Journal of Athletic Training 2017;52(9):826–833
doi: 10.4085/1062-6050-52.6.03

www.natajournals.org

Reliability of Computerized Neurocognitive Tests for Concussion Assessment: A Meta-Analysis

James L. Farnsworth II, MS, ATC*; Lucas Dargo, MS, ATC†; Brian G. Ragan, PhD, ATC‡; Minsoo Kang, PhD, FACSM§

*School of Education and Exercise Science, Buena Vista University, Storm Lake, IA; †Eli Lilly and Company, Indianapolis, IN; ‡Deceased; §Middle Tennessee State University, Murfreesboro

Objective: Although widely used, computerized neurocognitive tests (CNTs) have been criticized because of low reliability and poor sensitivity. A systematic review was published summarizing the reliability of Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) scores; however, this was limited to a single CNT. Expansion of the previous review to include additional CNTs and a meta-analysis is needed. Therefore, our purpose was to analyze reliability data for CNTs using meta-analysis and examine moderating factors that may influence reliability.

Data Sources: A systematic literature search (key terms: *reliability*, *computerized neurocognitive test*, *concussion*) of electronic databases (MEDLINE, PubMed, Google Scholar, and SPORTDiscus) was conducted to identify relevant studies.

Study Selection: Studies were included if they met all of the following criteria: used a test-retest design, involved at least 1 CNT, provided sufficient statistical data to allow for effect-size calculation, and were published in English.

Data Extraction: Two independent reviewers investigated each article to assess inclusion criteria. Eighteen studies involving 2674 participants were retained. Intraclass correlation coefficients were extracted to calculate effect sizes and determine overall reliability. The Fisher Z transformation adjusted for sampling error associated with averaging correlations. Moderator analyses were conducted to evaluate the effects of the length of the test-retest interval, intraclass correlation coefficient model selection, participant demographics, and study design on reliability. Heterogeneity was evaluated using the Cochran Q statistic.

Data Synthesis: The proportion of acceptable outcomes was greatest for the Axon Sports CogState Test (75%) and lowest for the ImPACT (25%). Moderator analyses indicated that the type of intraclass correlation coefficient model used significantly influenced effect-size estimates, accounting for 17% of the variation in reliability.

Conclusions: The Axon Sports CogState Test, which has a higher proportion of acceptable outcomes and shorter test duration relative to other CNTs, may be a reliable option; however, future studies are needed to compare the diagnostic accuracy of these instruments.

Key Words: test-retest design, cognitive function, head injuries, traumatic brain injuries

Many experts agree that, when combined with symptom and motor-control assessments, computerized neurocognitive tests (CNTs) can aid athletic trainers in managing and evaluating patients with sport-related concussions.[1–3] Current estimates suggest that 33% to 39% of athletic trainers include CNTs as part of their return-to-play protocol.[4–6] Although widely used, CNTs have been criticized because of low reliability and poor sensitivity. Reliability coefficients as low as 0.22 have been reported for the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT; ImPACT Applications, Inc, Pittsburgh, PA)[7] and 0.10 for the Automated Neuropsychological Assessment Metrics (ANAM; Vista LifeSciences, Parker, CO).[8]

Reliability is an extremely important concept in concussion testing due to individual serial-testing strategies.[9] *Reliability* refers to the consistency of the scores obtained from a test. When no concussion is present, an athlete's scores should not change in between testing periods. A change in scores when no concussion has occurred indicates measurement error. Unfortunately, no test is perfect, and some level of measurement error is expected with all tests. To address this concern, the *reliable change index* (RCI), a statistic that estimates the magnitude of differences in scores necessary to suggest true change, is often calculated. The size of an RCI depends on the reliability of the test and the desired level of confidence (eg, 80%, 90%, 95%). When reliability is low, the RCI will be large, and when reliability is high, the RCI will be small.

The 90% RCI reported for the visual memory section of the ImPACT ranges from 18.23 to 26.50 points.[10] This indicates that a change in score of at least 18 points is necessary to reflect a true change in visual memory. The intended purpose of the RCI is to minimize the risk of incorrect clinical decisions due to measurement error. Despite the large RCIs reported across each of the ImPACT outcome scores (visual memory, verbal memory, visual-motor speed, and reaction time), 40% to 80% of healthy individuals experienced significant change (ie, a false-positive diagnosis) on at least 2 of 3 trials during serial testing.[10] Although this problem is not limited to ImPACT, it highlights a major area of concern in concussion testing.

False-positive diagnoses are problematic because they can lead to unwarranted removal from competition and

subject patients to unnecessary medical procedures. Many of the commonly used CNTs have relatively high false-positive (ie, a healthy individual is diagnosed with a concussion) rates. Broglio et al[11] identified false-positive rates of 38% on ImPACT and 19% on the Headminder Concussion Resolution Index (Headminder; Headminder Inc, New York, NY). Nelson et al[12] found similar false-positive rates for ImPACT, as well as a false-positive rate of 52% for the Axon Sports CogState Test (Axon; CogState Ltd, Melbourne, Australia).

Although false-positive diagnoses can be a nuisance for athletes, false-negative diagnoses are a bigger concern because they can result in an athlete being returned to play prematurely, leading to further injury or worse. Louey et al[13] reported a false-negative (ie, a concussed athlete being diagnosed as healthy) rate of 17%. When the RCI for a test is high (ie, poor reliability), a high rate of false-negative diagnoses may occur due to the large change in scores needed to identify true change. Cognitive changes after concussion may be subtle, and even though a change in scores might be identified during postinjury testing, the difference in scores between test periods may not exceed the RCI; thus, an incorrect clinical decision would be made, placing the athlete at risk for further harm.

