Journal of the International Neuropsychological Society (2021), **27**, 23–34

doi:10.1017/S1355617720000594

Test–Retest Reliability of Concussion Baseline Assessments in United States Service Academy Cadets: A Report from the National Collegiate Athletic Association (NCAA)–Department of Defense (DoD) CARE Consortium

Megan N. Houston[1,*], Kathryn L. Van Pelt[2], Christopher D'Lauro[3], Rachel M. Brodeur[4], Darren E. Campbell[5], Gerald T. McGinty[3], Jonathan C. Jackson[3], Tim F. Kelly[6], Karen Y. Peck[6], Steven J. Svoboda[7], Thomas W. McAllister[8], Michael A. McCrea[9], Steven P. Broglio[10] and Kenneth L. Cameron[1]

[1]John A. Feagin Jr. Sports Medicine Fellowship, Keller Army Hospital, West Point, NY 10996, USA
[2]University of Kentucky, Lexington, KY 40526, USA
[3]United States Air Force Academy, Colorado Springs, CO 80840, USA
[4]United States Coast Guard Academy, New London, CT 06320, USA
[5]Logan Regional Orthopedics, Logan, UT 84341, USA
[6]United States Military Academy, West Point, NY 10996, USA
[7]MedStar Orthopaedic Institute, Washington, DC 20036, USA
[8]Department of Psychiatry, Indiana University, Indianapolis, IN 46202, USA
[9]Department of Neurosurgery, Medical College of Wisconsin, Milwaukee, WI 53226, USA
[10]Michigan Concussion Center, University of Michigan, Ann Arbor, MI 48109, USA

(Received December 9, 2019; Final revision April 14, 2020; Accepted May 12, 2020; First Published Online June 16, 2020)

Abstract

Objective: In response to advancing clinical practice guidelines regarding concussion management, service members, like athletes, complete a baseline assessment prior to participating in high-risk activities. While several studies have established test stability in athletes, no investigation to date has examined the stability of baseline assessment scores in military cadets. The objective of this study was to assess the test–retest reliability of a baseline concussion test battery in cadets at U.S. Service Academies. **Methods:** All cadets participating in the Concussion Assessment, Research, and Education (CARE) Consortium investigation completed a standard baseline battery that included memory, balance, symptom, and neurocognitive assessments. Annual baseline testing was completed during the first 3 years of the study. A two-way mixed-model analysis of variance (intraclass correlation coefficent $(ICC)_{3,1}$) and Kappa statistics were used to assess the stability of the metrics at 1-year and 2-year time intervals. **Results:** ICC values for the 1-year test interval ranged from 0.28 to 0.67 and from 0.15 to 0.57 for the 2-year interval. Kappa values ranged from 0.16 to 0.21 for the 1-year interval and from 0.29 to 0.31 for the 2-year test interval. Across all measures, the observed effects were small, ranging from 0.01 to 0.44. **Conclusions:** This investigation noted less than optimal reliability for the most common concussion baseline assessments. While none of the assessments met or exceeded the accepted clinical threshold, the effect sizes were relatively small suggesting an overlap in performance from year-to-year. As such, baseline assessments beyond the initial evaluation in cadets are not essential but could aid concussion diagnosis.

Keywords: Mild traumatic brain injury, Neuropsychological tests, Balance, Symptoms, Neurocognitive, Change score

INTRODUCTION

A signature wound of the Iraq and Afghanistan Wars (Snell & Halter, 2010) and declared a public health issue per the Centers for Disease Control and Prevention (Centers for Disease Control and Prevention, 2016), traumatic brain injuries (TBIs) can result in acute and long-term cognitive, behavioral, and physical effects. Since 2000, an estimated 383,947 TBIs have occurred in United States (U.S.) military service members, 82% of which were classified as mild TBIs or more commonly termed concussions (Defense and Veterans Brain Injury Center, 2018). Surprisingly, the majority of

