University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

Center for Brain, Biology and Behavior: Papers & Publications

Brain, Biology and Behavior, Center for

2020

Psychometric properties of the standardized assessment of concussion in youth football: Validity, reliability, and demographic factors

Arthur C. Maerlender

Eric Smith

P. Gunnar Brolinson

Jillian Urban

Steven Rowson

See next page for additional authors

Follow this and additional works at: https://digitalcommons.unl.edu/cbbbpapers

Part of the Behavior and Behavior Mechanisms Commons, Nervous System Commons, Other Analytical, Diagnostic and Therapeutic Techniques and Equipment Commons, Other Neuroscience and Neurobiology Commons, Other Psychiatry and Psychology Commons, Rehabilitation and Therapy Commons, and the Sports Sciences Commons

Authors

Arthur C. Maerlender, Eric Smith, P. Gunnar Brolinson, Jillian Urban, Steven Rowson, Amaris Ajamil, Eamon T. Campolettano, Ryan A. Gellner, Srinidhi Bellamkonda, Mireille E. Kelley, Derek Jones, Alex Powers, Jonathan Beckwith, Joseph J. Crisco, Joel Stitzel, Stefan Duma, and Richard M. Greenwald

Psychometric properties of the standardized assessment of concussion in youth football: Validity, reliability, and demographic factors

Arthur Maerlender,[1] Eric Smith,[2] P. Gunnar Brolinson,[2,3]
Jillian Urban,[4] Steven Rowson,[2] Amaris Ajamil,[5]
Eamon T. Campolettano,[2] Ryan A. Gellner,[2]
Srinidhi Bellamkonda,[6] Mireille E. Kelley,[4] Derek Jones,[4]
Alex Powers,[7] Jonathan Beckwith,[5] Joseph Crisco,[6]
Joel Stitzel,[4] Stefan Duma,[2] & Richard M. Greenwald [5]

1 University of Nebraska–Lincoln, Lincoln, Nebraska
2 Virginia Polytechnic Institute and State University, Blacksburg, Virginia
3 Edward Via College of Osteopathic Medicine, Blacksburg, Virginia
4 Wake Forest University, Winston-Salem, North Carolina
5 Simbex, Inc., Lebanon, New Hampshire
6 Brown University, Providence, Rhode Island
7 Wake Forest Baptist Medical Center, Winston-Salem, North Carolina

Corresponding author — Arthur Maerlender, amaerlender2@unl.edu Center for Brain, Biology and Behavior, University of Nebraska, Lincoln, NE, USA 68588

ORCID: https://orcid.org/0000-0002-8786-0905 Joseph Crisco

Abstract

The objective of this study was to determine the psychometrics (reliability, validity) of the original Standardized Assessment of Concussion (SAC) in a youth sample (ages 11 to 13). Demographic factors of race, level of vocabulary knowledge, mother's level of education were also considered. Over 150 youth football athletes completed the SAC and a brief battery of NIH Toolbox cognitive tests as part of a larger study on biomechanical factors in youth sport concussion. This was a within-subjects design (pre-season, post-season assessments), and correlational analysis of convergent and discriminant validity. Between groups analysis based on demographic differences was also employed. The pre-season SAC scores were not different by age; however, SAC scores were statistically different by race: $t(155) = 3.162$, $p = .002$, $d = .519$. Maternal level of education and participant vocabulary scores were related to racial group membership. Convergent and discriminant validity were established compared to NIH Toolbox tests of memory and speed. Pre–post-season tests for 108 participants established marginally acceptable test–retest reliability (ICC = .692). These data support the use of the original SAC in youth football although clinicians must be aware of racial differences in scores.

Keywords: Construct validity, demographic factors, standardized Assessment of Concussion (SAC), test–retest reliability, youth football

Introduction

The Standardized Assessment of Concussion (SAC) was originally developed by McCrea et al. (McCrea, Kelly, Kluge, Ackley, & Randolph, 1997; McCrea et al., 1998, 2001; McCrea, Randolph, & Berger, 2002) following recommendations from the American Academy of Neurology Practice Parameters issued in 1997 (Kelly & Rosenberg, 1997). The purpose was to develop a valid and reliable assessment of concussion that could be used on a sideline of a game field. The SAC has been shown to have a sensitivity of 80–94% and a specificity of 76–91% within 48 hours of injury (Barr & McCrea, 2001; Giza et al., 2013; McCrea, 2001; McCrea et al., 2005, 2003).