Schatz et al[14] and Ackerman and Kanfer[15] proposed that the low reliability reported in some studies may have been the result of inappropriate study designs resulting in test fatigue experienced by participants. When multiple CNTs are administered concurrently, participants may experience cognitive fatigue, which can negatively affect reliability estimates. Alsalaheen et al[10] contended that the differences in reliability coefficients were more likely due to differences in analytic methods between studies.

Intraclass correlation coefficients (ICCs) are the preferred method of examining reliability between sets of scores. Baumgartner et al[16] suggested that ICCs between 0.70 and 0.79 be considered *below-average acceptable*; 0.80 to 0.89, *average acceptable*; and 0.90 to 1.0, *above-average acceptable*. The ICC, which uses analysis-of-variance techniques to assess variances among sets of scores, includes different models for estimating reliability that depend on the study design. Many of the authors who examined the reliability of CNTs used different ICC models. Inappropriate models can artificially inflate reliability estimates.

In addition, test score reliability appears to depend on the length of time between test administrations.[11,17–19] For example, large differences in reliability coefficients were identified when the test-retest interval was increased from 1 hour to 1 week.[18] The controversy of CNT reliability is further complicated because of the wide range of reliability estimates (from as low as 0.10 to as high as 0.93) reported across studies.[7,8,11,12,17–30]

Because of the conflicting reports, a more thorough investigation of the reliability of CNT scores is necessary. Although a systematic review[10] summarizing the reliability of ImPACT scores has been previously published, this study was limited to a single CNT. Expanding the previous review to include additional CNTs would be beneficial for determining which instrument is the most reliable. Furthermore, no currently published studies have summarized reliability data for CNTs using meta-analytic techniques. Meta-analysis is an advanced statistical procedure used to combine the results from many independent studies into a single study. Meta-analysis can help to minimize biases associated with small sample sizes and allow for comparison of results across different moderator variables (eg, length of the test-retest interval, ICC model selection, participant demographics, study designs). Therefore, our purpose was to compile and analyze current reliability data for CNTs using meta-analytic techniques and to examine moderating factors that may influence the reliability of CNT scores.

METHODS

Literature Search

We conducted a systematic literature search in March 2016 to locate and identify relevant research for the current study. Combinations of the key words *reliability, computerized neurocognitive test,* and *concussion* were entered into the following electronic database search engines with no restrictions for year of publication: MEDLINE, PubMed, Google Scholar, and SPORTDiscus. Literature search findings from each set of key words were recorded and screened for inclusion and exclusion criteria. In addition to electronic database searches, we performed a manual search of reference lists from relevant articles to identify any potential studies missing from the online search. Studies were included in the analysis if they met all of the following inclusion criteria: (1) used a test-retest design; (2) involved at least 1 CNT (ANAM, Axon, Concussion Vital Signs [CNS-VS; CNS Vital Signs, LLC, Morrisville, NC], Headminder Concussion Resolution Index, or ImPACT); (3) author(s) reported sufficient descriptive or inferential statistical data to allow for calculation of effect sizes (ESs); and (4) published in the English language. The outcome of interest was the ICC between the initial baseline test assessment and the follow-up assessment. In this instance, the ICC measures the reproducibility (ie, reliability) between the baseline and follow-up assessments. The Pearson correlation coefficient (r) is another option for reporting reliability; however, r is less desirable because it measures the relative reliability between 2 time points, whereas ICC is a measure of absolute agreement. Therefore, for this meta-analysis, only studies that examined reliability using ICCs were included. Reliability studies using alternative statistics such as the κ statistic or regression were excluded from this review. Using these criteria, potentially relevant studies were screened by 2 independent reviewers, and full texts of all studies meeting the inclusion criteria were further assessed for methodologic quality and data extraction. Unpublished abstracts, dissertations, and theses were considered for inclusion in the study as long as they met the inclusion criteria. When disagreements occurred between the reviewers, consensus was achieved through discussion (see the Figure for a Preferred Reporting Items for Systematic Reviews and Meta-Analyses [PRISMA] diagram illustrating the review process).

Methodologic Quality

Two reviewers independently assessed the methodologic quality of studies using a modified version of the Downs and Black checklist,[31] which has been applied in a recent