*Correspondence and reprint requests to: Megan N. Houston, John A. Feagin Jr. Sports Medicine Fellowship, Keller Army Hospital, 900 Washington Road, West Point, NY 10996, USA. Tel: +1 845 938 6826. E-mail: megan.n.houston.ctr@mail.mil

these injuries are not related to injuries during military deployments (Cameron et al., 2012). With roughly 11,000–33,000 concussions per year in the U.S. military (Defense and Veterans Brain Injury Center, 2018) and 1.6–3.8 million sports and recreation-related concussions within the U.S. civilian population (Langlois et al., 2006), multiple organizations have suggested or endorsed baseline assessments for athletes (Broglio et al., 2014; Herring et al., 2011; McCrory et al., 2017) and service members (Department of Defense, 2015) prior to athletic participation and deployment. Intended to provide a pre-morbid standard to allow a better measure of impairment following injury, baseline assessments remain difficult to interpret and incorporate into the evaluation and management of concussion. Establishing clinical interpretation ranges and understanding the foundational psychometric properties of these baseline assessments are vital to the clinical management of concussion.

The initial baseline and post-concussion test batteries emerged as part of the Sports as a Laboratory Assessment Model (SLAM) in the late 1980s (Barth et al., 1989). The SLAM methodology was founded on the use of a pre–post neurocognitive test model to measure impairment post-concussion (Barth et al., 1989). Current concussion baseline batteries have expanded the pre–posttest model to include neurological, postural, and symptom assessments, in addition to the neurocognitive test. Thus, with repeat exposure to these assessments, pre-, and post-injury, establishing reliability metrics is critical to interpreting scores and differentiating between normal variation within a test and variation due to injury. For repeat exposure, test–retest reliability is an important metric and typically measured as an intraclass correlation coefficient (ICC) or Kappa coefficient. ICC and Kappa scores ≥0.75 are considered good or clinically acceptable (Portney & Watkins, 2009). Scores <0.75 reflect moderate (0.50–0.74) to poor (<0.50) reliability and do not meet the accepted threshold for clinical utility (Portney & Watkins, 2009). Unfortunately, the results of prior reliability analyses of baseline assessments have varied and involved relatively small homogenous cohorts (Bell et al. 2011; Broglio et al., 2018; Farnsworth et al., 2017; McLeod & Leach, 2012).

The clinical tools and individual assessments included in a baseline test battery for concussion may vary; however, most are multidimensional and include a computerized neurocognitive test, as well as, balance, memory, and symptom assessments. The reliability of computerized neurocognitive tests has been less than desirable (Farnsworth et al., 2017). In a recent meta-analysis (Farnsworth et al., 2017), reliability coefficients for the Immediate Post-Concussion Assessment and Cognitive Test (ImPACT) and Automated Neuropsychological Assessment Metric (ANAM) ranged from 0.10 to 0.87 across 13 studies. While the Concussion Assessment, Research and Education (CARE) Consortium reported more consistent test–retest reliability values for the ImPACT across 1- (ICCs = 0.50–0.72) and 2- (ICCs = 0.34–0.66) year time periods in their cohort of roughly 3000 athletes, these values are still less than optimal (Broglio et al., 2018). Other components of baseline concussion protocols have displayed similar variability across time. In the remaining CARE Consortium assessments (Broglio et al., 2017) that include the Standardized Assessment of Concussion (SAC), Balance Error Scoring System (BESS), Sport Concussion Assessment Tool (SCAT)-Symptom Evaluation, and Brief Symptom Inventory-18 (BSI-18), test–retest reliability was not ideal (Broglio et al., 2018). ICC values ranged from 0.34 to 0.51 and kappa statistics from 0.40 to 0.41 (Broglio et al., 2018). With a shorter test interval (e.g., 50–60 days), BESS reliability in children (ICC = 0.70) (Valovich McLeod et al., 2006) and young adults ($G = 0.64$) (Broglio et al., 2009) has been moderate. Lesser values have been published for the SAC (ICC = 0.46) (Valovich McLeod et al., 2006), BSI-18 (ICC = 0.37–0.69) (Lancaster et al., 2016), and symptom components of the SCAT-2 (ICC ≤0.50) (Chan et al., 2013), an earlier version of the SCAT, when administered at various time intervals (e.g., 7, 30, 45, 60, and 165 days). Thus, the time between administrations can influence reliability metrics consistent with diffusion drift models of cognition (Ratcliff et al., 2016). In addition to the time interval between test administrations, other reliability confounders include sex, age, and testing environment (Broglio et al., 2009; Lichtenstein et al., 2014). BESS test–retest reliability improved when male ($G = 0.92$) and female ($G = 0.91$) participants were analyzed independently, indicating that sex accounted for the largest source of variance in BESS scores (Broglio et al., 2009). Furthermore, younger athletes (10–12 years) and larger test groups (≥20 per room) have increased the frequency of invalid ImPACT test scores (Lichtenstein et al., 2014). Given the contradictory findings in published reliability values and the various confounders known to effect baseline performance, further investigation into the stability of these tests is necessary, particularly in high-risk populations that have been under-represented in the literature such as service academy enrollees.