Studies of the SAC in younger populations are instructive, but differences in ages, age-groupings, and other methodological considerations make it difficult to draw comparisons to the current study. Across ages from six to 18, mean SAC total scores ranged from 22.7 to 27, with younger students having lower scores than older students, and small but significant differences between males and females (Valovich McLeod, Barr, McCrea, & Guskiewicz, 2006; Valovich McLeod et al., 2004). Intraclass correlations obtained by Valovich McLeod for a sample of nine to 14 year-olds was .46 (Valovich McLeod et al., 2006).

While the previous consensus conferences recommended the use of the SAC and SCAT, the Berlin Consensus of 2016 recommended development of a child-version of the SCAT5, including a new version of the SAC (McCrory et al., 2017). This new test would cover ages five to 12. Brooks, Snedden, Mixis, Hetzel, and McGuine (2017) reported baseline child-SAC (SAC-C) total scores for 191 youth ages 11–13 of 26.1 ($SD = 2.1$) for a mixed-sex sample; no validity or reliability data were presented.

Therefore, the aims of this study were to assess the reliability and validity of the original SAC with middle-school aged football players ages 11–13 years. Reliability was assessed comparing pre-season to post-season assessment results; convergent validity was assessed at pre-season by comparing the SAC total score to a battery of neuropsychological tests for speed and memory. Means and standard deviations of the SAC are provided by race. By analyzing a well-characterized sample, it was intended to determine the appropriateness of the use of the original version in this age group of male football players. If found to be valid and reliable, extending the age of use down to 11 years would be useful for clinical practice.

Materials and methods

This study is part of a larger, ongoing NIH-funded project: "The biomechanical basis of concussions in youth football." Institutional Review Board approval was obtained through Virginia Tech as the lead University.

Participants

Participants for this analysis were male youth, ages 11 to 13 years. This age-band broadly encompasses the middle school years, including fifth graders ($n = 8$), sixth graders ($n = 52$), seventh graders ($n = 64$), and eighth graders ($n = 24$). A grade was not available for all participants. Each year, between 2015 and 2018, there were two teams each from three states (Rhode Island, Virginia, North Carolina), for a total of 24 team-seasons. For this analysis, only athletes in their first season of the project were included to minimize multiple test exposures (pre- and post-season).

Recruitment

Typically, participants were recruited at the initial team meetings or equipment hand-outs at the beginning of each football season by graduate research assistants. For some teams, coaches sent out approved e-mails to players' families; those interested attended an information session of presentations by study-staff at the end of which, informed consent and assent was completed.

Measures

Standardized Assessment of Concussion (SAC)

The SAC is an instrument designed for 6-minute administration to assess four neurocognitive functions — orientation, immediate memory, concentration, and delayed recall — for use by athletic medical care providers on the sidelines of an athletic event (McCrea, 2001; McCrea et al., 1998). Barr and McCrea (2001) reported test–retest reliability coefficient of $r = .55$ for a combined high school and college sample. Change score analysis indicated that a change of more than 3-points was significant. Using that cutoff comparing concussed to non-concussed, the sensitivity was .96 while specificity was .76.

NIH toolbox

Four tests were chosen were used as a brief neuropsychological tests battery to assess memory and speed functions. NIH Toolbox tests have excellent reliability and validity as reported by Tulsky et al. (2017). Before the 2016 season, the Toolbox testing platform was changed from a desk-top computer to an iPad. Differences between test platforms were studied and reported to be no different (National Institutes of Health & Northwestern University, 2017).