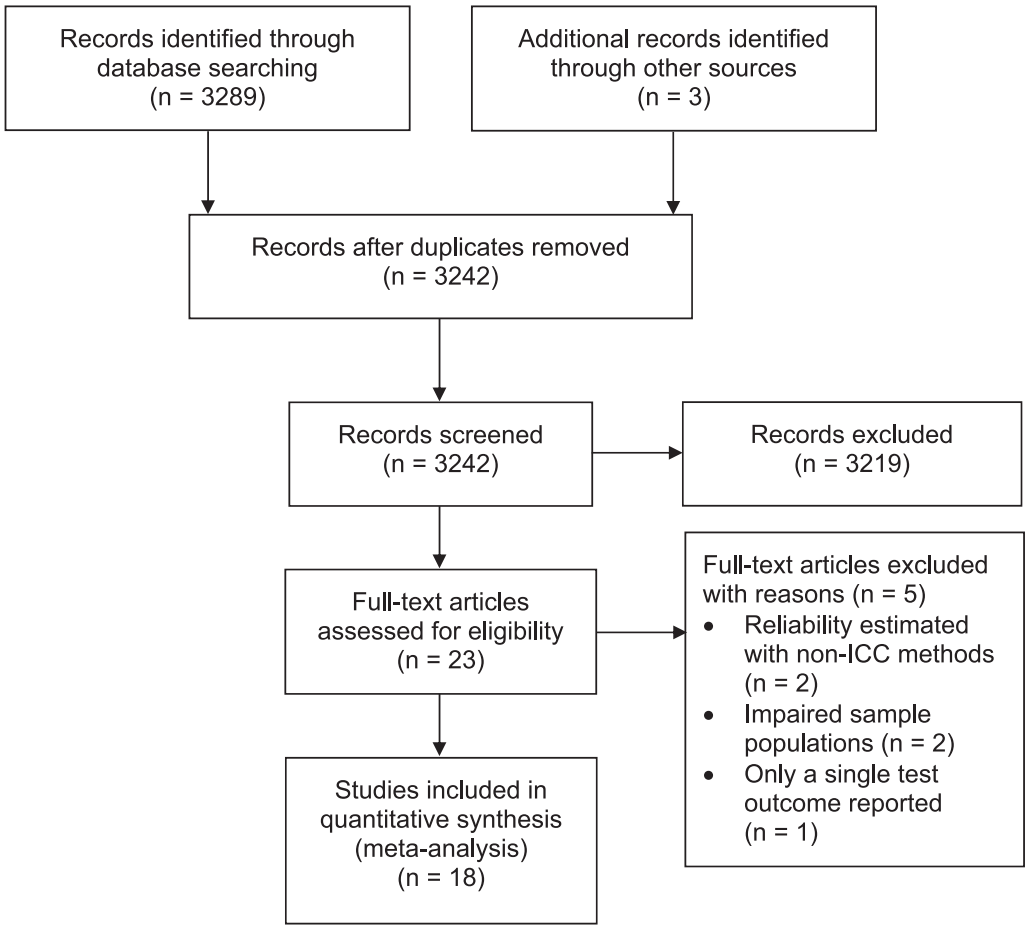

Figure. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow chart. Abbreviation: ICC, intraclass correlation coefficient.

meta-analysis[32] and systematic review.[33] The original checklist contained 27 items to measure the quality of intervention-based studies. Thirteen of the original items were irrelevant to our study design and were removed. The modified checklist consists of 14 items in 3 domains (ie, reporting, external validity, and internal validity) and scores range from 0 to 14 (higher scores indicate better-quality studies). Any study with low methodologic quality (ie, a score greater than 1.5 × interquartile range [IQR] above the upper quartile or a score lower than 1.5 × IQR below the lower quartile) was further examined to determine inclusion in or exclusion from our investigation.

Coding Procedures and Data Extraction

Before coding, we developed a standardized coding form to simplify the extraction process and maintain consistency between the reviewers. Each study was analyzed, and the following data were extracted: the number of participants in each test sample, the type of ICC model used for analyzing test-retest reliability, the average length of the test-retest interval (ie, the average number of days between the first and second testing sessions), the specific CNT(s) used in each study, and the number of CNT(s) administered concurrently in a single session (ie, the number of CNTs each participant completed). Intraclass correlation coefficients were obtained for the reported outcomes of each CNT. When the ICC model was not specified, the author(s) of the study were contacted to determine which model was used. To avoid dependency concerns in studies that reported ICCs for multiple retesting time points, we used only the first time point. If ICCs were reported for multiple subgroups (eg, athletes versus general population, intercollegiate versus high school), each subgroup was assumed to be independent and included in a single meta-analysis.

Key word searches identified 3289 records. Manual searches identified an additional 3 records. After duplicate studies and studies that did not meet the inclusion criteria were removed, 23 studies were available for full review. After reviewing full-text articles for each study, we removed an additional 5 studies from the analysis (see the Figure). This resulted in a final sample of 18 studies. It should be noted that some of these researchers assessed multiple independent samples or administered multiple CNTs to a single sample. Therefore, the final number of samples included in the analysis was 27. Detailed study characteristics are provided in Table 1. Quality of the studies was relatively high, with a median quality index of 13 (upper quartile = 14, lower quartile = 11, IQR = 3). Of the 18 studies, 17 had quality scores within acceptable limits. The remaining study, which had a quality index of 6 (acceptable range = 6.5–14), was an unpublished abstract in which space was limited. After careful consideration and consensus among the authors, we retained the abstract because it contained all the necessary information to

Table 1. Study Characteristics[a]