While test stability is critical to the serial administration of concussion assessments in the test–retest paradigm, much of the focus has shifted toward the clinical interpretation of scores. Thus, a variety of methods have been proposed to distinguish between normal test variation and clinically meaningful change. In the realm of concussion, reliable change indices have been used (Barr & McCrea, 2001; Hinton-Bayre et al., 1999; Iverson, 2011). Reliable change indices have broad application to concussion because it identifies clinical change independent of measurement error (Iverson, 2011). By creating a confidence interval around the index, clinicians can estimate the measurement error on retest scores. The technique has been applied to a variety of concussion assessment tools to establish criteria that indicate significant change on neurocognitive (Barr & McCrea, 2001; Hinton-Bayre et al., 1999) and postural control assessments (Broglio et al., 2008; Valovich McLeod et al., 2006).

Administering a baseline test battery pre-injury enables the medical professionals treating concussed patients to not only measure impairment but to apply individualized performance metrics when making return-to-play or return-to-duty decisions. While some work evaluating the test–retest

https://doi.org/10.1017/S1355617720000594 Published online by Cambridge University Press

reliability of common concussion assessment tools has been completed in traditional athletic populations, to date, no investigation has examined the reliability of these test batteries among military service members or service academy cadets. Thus, the objective of the current study is to describe the test–retest reliabilities of the ImPACT, BESS, SAC, SCAT-Symptom Evaluation, and BSI-18 among service academy cadets. Furthermore, this study aims to establish clinical interpretation ranges for this particular baseline battery in the military service academy population.

METHODS

As part of the CARE Consortium, the U.S. Service Academies took part in a multi-site study investigating the natural history of concussion. All service academy cadets from the United States Military Academy (West Point), Air Force Academy, and Coast Guard Academy were eligible and invited to participate. Prior to data collection, each institution's local Institutional Review Board (IRB) and the U.S. Army Human Research Protection Office (HRPO) approved the study protocol and all participants provided written informed consent.

The CARE study initially launched in 2014, thus the complete methodology has been described in an earlier article (Broglio et al., 2017). From 2014 to 2016 or years 1 (Y1), 2 (Y2), and 3 (Y3) of the study, participants completed an annual baseline assessment. In summary, each athlete or cadet enrolled completed a detailed demographic questionnaire that included medical history, as well as a comprehensive baseline assessment. Each Service Academy addresses neurocognitive function, neurological status, motor control, and symptom domains using the same CARE Level A assessments. Their primary test battery includes the SAC, BESS, BSI-18, SCAT – Symptom Evaluation (Version 3), and the ImPACT. The outcomes associated with each test are briefly described below.

- The *SAC* is an acute measure of cognitive function comprised of four components: orientation, immediate memory, concentration, and delayed recall (McCrea et al., 1998). Total SAC score out of 30 was used as the outcome variable for this study.
- The BESS is a measure of posture stability that consists of three stances (double limb, single limb, tandem) performed on two surfaces (firm and foam) for a total of six balance trials (Riemann et al., 1999). Total BESS score out of 60 was used as the outcome variable for this study.
- The *BSI-18* captures symptoms related to anxiety, mood, and depression to measure psychological distress (Meachen et al., 2008). BSI-Total score out of 72 was used as the outcome variable for this study.
- The *SCAT-Symptom Evaluation* captures physical, cognitive, sleep, and affective symptoms related to concussion and is summarized as the total number of symptoms and symptom severity (McCrory et al., 2013). SCAT Total number of symptoms out of 22 and SCAT symptom severity out of 132 were used as outcome variables for this study.
- The *ImPACT* is a neurocognitive computer assessment (Iverson et al., 2003). *ImPACT* composite scores for verbal memory, visual memory, motor speed, and reaction time were include as outcome variables.