The two for memory tests chosen were List Sorting and Rey Auditory Verbal Learning Test. The two processing speed tests were Pattern Comparison and Flanker. The Picture Vocabulary test was used as a proxy for level of cognitive development and not included in outcome analyses except as a covariate (Weintraub et al., 2013). Tests were individually administered by trained research technicians according to standard protocols in laboratories at the respective sites. Testing was conducted at the beginning and end of the season. The tests are described as shown in the following sections.

List sorting

The List Sorting test requires immediate recall and sequencing of different visually and orally presented stimuli. Pictures of different foods and animals are displayed with accompanying audio recording and written text (e.g., "elephant"), and the participant is asked to say the items back in size order from smallest to largest, first within a single dimension (either animals or foods, called 1-List) and then on two dimensions (foods, then animals, called 2-List).

Rey Auditory Verbal Learning Test (RAVLT)

The Rey is a word-list learning task in which 15 unrelated words are presented orally by the examiner or via audio recording over three consecutive learning trials. After each presentation, the participant is asked to recall as many of the words as he/she can.

Pattern comparison

This test measures speed of processing by asking participants to discern whether two side-by-side pictures are the same or not. The items are presented one pair at a time on the computer screen, and the participant is given 90 seconds to respond to as many items as possible.

Flanker

The Flanker task measures both a participant's attention and inhibitory control and is thus a measure of efficiency of processing. The test requires the participant to focus on a given stimulus while inhibiting attention to stimuli flanking it. Sometimes the middle stimulus is pointing in the same direction as the "flankers" (congruent) and sometimes in the opposite direction (incongruent). Twenty trials are conducted.

Picture vocabulary

The NIH Toolbox Picture Vocabulary Test is a measure of receptive vocabulary administered in a computer-adaptive test (CAT) format. The participant is presented with four pictures on the iPad screen and an audio recording saying a word. The participant is instructed to touch the picture that most closely shows the meaning of the word. After the participant makes a choice, another set of pictures automatically appears with the next item and associated audio file.

Analyses

All analyses were carried out using SPSS v.22 (IBM, Chicago, IL). Sample demographics were analyzed for systematic effects on the SAC total score: age, race, developmental-psychiatric diagnoses, medication status, number of previous concussions, maternal levels of education, and years of playing contact sports. ANOVA, *t*-tests, or chi-square tests were used depending on the nature of the data. Those with LD or ADHD were not excluded from the larger sample but rather were examined for differences.

The primary dependent measure was the SAC total score (the sum of the four subtests scores: maximum score of 30). The raw score distributions were explored for outliers (SAC total score $>3\,SD$ from the sample mean).

To explore construct validity, convergent and discriminant validity were assessed with Pearson correlation coefficients calculated with SAC total scores and the four neuropsychological tests; significance was adjusted for multiple comparisons (Bonferroni). Principal component analysis was conducted to verify the factor structure of the five tests (SAC, List Sorting, Pattern Comparison, RAVLT, Flanker).

Test–retest reliability was assessed using a reduced sample that had post-season test results. Intraclass correlation (ICC) was calculated for the SAC using a two-way mixed effects model and consistency definition (Koo & Li, 2016; Shrout & Fleiss, 1979).

The interval between test sessions was used as a covariate to adjust for differences in testing time-periods. To assess if race affected the change scores, two approaches were used. First, repeated measure ANOVA was run (pre-season, post-season SAC by race). Due to differences in preseason scores that would alter regression to the mean, regression-based reliable change scores were computed using Chelune's standard error of difference method using regression parameters from a model with pre-season score predicting post-season score (Chelune, 2003). This approach yields a standard score (z-score) that represents the amount of change after controlling for practice effects and regression to the mean. A z-score of 0 would indicate no real change. This approach is highly recommended for assessing change (Heilbronner et al., 2010). The independent samples *t*-test was used to determine if the amount of change differed by racial group.

The formula for regression-based change scores is:

z-score = T2 − (T1 * beta + constant) / SE of the estimate

where T2 = post-season SAC score, T1 = pre-season SAC score.