Study	N	Test(s) Evaluated	Length of Test-Retest Interval	Participant Type	Intraclass Correlation Coefficient Model	Publication Status	Quality Index[c]
Broglio et al[11] (2007)[b]	73	ImPACT, Axon, CRI	45 d	General	(2,1)	P	11
Bruce et al[7] (2014)	305	ImPACT	1 y	Athletes	(2,1)	P	12
Cole et al[21] (2013)[b]	44, 53, 39, 50	ImPACT, Axon, CNS-VS, ANAM	21–42 d	General	(2,1)	P	13
Collie et al[18] (2003)	60	Axon	1 h	General	NR	P	11
Cousino and Kaminski[22] (2006)	14	ANAM	1 d	Athletes	(3,k)	A	13
Elbin et al[17] (2011)	369	ImPACT	0.5–2.35 y	Athletes	(3,k)	P	14
Irwin et al[30] (2014)[b]	92,73	ImPACT	3–4 mo	Athletes	NR	A	6[d]
Louey et al[13] (2014)	235	Axon	1 y	Athletes	(2,k)	P	14
Littleton et al[40] (2015)	40	CNS-VS	6–11 d	General	(2,1)	P	11
MacDonald and Duerson[24] (2015)	117	Axon	50–52 wk	Athletes	(2,1)	P	14
Nakayama et al[25] (2014)	85	ImPACT	45 d	General	(2,1)	P	13
Nelson et al[12] (2016)[b]	166	ImPACT, Axon, ANAM	7 d	Athletes	(3,1)	P	14
Register-Mihalik et al[26] (2012)	40	ImPACT	1–3 d	Athletes	(2,1)	P	13
Resch et al[27] (2013)[b]	46, 45	ImPACT	45 d	General	(1,1)	P	11
Segalowitz et al[20] (2007)	29	ANAM	7 d	General	NR	P	11
Schatz[19] (2010)	95	ImPACT	2 y	Athletes	(3,k)	P	14
Schatz and Ferris[28] (2013)	25	ImPACT	4 wk	General	(3,k)	P	13
Straume-Naesheim et al[29] (2005)	232	Axon	Consecutive	Athletes	NR	P	14

Abbreviations: A, abstract presented at a research conference; ANAM, Automated Neuropsychological Assessment Metrics (Vista LifeSciences, Parker, CO); Axon, Axon Sports CogState Test (CogState Ltd, Melbourne, Australia); CNS-VS, Concussion Vital Signs (CNS Vital Signs, LLC, Morrisville, NC); CRI, Headminder Concussion Resolution Index (Headminder, New York, NY); ImPACT, Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT Applications, Inc, Pittsburgh, PA); NR, not reported; P, published manuscript in peer-reviewed journal.

[a] Median quality index = 13 (upper quartile = 14, lower quartile = 11, interquartile range = 3, acceptable range = 1.5 × interquartile range: 6.5–14).

[b] Study included multiple independent samples.

[c] Modified Downs and Black Checklist[30] scored out of 14 points, where a higher value indicates a higher-quality study.

[d] Low score is attributed to publication status (abstract). After discussion, we retained the article because it met all of the inclusion criteria for the meta-analysis.

calculate ES estimates. Publication bias was determined by performing the Egger test.

Data Analysis

All analyses were performed with R software (version 3.2.4; R Foundation, Vienna, Austria)[34] using the metafor (version 1.9–8)[35] package. Our measure of ES was the ICC, which represents the reproducibility of CNT scores. The Fisher Z transformation was conducted on the ICCs to adjust for sampling error associated with averaging correlations.[36] The transformed average Z coefficients were then transformed back to ICCs to allow for interpretation of the results. A recent Monte Carlo simulation confirmed that the use of back-transformed average Z coefficients are less biased than averaging correlation coefficients.[37] Although the previous authors examined only the Pearson correlation (r), it should be noted that both ICC and r are bounded measures (ie, 0 to 1), whereas the transformed average Z coefficient is an unbounded measure. The Z transformation can also be used to build confidence intervals for the ICCs.[38]

We selected a random-effects model for the current investigation due to the variability among studies. Effect sizes were computed for each outcome on each CNT (eg, 4 ESs were computed for ImPACT: [1] verbal memory, [2] visual memory, [3] visual-motor speed, and [4] reaction time). Effect sizes were estimated using the escalc function, whereas the random-effects model was calculated using the rma function. A restricted maximum-likelihood estimator was used because it has been demonstrated to be unbiased and efficient [model specification: rma(yi, vi, measure = GEN, method = "REML")].[39] A detailed description of the metafor package and the available functions is online at the comprehensive R archive network Web page (https://cran.r-project.org).

To compare the reliability of CNTs, we calculated an average ES by averaging the ESs of the outcomes for each CNT. Furthermore, the proportion of outcomes with acceptable reliability was calculated for each CNT. Only 2 studies[21,40] examined the reliability of the CNS-VS test, whereas only a single study[11] examined the reliability of Headminder. The samples for these CNTs were not included in the overall meta-analysis because of the small sizes. However, these studies were included in moderator analyses.

Due to the high variability in reported ICCs among CNTs, it is important to assess the potential sources of bias in reliability estimates. Understanding these sources of error can help to minimize testing error and improve future studies. Given the small number of studies for some CNTs, it was not possible to evaluate each moderator for each outcome individually. As a result, the ICCs for each outcome were combined into a single analysis. Although combining related outcomes can result in biased estimates of ES,[41] the significance of the moderators can still be determined using these methods with a meta-regression analysis. In this manner, it is possible to determine which variables influence the reliability of outcome scores.