The current project aims to describe the test–retest reliabilities for the CARE Consortium's Level A baseline test battery among cadets enrolled at the U.S. Service Academies. Thus, in June of 2017, all annual CARE Level A Service Academy baseline assessments were pulled from the repository. Between January 2014 and March 2017, the Service Academies participating in the CARE Consortium captured 16,061 baseline assessments. Participants were included in the current analyses if they had at least two baseline assessments recorded in the database. Participants with only one baseline record or those that sustained a concussion between annual baselines were excluded.

Statistical Analyses

Prior to any analysis, the data were cleaned. ImPACT scores were removed if the ImPACT system deemed the test session invalid. Overall, 41 cadets had ImPACT scores flagged as invalid. For these cadets, the invalid results were removed from subsequent analyses, but the other baseline assessments, if intact, were analyzed. Additionally, BSI-18 scores that were non-integer values were also excluded as scoring only allows for integers and therefore these were deemed data entry errors. All other assessments (i.e., SAC, BESS, SCAT-Symptoms) were checked by verifying scores fell within the possible test score limits. For example, the BESS total score was checked by determining all scores fell between 0 and 60. No tests were excluded for the SAC, BESS, or SCAT-Symptoms. All statistical analyses were conducted in R Version 3.6.1 Statistical Software Package (Vienna, Austria). Distribution metrics were calculated as mean, median, and quartiles. Test–retest reliability was calculated between Y1 and Y2 and Y1 and Y3. A two-way mixed-model analysis of variance ($ICC_{3,1}$) (Shrout & Fleiss, 1979) was used for the SAC, BESS, BSI-18, and ImPACT scores to assess stability in these measures over time. Both "consistency" (ICCc) and "agreement" (ICCa) definitions were estimated for ICC. Both methods were estimated to provide a description of how well tests were rated in a consistent manner (e.g., high scores in both years) versus absolute agreement (e.g., getting the exact same score both years). The "psych" package was used to calculate both ICCs. Since many participants select zero symptoms at baseline, we considered the SCAT-Symptom Evaluation to be a categorical assessment. Thus, Cohen's Kappa using linear weights was used to calculate test–retest reliability for the SCAT-Symptom Evaluation scores. The "rel" package was used to calculate weighted Kappa statistics. Both ICC and Kappa values are scored on a 0.0 to 1.0 scale with greater scores indicating more stable performance (Koo & Li, 2016). It is important to note that ICC calculations do not require the assumption of normality and therefore are appropriate estimates of

https://doi.org/10.1017/S1355617720000594 Published online by Cambridge University Press

reliability for SAC, BESS, BSI-18, and ImPACT scores (Mehta et al., 2018).

The initial analyses included all cadets. However, secondary analyses were stratified by level of sports participation (varsity and non-varsity athletes). A secondary sub-analysis stratified cadets into freshmen and upperclassmen to determine whether baseline assessments completed during the transition from high school or prior military service to a Service Academy environment influenced score reliability and variability. ICC and Kappa values were not computed for analyses in which the sample size was less than 100.

Bland–Altman plots were also generated to visualize the agreement between tests at both time points. The Bland–Altman plot is a scatter plot showing the difference in assessment scores (e.g., Y2–Y1 and Y3–Y1) on the Y-axis and the mean of both assessments on the X-axis. The mean of the difference in assessment scores provides an indication of the level of bias. A positive mean bias (>0) indicates that Y2 or Y3 scores are greater than Y1, and a negative mean bias (<0) indicates that Y2 and Y3 scores are less than Y1. The level of agreement is defined by 95% confidence intervals around the mean difference. To estimate the level of agreement bounds, the Bland–Altman analysis assumes homoscedasticity. The assumption of homoscedasticity was evaluated using the Goldfeld–Quandt test (Hoffman, 2015) with the "lmtest" package in R. If the assumption of homoscedasticity was violated, then the level of agreement bounds was estimated to enable the bounds to increase/decrease with increasing mean of both assessments (Grilo & Grilo, 2012).