Results

By race/ethnicity, the majority of players were either African American or Caucasian (see Table 1). There were 17 participants of different races/ethnicities and since those samples were small, they were eliminated from further analysis. The deleted cases by category were: Asian = 1, Asian-Caucasian = 1, Asian-African American = 1, Hispanic = 6, Hispanic- African American = 6, Native American-Asian-African American = 1, Native American-Caucasian = 1. From the initial sample, one participant was dropped as the SAC values were deemed to be extreme. Demographic characteristics of the final sample are presented in **Table 1.** Note that not all cases provided complete demographic data.

Age

There were no significant differences in SAC total score by age group; therefore, the ages were combined into an 11–13 year-old age-band: $F(2,153) = .036$, $p = .965$. The distribution of individuals was similar across grades for the two primary races (chi-square test of homogeneity): $X^2(3) = 3.522$, $p = .318$.

Race

Independent samples *t*-test of SAC scores by race revealed a significant difference between groups: $t(155) = 3.162$, $p = .002$, $d = .519$. In addition there was a difference in test interval (pre- to post-season) by race: $t(107) = 3.809$, $p < .001$.

Maternal education

A total of 150 cases reported mother's levels of education (see Table 1). There was a weak but significant correlation (Spearman $r_s = .332$, $p < .001$). When binned by categories of education less than Bachelor's

Table 1. Demographic characteristics.

	African American	Caucasian	Total
Age 11	31	27	58
Age 12	28	41	69
Age 13	9	20	29
SAC Total Test 1 mean (*SD*)	24.01 (3.04)	25.42 (2.47)	24.81 (2.81)
SAC Total Test 2 mean (*SD*)	25.55 (2.69)	2 6.31 (2.01	26.00 (2.33)
Picture Vocabulary SS* mean (*SD*)	94.72 (11.53)	107.27 (15.4)	102.00 (15.21)
	(*n* = 47)	(*n* = 65)	(*n* = 112)
ADHD	9	11	20
Language Disability	0	1	1
Reading Disability	3	4	7
Writing Disability	1	1	2
Math Disability	1	2	3
Autism	1	1	2
Taking medication	4	11	15
Average years of contact sport	4.55 (*N* = 64)	3.71 (*N* = 85)	3.97 (*N* = 149)
Previous concussions	1	2	3
Maternal levels of education			
<High school (Low)	0	1	1
High school grad/GED (Low)	19	6	25
Some college (Low)	22	15	38
Bachelor's degree (High)	13	33	49
Master's Degree (High)	6	21	27
PhD (High)	2	7	7
Professional degree (High)	2	2	4
Numbers of High vs. Low*	23/41	63/22	107/63

ADHD = Attention Deficit Hyperactivity Disorder. Test interval is the time from pre-season to post-season assessment.

*There was a significant difference by race: *p* < .001.

degree and above, there was a significant difference between the numbers of African American and Caucasian mothers' at these levels of education: $X^2(1) = 20.890$, *p* < .001.

In addition, maternal education and vocabulary scores were significantly correlated: $r_s = .361$, *p* < .001. Maternal education and vocabulary score were entered into a regression with SAC total score as the dependent variable (step-wise entry). The assumption of normality was determined to be reasonable for standardized and unstandardized residuals (Shapiro-Wilk = .978, *p* = .088. Both maternal education

and vocabulary were significant and accounted for 24% variance (R^2) in SAC total score: $t(103) = 16.213$ $p < .001$).

Exploration of these differences revealed that for lower maternal education level there was no significant difference in participants' vocabulary score; that difference only emerged in the higher maternal education group: $t(84) = -2.811$, $p = .006$). In the hi-maternal education group, the mean African American vocabulary score was 98.49 with standard deviation (*SD*) of 10.48 ($n = 20$), while the Caucasian mean was 110.01 (*SD* = 14.10; $n = 46$).

When SAC scores were compared in the high-maternal education group, the difference between races was significant: equal variances not assumed, $t(48.1) = -3.677$, $p = .001$), but scores were not different by group in the low maternal education group: equal variances not assumed, $t(44.6) = .032$, $p = .975$).