Table 2. Overall Reliability Coefficients for Computerized Neurocognitive Tests

Outcome	k	N	n	Q_{total}	df	*P* Value	I^2 (%)	Effect Size[a]	95% Confidence Interval
ImPACT									
Verbal memory	11	16	1391	48.73	15	<.01	70.53	0.52	0.44, 0.60
Visual memory	11	16	1391	48.73	15	<.01	67.61	0.56	0.48, 0.62
Visual-motor speed	11	16	1391	66.27	15	<.01	76.89	0.77	0.72, 0.81
Reaction time	11	16	1391	46.88	15	<.01	65.41	0.65	0.59, 0.71
Axon/CogSport									
Processing speed	7	7	1022	93.93	6	<.01	94.02	0.73	0.58, 0.83
Attention	7	7	1022	61.38	6	<.01	88.49	0.70	0.56, 0.78
Learning accuracy	7	7	1022	304.18	6	<.01	96.86	0.56	0.25, 0.75
Working memory	7	7	1022	23.90	6	<.01	72.68	0.75	0.69, 0.80
ANAM									
Code substitution—learning	3	3	247	1.10	2	.58	69.28	0.67	0.59, 0.73
Code substitution—delayed	3	3	247	2.84	2	.24	79.64	0.73	0.67, 0.78
Matching to sample	4	4	261	0.99	3	.81	45.21	0.71	0.65, 0.77
Mathematical processing	4	4	261	1.41	3	.70	80.63	0.72	0.65, 0.77
Procedural reaction time	2	2	218	0.99	1	.32	89.00	0.60	0.50, 0.68
Simple reaction time	4	4	261	2.21	3	.53	0.00	0.54	0.44, 0.62
Simple reaction time (repeated)	4	4	261	6.09	3	.11	50.89	0.55	0.38, 0.68

Abbreviations: ANAM, Automated Neurological Assessment Metrics (Vista LifeSciences, Parker, CO); Axon/CogSport, Axon Sports CogState Test (CogState Ltd, Melbourne, Australia); df, degrees of freedom; ImPACT, Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT Applications, Inc, Pittsburgh, PA); k, total number of studies; N, total number of samples obtained from all studies; n, number of participants included across all samples; Q_{total}, Cochran Q heterogeneity statistic (χ^2 distribution with N–1 df); I^2, heterogeneity statistic (100% × [Q–df]/Q).

[a] Effect sizes were calculated for each computerized neurocognitive test using a random-effects meta-analysis model. Effect sizes were calculated from Fisher Z transformed intraclass correlation coefficients. Fisher Z scores were then back-transformed to intraclass correlation coefficients to be interpretable.

We used mixed-effects models with meta-regression procedures to examine the effects of moderator variables on the reliability of CNTs (model specification: rma[yi, vi, mods = ~moderatorvariable measure = GEN, method = "REML"]). Yet due to sample-size limitations, only the effects of the moderators were examined. A separate meta-regression analysis examined each of the following moderator variables: (1) length of the test-retest interval, (2) ICC model selection, (3) participant demographics (eg, athlete population versus general population), and (4) study design (eg, number of CNTs completed by each individual in a single study). A wide range of test-retest intervals was reported in some studies, so the average test-retest interval was used. Heterogeneity was evaluated using the Cochran Q statistics (Q_{model} and Q_{error}), which are based on the χ^2 distribution, with N–1 degrees of freedom, where N represents the total number of samples included in the analysis. In general, a significant Q_{model} suggests that the ES estimates are significantly different across studies. When both Q_{model} and Q_{error} are significant, the moderator variables explain some but not all of the variations in ES estimates across studies. A nonsignificant Q_{model} suggests that there is no difference in ES estimates across studies.

RESULTS

Overall Reliability

Effect-size estimates (ICCs), Q statistics (Q_{total}), and I^2 for CNT outcomes are provided in Table 2. Stem-and-leaf plots illustrating the distribution of ESs for each CNT are shown in Table 3. Evidence of publication bias was examined using the Egger test ($P = .15$); no such bias was identified. Seventy-five percent (3 of 4) of the outcomes for Axon were below-average acceptable (0.70–0.79). Twenty-five percent (1 of 4) of the outcomes for ImPACT were below-average acceptable. Forty-three percent (3 of 7) of the outcomes for ANAM were below-average acceptable. All other outcomes had poor reliability (<0.70).

Moderator Analyses

Intraclass Correlation Coefficient Model Selection. Effect-size estimates for studies using average-measure ICC models (ICC = 0.76; 95% confidence interval [CI] = 0.70, 0.80) were significantly higher than studies using single-measure ICC models (ICC = 0.61; 95% CI = 0.58, 0.65; $Q_{model} = 18.40$, degrees of freedom [df] = 2, $P < .01$; $Q_{error} = 697.82$, df = 112, $P < .01$).

Length of Test-Retest Interval. No differences were identified in ES estimates based on average length of the test-retest interval ($Q_{model} = 0.70$, df = 1, $P = .40$; $Q_{error} = 866.51$, df = 110, $P < .01$).

Study Population. No differences were identified in ES estimates based on the population (athletes versus general population) included in the study ($Q_{model} = 1.31$, df = 1, $P = .25$; $Q_{error} = 919.34$, df = 113, $P < .01$).

Number of Computerized Neurocognitive Tests in the Study Protocol. No differences were identified in ES estimates based on the number of CNTs evaluated in a single testing session ($Q_{model} = 2.19$, df = 1, $P = .14$; $Q_{error} = 903.17$, df = 113, $P < .01$).