Cohen's *d* effect sizes were calculated to evaluate the magnitude of change over time between annual baseline assessments. Effect size estimates of <0.2, 0.5, and 0.8 were deemed small, medium, and large, respectively (Cohen, 1977). For clinical interpretation, rather than estimating the percentiles of the distribution under the assumption of normality (i.e., reliable change indices), non-parametric confidence intervals based on the observed distributions were applied to estimate the degree of certainty of change on each assessment or change scores. More specifically, these are one-sided confidence intervals used to represent change that may occur as a result of concussion. Following a suspected concussion, we would anticipate that poorer performance would be reflected by lesser scores on the SAC and verbal memory, visual memory, and visual motor speed sections of *ImPACT*, but greater scores on BESS, BSI-18, SCAT – Symptom Evaluation outcomes, and *ImPACT* reaction time.

RESULTS

Reliability Analyses

At the time of analysis, 4875 cadets (76.9% male) had completed the annual CARE baseline Level A test battery during back-to-back years without sustaining a concussion during the follow-up period and 207 cadets had additionally completed the baseline battery during Y1 and Y3. Of the 4875 cadets that completed baselines in Y1 and Y2, 28.45% ($n = 1387$) participated in varsity sports, 44.91% ($n = 2177$) were freshmen, and 18.79% ($n = 908$) reported a prior concussion. All of the cadets who completed baselines in Y1 and Y3 participated in varsity athletics, 86.47% were male ($n = 179$), 34.30% ($n = 71$) were freshmen, and 33.33% ($n = 69$) reported a prior concussion.

Distribution metrics and reliability analyses results for the clinical concussion assessments are reported in Table 1. The metrics and ICC values for the ImPACT appear in Table 2. ICCc values (showing score consistency) from Y1 to Y2 ranged from 0.28 to 0.67 and Y1 to Y3 from 0.17 to 0.57. ICCa values (showing score agreement) were similar ranging from 0.28 to 0.67 in Y1 to Y2 and 0.15 to 0.57 in Y1 to Y3. Kappa values for the SCAT Symptom Evaluations ranged from 0.16 to 0.21 from Y1 to Y2 and 0.29 to 0.31 from Y1 to Y3. Overall, the reliability analyses indicated poor consistency. ImPACT visual memory and visual motor speed were the only assessments greater than 0.50.

Bland–Altman analyses revealed statistically significant, but clinically insignificant bias in all clinical assessments for both time points except for the BESS in Y1 and Y3 (Table 1). Cadets performed better on the SAC in Y2 and Y3 compared to Y1. BESS performance was worse in Y2 compared to Y1. Fewer symptoms were reported on the BSI-18 and SCAT Symptom Evaluations in Y2 and Y3 compared to Y1. For ImPACT, only verbal memory and visual memory performance demonstrated significant positive bias (Table 2). Better performance was observed for verbal memory scores in Y2 and Y3 compared to Y1 and visual memory performance improved in Y2 compared to Y1. The Bland Altman plots for level of agreement between Year 1 and Year 2 are presented in Figures 1 and 2. Figures 3 and 4 display Bland–Altman plots of a subset of clinical assessments where the assumption of homoscedasticity was not met.

Distribution and reliability analyses were conducted separately for varsity (Tables S1 and S2) and non-varsity cadet-athletes (Tables S4 and S5). The sub-analyses by class year, freshman versus upperclassmen, are presented in Supplementary Tables S7–S8 and S10–S11. Due to an inadequate sample size, the 2-year test interval could not be calculated for the non-varsity athletes or upperclassmen. Overall, the separate sub-analyses yielded similar ICC and Kappa values to the combined sample. However, SCAT Symptom Number and SCAT Symptom Severity decreased between Y1 and Y2 for freshmen (Kappa = 0.10–0.13) compared to upperclassmen (Kappa = 0.27–0.30). Overall, none of the reliability metrics neared 0.75 to suggest good stability from year-to-year for annual baseline concussion assessments.