ADHD

Due to the differences in SAC scores by race, standard scores were created for each group. In that way, the groups could be combined for this analysis. There was a significant difference in mean SAC total scores between those with ADHD and those without; equal variances were not assumed: $t(21.50) = -2.224$ $p = .036$; 95th CI of difference: 1.361 to -0.502. As ADHD cases were evenly represented in both race groups, they were left in the sample (X^2: ns).

Vocabulary level

Vocabulary score was predictive of SAC z-score ($R^2 = 0.082$, $t = -3.227$, $p = .002$). Vocabulary scores by race were also significantly different although vocabulary means for both were in the average range ($t(110) = -4.6704$, $p < .001$): African American mean = 94.72, *SD* = 11.53, Caucasian mean = 107.28, *SD* = 15.44.

Test interval

Only 111 cases had testing dates for both assessments. The median days between tests for the African American sample was 97 ($n = 46$) and for the Caucasian sample was 126 ($n = 65$). The difference was significant: Mann-Whitney U = 2141.0, $p < .001$; however, it did not contribute to any analyses.

None of the remaining demographic variables had large enough samples to evaluate or demonstrated significant relationships with SAC total scores: that is, psychiatric or medical conditions, medication usage, number of concussions, or number of years playing contact sports. See **Table 2** for means and standard deviations of SAC scores by ADHD, Race, and Level of Maternal Education.

Construct validity

Although 156 participants had pre-season SAC data, due to missing neuropsychological test data the numbers fluctuated for each validity analysis.

The SAC total score correlated significantly with List Sorting and RAVLT, the two Toolbox memory tests. Pearson correlation showed significant relationships between the two memory scores [List Sort, RAVLT, and SAC total, reflecting convergent validity for memory (List Sort $r = 0.475$, RAVLT $r = 0.442$, both $p < .001$]. Discriminant validity was demonstrated by non-significant relationships between SAC and speed scores (Flanker $r = 0.103$, $p = .202$, Pattern Comparison $r = 0.018$, $p = .848$).

To confirm the relationship of SAC scores to the neuropsychological scores, a principal components analysis was conducted with the

Table 2. SAC total scores by ADHD, race, and level of maternal education.

Group	Characteristic	N	Mean	SD
Full Sample				
	ADHD	20	23.00	3.97
	No ADHD	136	25.08	2.51
AfAm all		64	23.98	3.00
	LoMomEd	41	23.71	3.08
	HiMomEd	23	24.48	2.84
Caucasian all		85	25.40	2.48
	LoMomEd	22	23.68	3.05
	HiMomEd	63	25.40	2.48

ADHD = Attention Deficit Hyperactivity Disorder LoMomEd = maternal education less than bachelor's degree HiMomEd = maternal education level: bachelor's degree or higher

five pre-season test scores. The SAC scores loaded with the two memory tests as expected. **Table 3** presents the rotated factor loadings.

Reliability

One-hundred and eight **(**108) subjects had both pre-season (baseline) and post-season results (69% of total). The mean test interval was 3.5 months, although as noted, there was a significant difference between African American and Caucasian test intervals. Because the difference between intervals did not affect scores the interval was not included as a covariate. The correlation between pre- and post-season standard scores was significant: $r = .534$, $p < .001$. The ICC of SAC pre- and post-season standard scores was also significant: ICC = 0.692, 95% CI = 0.550–0.790.

Change metrics

In terms of raw score changes (pre-season to post season), the African American sample showed an average change of $z = 1.34$ points ($SD = 2.69$) points on retesting, while the Caucasian sample showed a change of $z = 0.969$ points ($SD, = 2.16$). Given the significant differences in pre-season scores, the possibility existed that practice effects and regression to the mean might also be different, thus impacting score changes by race. Regression models for each group were run (pre-season score as the independent and post-season score as the dependent variables), and from those parameters, regression-based

Table 3. Principal component analysis rotated factor matrix, with Varimax rotation.