Table 3. Stem-and-Leaf Plot of Reliability Coefficients Across Computerized Neurocognitive Tests

Axon		ImPACT		ANAM	
Stems	Leafs	Stems	Leafs	Stems	Leafs
0.1		0.1		0.1	0
0.2	2	0.2	2 3 6 9	0.2	5
0.3	1	0.3	2 8 8 9	0.3	8
0.4	0 3 4 5 9	0.4	0 2 5 5 6 6	0.4	0 4 4 7
0.5	5 6	0.5	0 0 1 2 2 3 3 4 6 7 8 9	0.5	1 2 5 5 8 9
0.6	0 5 5 6 6 7 9	0.6	0 0 0 0 0 2 5 5 7 7 9	0.6	0 0 1 2 6 7 8 9
0.7	1 3 4 6 7 8	0.7	0 1 2 4 4 5 5 6 6 6 6 6 8 9	0.7	0 2 2 3 3 4 9
0.8	1 3 5 6	0.8	1 4 5 6 7	0.8	
0.9	0 3	0.9		0.9	

Abbreviations: ANAM, Automated Neurological Assessment Metrics (Vista LifeSciences, Parker, CO); Axon, Axon Sports CogState Test (CogState Ltd, Melbourne, Australia); ImPACT, Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT Applications, Inc, Pittsburgh, PA).

DISCUSSION

This meta-analysis provides a comprehensive evaluation of the reliability of CNT scores, with combined data from 18 studies, consisting of 27 data samples and 2674 participants. Debate is ongoing among experts regarding the clinical utility of CNTs as part of the clinical decision-making process. Although many studies have been published examining reliability data for the various commercially available CNTs, their large variability can make it difficult to determine which CNT is the most reliable. Athletic trainers often have limited budgets and must choose a single test for use in the clinical setting. The goal of our study was to provide a more in-depth evaluation of CNTs and supply athletic trainers with accurate information for making evidence-based decisions regarding the use of CNTs.

One of the main reasons that direct comparisons of CNTs are difficult is that each test evaluates different domains of cognitive function. This situation is further complicated because some instruments report similar domains, yet these domains are assessed using different tasks. Thus, it can be challenging for athletic trainers to determine which test is the most effective tool. Effect-size estimates across CNT outcomes in this study ranged from 0.52 to 0.77. The majority of the outcomes examined in this meta-analysis (53%) had less than desirable reliability. This is alarming considering the widespread use of these tests in clinical practice. For this reason, the National Athletic Trainers' Association recommends the use of a multidimensional concussion-evaluation protocol.[3] Our results support this recommendation; overreliance on CNTs could result in false-positive and false-negative diagnoses due to low reliability.

It should also be noted that, although reliability is a clear concern for CNTs, such tests are not alone in this regard, particularly in the context of concussion evaluation and management. An examination of the Balance Error Scoring System, a commonly used balance assessment, indicated that the interrater and intrarater reliability ICCs for the total scores were 0.57 and 0.74, respectively.[42] Furthermore, the reliability of scores appeared to be influenced by sex.[43] When multiple baseline assessments were used, the reliability of scores improved.[43] Use of a double baseline for concussion testing may be 1 method of improving the reliability of scores.

Comparisons of the CNTs examined in this study suggest that Axon may be the most reliable. First, the Axon test had the highest proportion of outcomes with acceptable reliability (3 of 4 [75%]). Second, compared with the ImPACT, administration time for Axon is considerably shorter. Axon takes approximately 8 to 10 minutes to complete and contains 4 tasks to assess processing speed, attention, learning accuracy, and working memory. The ImPACT consists of 4 composite scores measured using 6 modules and takes twice as long to complete (approximately 20 minutes). Athletic trainers often work with large groups of athletes across multiple teams. Therefore, baseline testing all athletes can take considerable time. The shortened administration time of Axon would allow more individuals to be tested in the same period.

Learning accuracy was the lone Axon outcome with poor reliability. The learning accuracy task, which is associated with delayed memory, requires participants to recall whether their card has been displayed previously. Axon also requires participants to press a key when their card has turned over, determine if the color of the current card is red, or state whether their card is the same as the most recent card to assess processing speed, attention, and working memory, respectively. By comparison, the learning accuracy task seems significantly more challenging. The increased difficulty of the learning accuracy task could explain the low reliability for this particular outcome, especially if the task is too challenging for the patients being assessed.

Another complication that arises when comparing the efficacy of CNTs is related to study design. To increase power and account for a small sample size, a within-subject study design is often used to compare CNTs. This practice of examining multiple CNTs in a single study[11,14,21] has been questioned by some due to the high risk of fatigue from extended test protocols.[14] Only 1 of the 3 studies counterbalanced to offset these potential biases. We found no differences in ES estimates among studies evaluating a single CNT or multiple CNTs for a single population. These findings are in contrast to those of Schatz et al,[14] who proposed that the low reliability in some studies is related to cognitive fatigue and low methodologic scrutiny. It is likely that the differences in ES estimates identified by the studies in question[11,21] are related to differences in analytic methods rather than to differences in study design.

Many of the studies included in this meta-analysis used different ICC models for analyzing test-retest reliability. In general, ICCs derived from models using average measures will be higher than ICCs derived from single-measure models. Intraclass correlation coefficient model selection should depend on the type of data used for composite scores and the intended use of the instrument. When a double baseline approach is applied to minimize potential learning and practice effects, average-measure ICC models are used to account for the fact that multiple assessments are being incorporated into a single time point. In most cases, however, CNTs are administered only once at each time point in the test-retest design. This is equivalent to assessing the reliability of a single rater, where the single-measure ICC model would be the most appropriate.

