matching to Sample (M2S) accuracy < 56%	PS speed < AT speed	
Code Substitution-Delayed (CDD) accuracy < 56%	PS speed < WM speed	
Simple Reaction Time 2 (SR2) accuracy < 56%		
Questionable Effort		

Note. ANAM and ImPACT criteria are labeled as invalidity criteria, whereas Axon's criteria are scaled in the direction of validity. ANAM's questionable effort index was derived to differentiate between individuals instructed to put forth suboptimal versus good effort and is described by Roebuck-Spencer et al.[32]

Table 3

Percentage of invalid baseline tests and frequency of failures on each validity criteria

	ANAM *n* = 1,313		Axon *n* = 1,355		ImPACT *n* = 1,434	
% of baselines flagged as invalid	10.7%		11.3%		2.7%	
Criteria failed (% of invalid profiles only)	Consider Retest		Integrity Checks			
	SRT acc < 56%	1.4%	PS acc > 80%	8.5%	ICC > 30	7.7%
	CDS acc < 56%	0.7%	AT acc > 80%	20.9%	WML < 69%	7.7%
	PRO acc < 56%	8.6%	LN acc > 53%	83.0%	DML < 50%	46.2%
	MTH acc < 56%	23.6%	WM acc > 53%	14.4%	XO total incorrect > 30	7.7%
	M2S acc < 56%	35.0%	PS speed < AT speed	9.2%	TL total correct < 8	43.6%
	CDD acc < 56%	38.6%	PS speed < WM speed	4.6%		
	SR2 acc < 56%	0.0%				
	Questionable Effort	12.9%				
Total # of above validity criteria failed	1	89.0%	1	79.1%	1	89.7%
	2	10.7%	2	8.5%	2	7.7%
	3	2.9%	3	3.9%	3	2.6%
	4	1.4%	4	0.7%	4+	0.0%
	5+	0.0%	5	1.3%		
			6	3.9%		

Note. acc = accuracy; SRT = simple reaction time; CDS = code substitution-learning; PRO = procedural reaction time; MTH = mathematical processing; M2S = matching to sample; CDD = code substitution-delayed; SR2 = simple reaction time 2; PS = processing speed; AT = attention; LN = learning; WM = working memory; ICC = impulse control composite; WML = word memory learning; DML = design memory learning; XO = Xs and Os; TL = three letters

Table 4

Results of univariate logistic regression on predictors of CNT validity

	ANAM			Axon			ImPACT		
Categorical variables	**Odds ratio**	**95% *CI***	***p***	**Odds ratio**	**95% *CI***	***p***	**Odds ratio**	**95% *CI***	***p***
Gender (male)	1.18	(.74, 1.89)	.570	1.33	(.88, 1.99)	.373	.84	(.41, 1.70)	.709
Level of competition (HS)	1.16	(.81, 1.67)	.544	.85	(.61, 1.20)	.489	2.12	(1.11, 4.05)	.092
ADHD (yes)	2.25	(1.36, 3.72)	**.005**	3.18	(2.02, 5.02)	**< .001**	1.73	(.60, 5.00)	.461
LD (yes)	3.36	(1.80, 6.25)	**< .001**	1.84	(.91, 3.74)	.226	2.02	(.47, 8.70)	.461
CNT order (2nd)	1.35	(.95, 1.92)	.232	1.91	(1.35, 2.72)	**.001**	1.40	(.73, 2.67)	.461
Taken the CNT before (no)	-	-	-	-	-	-	1.22	(.63, 2.34)	.679
Continuous variables	Estimate	95% *CI*	*p*	Estimate	95% *CI*	*p*	Estimate	95% *CI*	*p*
Age	−.05	(−.15, .04)	.501	−.03	(−.11, .06)	.715	−.17	(−.35, .00)	.153
Number prior concussions	−.05	(−.24, .13)	.631	.04	(−.14, .22)	.768	.17	(−.14, .49)	.461
GPA	−.55	(−.88, −.22)	**.005**	−.83	(−1.14, −.52)	**< .001**	−.96	(−1.58, −.35)	**.035**
WTAR standard score	−.05	(−.06, −.03)	**< .001**	−.04	(−.06, −.03)	**< .001**	−.02	(−.05, .01)	.268

Note. *p*-values were adjusted for multiple comparisons using the false discovery rate method.[4] CNT = computerized neurocognitive tests; HS = high school (versus college); GPA = grade point average; WTAR = Wechsler Test of Adult Reading