Variables	1	2
RAVLT	0.723	0.293
List Sort	0.799	0.179
SAC Total	0.851	−0.031
Flanker	0.237	0.795
Pattern Comparison	0.039	0.883

RAVLT = Rey Auditory Verbal Learning Test; SAC Total = Standardized Assessment of Concussion total score.

z-scores were calculated as described above. An Independent Samples t-test of those *z*-scores by group was non-significant: $t(106) = -.043$, $p = .966$. Using 1.28 as the 90% cutoff (1-tail because of the presumption of improvement), five African American scores (11%), and five Caucasian scores (8%) were below the threshold.

Discussion

The SAC is a well-known brief screening of cognitive status that is used following concussion to aid in diagnosis. Although a newer child version has been developed, the analyses presented here sought to characterize the original SAC in a middle-school sample to determine if it could satisfy validity and reliability requirements and thus be appropriately used for this age group. Additionally, there have been few studies of socioeconomic or demographic differences in performance on this measure. The data presented here demonstrated mean scores that were generally consistent with previous studies when interpolated for age (Valovich McLeod, Barr, McCrea, & Guskiewicz, 2006; Valovich McLeod et al., 2004; Valovich McLeod, Bay, Lam, & Chhabra, 2012). Surprisingly, age-bands in youth studies rarely match typical grade breakdowns (i.e., elementary, middle school).

Statistically significant differences were found in SAC total scores between African American and Caucasian students taking the SAC for the first time. Of note, there was not a ceiling effect as only two students (1.3%) obtained perfect scores of 30. The floor value of 15 was also adequate. Both convergent and discriminant construct validity were demonstrated in a comparison with NIH Toolbox cognitive tests of memory (convergent) and speed (discriminant) factors with significant correlations between the SAC total and the neuropsychological memory scores, but not neuropsychological speed scores. Principal component analysis supported these findings.

In the test–retest reliability analysis, practice effects were inferred by significantly higher post-season scores. Test-retest reliability was demonstrated over a 3–4 month retest period. Correlations between tests were significant indicating good stability. There was no effect of test interval on score change.

The findings of racial differences in performance is not unexpected when the effect of maternal education and vocabulary levels are taken
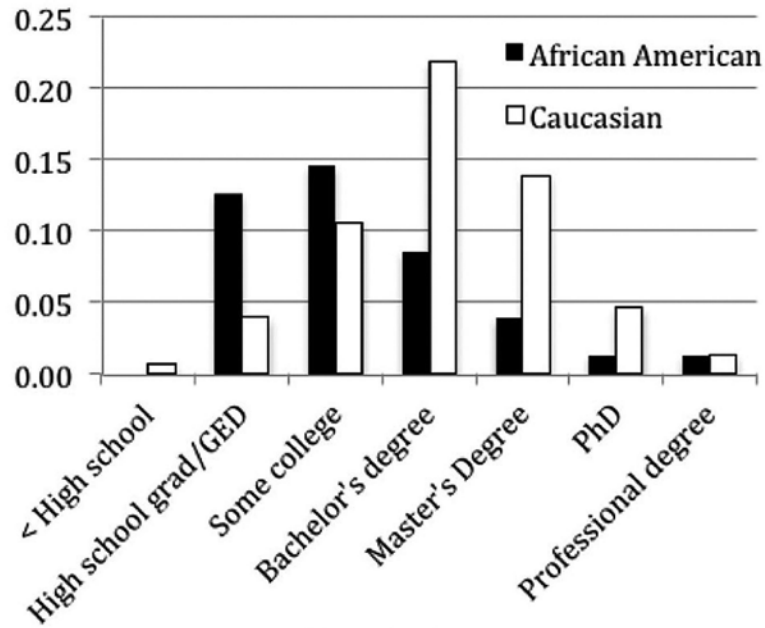
Figure 1. Percentage of cases by race and maternal education.

into account. By percentage, the mothers of Caucasian students in this sample had twice as many Bachelor's and post Bachelor's degrees as the African American group (73% vs. 36%); **Figure 1** characterizes the levels of education by race. The level of vocabulary knowledge was disproportionately influential by levels of maternal education: the difference in vocabulary score by race was only present in the high maternal education group. Thus, the high proportion of less educated mothers in the African American sample is likely a causative factor in lower scores here (Richels, Johnson, Walden, & Conture, 2013).