A systematic review[10] of the ImPACT reliability studies demonstrated that, when ES estimates are recalculated using average-measure ICC models, coefficients increased by as much as 0.17. Only a single study[13] was published on Axon using average-measure ICC models, and ICCs ranged from 0.83 to 0.93 across the 4 tasks. In this meta-analysis, estimated ESs were different between studies using single-measure and average-measure ICC models. The type of ICC model used accounted for 17% of the variation in ICCs across study outcomes. Inappropriate model selection may result in biased estimates of reliability, which could have contributed to the conflicting evidence reported across studies.

Limitations

Our study was limited by the quantity and quality of the research examining the test-retest reliability of CNTs. The overall sample size was relatively small, which may have influenced the power of the results. In addition, some authors failed to designate which ICC models were used to estimate reliability data. Although most investigators described the ICCs used for their studies through online communication, some were unsure which models were used due to the length of time since the study was published. Additionally, some authors did not respond, resulting in their studies being excluded from moderator analyses due to insufficient information.

Our results suggest that the Axon CNT may be superior to other tests, but it should be noted that ImPACT was included in almost double the number of studies as Axon (11 versus 6). It is possible that Axon's indices would be just as unreliable as those of ImPACT if additional studies were to be completed in the future. Furthermore, some studies were published examining the reliability of Headminder and CNS-VS, yet the number of studies investigating these instruments was too small to allow for comparisons with the more popular CNTs. More work is needed to examine the reliability of these instruments.

Practice effects are another potential area of concern that could influence the reliability of CNT scores. Some studies included multiple retesting time points; however, the number of studies that did this was rather small. In addition, the interval between retesting time points was not consistent. This combined with the small sample sizes would make it challenging to separate practice effects from effects related to the test-retest interval. As a result, we were not able to assess this in the current study. Future research is needed to investigate potential practice effects among CNTs.

Last, we used a univariate meta-analysis approach to analyze the data from each outcome independently. Multivariate data, such as those seen with CNT outcomes, should be analyzed under a multivariate model. To conduct a multivariate meta-analysis, the correlations between outcomes are required to calculate the covariance matrix necessary for analyzing the multilevel data. Unfortunately, no studies published currently, to our knowledge, reported correlations between outcomes, which prohibits the use of a multivariate meta-analysis. In addition, it has been suggested that a large number of studies are needed to produce reliable results with a multivariate meta-analysis.[44] Three potential solutions to this problem are (1) ignoring the dependencies and analyzing the data anyway, (2) averaging the ICC values across studies, or (3) conducting a separate analysis for each independent outcome.[44] Currently, no published investigations, to our knowledge, have examined the correlations between outcomes for each CNT; the effect of ignoring these potential dependencies on estimated ESs is unknown. Therefore, for this study, a combination of methods (2) and (3) was used to calculate ES. First, the ESs were estimated for each outcome independently. Second, the ESs were averaged for each test to determine which test was more reliable.

CONCLUSIONS

Despite limitations, this meta-analysis provides compelling evidence that the reliability of CNTs is less than desirable. Although no significant differences were identified in average ESs across CNTs, the Axon test, which has a higher proportion of acceptable outcomes and shorter test duration relative to other CNTs, may be a reliable option among popular CNTs. Future studies, however, are needed to compare the diagnostic accuracy of these instruments.

REFERENCES

1. Echemendia RJ, Iverson GL, McCrea M, et al. Advances in neuropsychological assessment of sport-related concussion. *Br J Sports Med*. 2013;47(5):294–298.
2. 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*. 2013; 47(5):250–258.
3. Broglio SP, Cantu RC, Gioia GA, et al. National Athletic Trainers' Association position statement: management of sport concussion. *J Athl Train*. 2014;49(2):245–265.
4. Covassin T, Elbin R III, Stiller-Ostrowski JL. Current sport-related concussion teaching and clinical practices of sports medicine professionals. *J Athl Train*. 2009;44(4):400–404.
5. Covassin T, Elbin RJ III, 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.
6. Meehan WP III, d'Hemecourt P, Collins CL, Taylor AM, Comstock RD. Computerized neurocognitive testing for the management of sport-related concussions. *Pediatrics*. 2012;129(1):38–44.
7. Bruce J, Echemendia R, Meeuwisse W, Comper P, Sisco A. 1 year test-retest reliability of ImPACT in professional ice hockey players. *Clin Neuropsychol*. 2014;28(1):14–25.
8. Brunner HI, Klein-Gitelman MS, Zelko F, et al. Validation of the Pediatric Automated Neuropsychological Assessment Metrics in

childhood-onset systemic lupus erythematosus. *Arthritis Care Res (Hoboken)*. 2013;65(3):372–381.
9. Ragan BG, Herrmann SD, Kang M, Mack MG. Psychometric evaluation of the Standardized Assessment of Concussion: evaluation of baseline score validity using item analysis. *Athl Train Sports Health Care*. 2009;1(4):180–187.
10. Alsalaheen B, Stockdale K, Pechumer D, Broglio SP. Measurement in the Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT): systematic review. *J Head Trauma Rehabil*. 2016;31(4): 242–251.
11. 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.
12. 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*. 2016;22(1):24–37.
13. 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*. 2014:29(5):432–441.
14. Schatz P, Kontos A, Elbin R. Response to Mayers and Redick: "Clinical utility of ImPACT assessment for postconcussion return-to-play counseling: psychometric issues." *J Clin Exp Neuropsychol*. 2012;34(4):428–434.
15. Ackerman PL, Kanfer R. Test length and cognitive fatigue: an empirical examination of effects on performance and test-taker reactions. *J Exp Psychol Appl*. 2009;15(2):163–181.
16. Baumgartner TA, Mahar MT, Jackson AS, Rowe DA. *Reliability and Objectivity: Measurement for Evaluation in Kinesiology*. 9th ed. Burlington, MA: Jones & Bartlett; 2015:90–113.
17. Elbin R, Schatz P, Covassin T. One-year test-retest reliability of the online version of ImPACT in high school athletes. *Am J Sports Med*. 2011;39(11):2319–2324.
18. 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*. 2003;13(1):28–32.
19. Schatz P. Long-term test-retest reliability of baseline cognitive assessments using ImPACT. *Am J Sports Med*. 2010;38(1):47–53.
20. Segalowitz SJ, Mahaney P, Santesso DL, MacGregor L, Dywan J, Willer B. Retest reliability in adolescents of a computerized neuropsychological battery used to assess recovery from concussion. *NeuroRehabilitation*. 2007;22(3):243–251.
21. 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.
22. Cousino E, Kaminski T. Test-retest reliability analysis involving five subtests from the Automated Neuropsychological Assessment Metrics [abstract]. *J Athl Train*. 2006;41(suppl 2):S–95.
23. Gualtieri CT, Johnson LG. Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. *Arch Clin Neuropsychol*. 2006;21(7):623–643.
24. MacDonald J, Duerson D. Reliability of a computerized neurocognitive test in baseline concussion testing of high school athletes. *Clin J Sport Med*. 2015;25(4):367–372.
25. Nakayama Y, Covassin T, Schatz P, Nogle S, Kovan J. Examination of the test-retest reliability of a computerized neurocognitive test battery. *Am J Sports Med*. 2014;42(8):2000–2005.
26. Register-Mihalik JK, Kontos DL, Guskiewicz KM, Mihalik JP, Conder R, Shields EW. Age-related differences and reliability on computerized and paper-and-pencil neurocognitive assessment batteries. *J Athl Train*. 2012;47(3):297–305.
27. Resch J, Driscoll A, McCaffrey N, et al. ImPact test-retest reliability: reliably unreliable? *J Athl Train*. 2013;48(4):506–511.
28. Schatz P, Ferris CS. One-month test-retest reliability of the ImPACT test battery. *Arch Clin Neuropsychol*. 2013:28(5):499–504.
29. Straume-Naesheim T, Andersen T, Bahr R. Reproducibility of computer based neuropsychological testing among Norwegian elite football players. *Br J Sports Med*. 2005;39(suppl 1):i64–i69.
30. Irwin CC, Li Y, Bene E, et al. Popular concussion assessments' reliability using high school and collegiate athletes. Presented at: American Alliance for Health, Physical Education, Recreation and Dance national convention and expo; April 1–5, 2014; St Louis, MO.
31. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health*. 1998;52(6):377–384.
32. Kim Y, Park I, Kang M. Convergent validity of the international physical activity questionnaire (IPAQ): meta-analysis. *Public Health Nutr*. 2013;16(3):440–452.
33. Prince SA, Adamo KB, Hamel ME, Hardt J, Gorber SC, Tremblay M. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *Int J Behav Nutr Phys Act*. 2008;5:56.
34. *R: A Language and Environment for Statistical Computing* [computer program]. Version 3.2.4. Vienna, Austria: R Foundation for Statistical Computing; 2016.
35. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Software*. 2010;36(3):1–48.
36. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press; 1985.
37. Silver NC, Dunlap WP. Averaging correlation coefficients: should Fisher's Z transformation be used? *J Appl Psychol*. 1987;72(1):146–148.
38. Carrasco JL, Jover L. Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*. 2003;59(4):849–858.
39. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *J Educ Behav Stat*. 2005; 30(3):261–293.
40. Littleton AC, Register-Mihalik JK, Guskiewicz KM. Test-retest reliability of a computerized concussion test: CNS Vital Signs. *Sports Health*. 2015;7(5):443–447.
41. Becker BJ. Multivariate meta-analysis. In: Tinsley HEA, Brown SD, eds. *Handbook of Applied Multivariate Statistics and Mathematical Modeling*. San Diego, CA: Academic Press; 2000:499–525.
42. Finnoff JT, Peterson VJ, Hollman JH, Smith J. Intrarater and interrater reliability of the Balance Error Scoring System (BESS). *PM R*. 2009;1(1):50–54.
43. Broglio SP, Zhu W, Sopiarz K, Park Y. Generalizability theory analysis of Balance Error Scoring System reliability in healthy young adults. *J Athl Train*. 2009;44(5):497–502.
44. Ahn S, Lu M, Lefevor GT, Fedewa AL, Celimli S. Application of Meta-Analysis in Sport and Exercise Science. In: Ntoumanis N, Myers ND, eds. *An Introduction to Intermediate and Advanced Statistical Analyses for Sport and Exercise Scientists*. Hoboken, NJ: John Wiley & Sons; 2015:233–251.

Address correspondence to James L. Farnsworth II, MS, ATC, School of Education and Exercise Science, Buena Vista University, 610 West 4th Street, Storm Lake, IA 50588. Address e-mail to farnsworth@bvu.edu.