Manly and Echemendia (2007) outline several caveats in understanding race-based normative standards and assert that the norms used should be defined by the clinical question. Certainly, pointing to lower scores by race as an effect of race alone is a gross misunderstanding of the complexities involved. As is shown here, the story is not so simple.

These findings also demonstrate that the original SAC is psychometrically adequate for use with young males down to age 11, and thus extends the utility of the SAC. It was previously recommended that a child's version of the SAC should be developed because it was thought the "adult" version might be too difficult (McCrory et al.,

2017). However, the difficulty level of the original version may actually provide a better ceiling and thus generate adequate score distributions. Clinicians do not have to be concerned with the need for having two different forms available down through middle school or to adjust for a student's age.

Study limitations include the narrow subject pool: this was an all-male sample of football players. As such, the statistics generated are only appropriate for males. While we have no evidence that there would be a difference between football athletes and other contact or collision sport athletes, clinicians should be aware of the possible differences. The analysis did not include injury sensitivity analysis due to a small number of concussions over this period (10). While identifying differences between two races/ethnicities, it was not possible to compare with Asians, American Indians, Pacific Islanders, or Hispanic students. Those samples sizes were simply too small for analysis. Finally, these were all students from the East coast. It is not expected that youth from other regions would provide different scores, but that has yet to be determined. Clearly, more complete sampling would give a broader perspective.

As a brief cognitive status examination, the SAC relates primarily to memory functions, which are known to be impacted by concussion. More detailed analysis of the cognitive underpinnings of the SAC, as well as its limitations, is needed. The apparent lack of relationship to speed tests in this age-cluster indicates a limitation of the SAC for concussion identification.

For male youth who play contact sports down to age 11, the SAC is an appropriate tool for assessing general cognitive status soon after injury. While there is currently no data available to specify the diagnostic utility of the SAC in this age cluster, using the SAC within the first 48 hours after injury is recommended based on McCrea's studies of high school and college athletes. Baseline assessment with the SAC is also recommended (McCrea et al., 2003, 2002).

Disclosure — No author has any interest to declare related to this work.

References

Barr, W., & McCrea, M. (2001). Sensitivity and specificity of standardized neurocognitive testing immediately following sports concussion. *Journal of the International Neuropsychological Society*, *7*(6), 693–702. doi:10.1017/S1355617701766052

Brooks, M. A., Snedden, T. R., Mixis, B., Hetzel, S., & McGuine, T. A. (2017). Establishing baseline normative values for the child sport concussion assessment tool. *JAMA Pediatrics*, *171*(7), 670. doi:10.1001/jamapediatrics.2017.0592

Chelune, G. (2003). Assessing reliable neuropsychological change. In *Prediction in forensics and neuropsychology: Sound statistical practices* (pp. 123–147). Mahwah, NJ: Lawrence Erlbaum.

Giza, C. C., Kutcher, J. S., Ashwal, S., Barth, J., Getchius, T. S. D., Gioia, G. A., … Zafonte, R. (2013). Summary of evidence-based guideline update: Evaluation and management of concussion in sports. *Neurology*, *80*(24), 2250–2257. doi:10.1212/WNL.0b013e31828d57dd

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., & Hart, R. P. (2010). Official position of the American Academy of clinical neuropsychology on serial neuropsychological assessments: The utility and challenges of repeat test administrations in clinical and forensic contexts. *The Clinical Neuropsychologist*, *24*(8), 1267–1278. doi:10.1080/13854046.2010.526785

Kelly, J. P., & Rosenberg, J. H. (1997). Diagnosis and management of concussion in sports. *Neurology*, *48*(3), 575–580. doi:10.1212/WNL.48.3.575

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. doi:10.1016/j.jcm.2016.02.012

Manly, J., & Echemendia, R. (2007). Race-specific norms: Using the model of hypertension to understand issues of race, culture, and education in neuropsychology. *Archives of Clinical Neuropsychology*, *22*(3), 319–325. doi:10.1016/j.acn.2007.01.006

McCrea, M. (2001). Standardized mental status assessment of sports concussion. *Clinical Journal of Sport Medicine*, *11*(3), 176–181. doi:10.1097/00042752-200107000-00008

McCrea, M., Barr, W. B., Guskiewicz, K., Randolph, C., Marshall, S. W., Cantu, R., … Kelly, J. P. (2005). Standard regression-based methods for measuring recovery after sport-related concussion. *Journal of the International Neuropsychological Society*, *11*(1), 58–69. doi:10.1017/S1355617705050083

McCrea, M., Guskiewicz, K. M., Marshall, S. W., Barr, W., Randolph, C., Cantu, R. C., … Kelly, J. P. (2003). Acute effects and recovery time following concussion in collegiate football players: The NCAA concussion study. *JAMA*, *290*(19), 2556. doi:10.1001/jama.290.19.2556

McCrea, M., Kelly, J. P., Kluge, J., Ackley, B., & Randolph, C. (1997). Standardized assessment of concussion in football players. *Neurology*, *48*(3), 586–588. doi:10.1212/WNL.48.3.586

McCrea, M., Kelly, J. P., Randolph, C., Kluge, J., Bartolic, E., Finn, G., & Baxter, B. (1998). Standardized assessment of concussion (SAC): On-site mental status evaluation of the athlete. *Journal of Head Trauma Rehabilitation*, *13*(2), 27–35.

McCrea, M., Randolph, C., & Berger, L. (2002). Immediate neurocognitive effects of concussion. *Neurosurgery*, *50*(5), 11. doi:10.1227/00006123-200205000-00017

McCrory, P., Meeuwisse, W., Dvořák, J., Aubry, M., Bailes, J., Broglio, S., ... Vos, P. E. (2017). Consensus statement on concussion in sport-the 5th international conference on concussion in sport held in Berlin, October 2016. *British Journal of Sports Medicine*, *51*(11), 838–847. doi:10.1136/bjsports-2017-097699

National Institutes of Health and Northwestern University. (2017). NIH toolbox® for assessment of neurological and behavioral function administrator 's manual table of contents.
 http://www.healthmeasures.net/images/nihtoolbox/Technical_Manuals/ Cognition/Toolbox_Cognition_Battery_Composite_Scores_Technical_Manual.pdf

Richels, C. G., Johnson, K. N., Walden, T. A., & Conture, E. G. (2013). Socioeconomic status, parental education, vocabulary and language skills of children who stutter. *Journal of Communication Disorders*, *46*(4), 361–374. doi:10.1016/j. jcomdis.2013.07.002

Shrout, P., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. doi:10.1037/0033-2909.86.2.420

Tulsky, D., Carlozzi, N., Holdnack, J., Heaton, R. K., Wong, A., Goldsmith, A., ... Heinemann, A. W. (2017). Using the NIH Toolbox Cognition Battery (NIHTB-CB) in individuals with traumatic brain injury. *Rehabilitation Psychology*, *62*(4), 413–424. doi:10.1037/rep0000174

Valovich McLeod, T. C., Barr, W. B., McCrea, M., & Guskiewicz, K. M. (2006). Psychometric and measurement properties of concussion assessment tools in youth sports. *Journal of Athletic Training*, *41*(4), 399–408.

Valovich McLeod, T. C., Bay, R. C., Lam, K. C., & Chhabra, A. (2012). Representative baseline values on the Sport Concussion Assessment Tool 2 (SCAT2) in adolescent athletes vary by gender, grade, and concussion history. *The American Journal of Sports Medicine*, *40*(4), 927–933. doi:10.1177/0363546511431573

Valovich McLeod, T. C., Perrin, D. H., Guskiewicz, K. M., Shultz, S. J., Diamond, R., & Gansneder, B. M. (2004). Serial administration of clinical concussion assessments and learning effects in healthy young athletes. *Clinical Journal of Sport Medicine*, *14*(5), 287–295. doi:10.1097/00042752-200409000-00007

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., ... Gershon, R. C. (2013). Cognition assessment using the NIH Toolbox. *Neurology*, *80*(11), S54–S64.