Clinical Interpretation Ranges

Cohen's *d* effect sizes are reported in Tables 1 and 2. Across all measures, the observed effects were small. From Y1 to Y2, effects ranged from 0.04 to 0.38 and from Y1 to Y3 from 0.01 to 0.44. The smallest effects (0.01–0.04) were observed for the BESS. The largest effects, still interpreted as small

https://doi.org/10.1017/S1355617720000594 Published online by Cambridge University Press

Table 1. Measures of central tendency, reliability, and effect sizes for clinical concussion measures for all cadets

	n	Mean (SD)	Median (Q1–Q3)	ICCc (lower, upper)	ICCa (lower, upper)	Bias (lower, upper)	Cohen's *d*
SAC							
Year 1	4536	27.66 (1.81)	28 (27–29)	0.34 (0.31, 0.36)	0.32 (0.27, 0.37)	0.53* (−3.31, 4.38) λ	0.31
Year 2	4536	28.19 (1.61)	28 (27–29)				
Year 1	205	27.64 (1.86)	28 (27–29)	0.17 (0.03, 0.3)	0.15 (0.02, 0.28)	0.69* (−3.56, 4.94)	0.41
Year 3	205	28.33 (1.47)	28 (28–30)				
BESS							
Year 1	4476	13.6 (6.38)	13 (9–17)	0.28 (0.25, 0.31)	0.28 (0.25, 0.3)	0.29* (−15.16, 15.75)	0.04
Year 2	4476	13.89 (6.74)	13 (9–18)				
Year 1	201	14.83 (7.11)	13 (10–18)	0.24 (0.11, 0.37)	0.25 (0.11, 0.37)	−0.08 (−16.71, 16.55)	−0.01
Year 3	201	14.75 (6.69)	14 (10–18)				
BSI-18 total							
Year 1	4797	41.88 (8.51)	36 (36–47)	0.34 (0.32, 0.37)	0.32 (0.25, 0.38)	−2.80* (−19.31, 13.71)	−0.38
Year 2	4797	39.08 (5.94)	36 (36–42)				
Year 1	206	40.82 (7)	36 (36–45)	0.27 (0.14, 0.4)	0.25 (0.11, 0.38)	−2.66* (−16.91, 11.60) λ	−0.44
Year 3	206	38.16 (4.9)	36 (36–39)				
SCAT symptom number					Weighted Kappa (lower, upper)		
Year 1	4815	3.28 (4.34)	2 (0–5)	–	0.21 (0.19, 0.23)	−1.36* (−9.96, 7.24) λ	−0.37
Year 2	4815	1.92 (2.92)	1 (0–3)				
Year 1	207	2.28 (3.02)	1 (0–3)	–	0.29 (0.19, 0.40)	−0.86* (−6.28, 4.57) λ	−0.32
Year 3	207	1.42 (2.21)	0 (0–2)				
SCAT symptom severity							
Year 1	4815	6.16 (10.26)	2 (0–7)	–	0.16 (0.14, 0.18)	−3.12* (−23.62, 17.38) λ	−0.38
Year 2	4815	3.04 (5.57)	1 (0–4)				
Year 1	207	3.57 (5.54)	2 (0–4)	–	0.31 (0.20, 0.42)	−1.29* (−11.51, 8.92) λ	−0.26
Year 3	207	2.27 (4.34)	0 (0–3)				

BESS, Balance Error Scoring System; BSI-18, Brief Symptom Inventory-18; SAC, Standardized Assessment of Concussion; SCAT, Sport Concussion Assessment Tool; ICCa, Intraclass Correlation Coefficient-Agreement; ICCc, Intraclass Correlation Coefficient-Consistency. λ Levels of agreement change with size of mean. *Bias confidence interval does not cover 0 mean difference.

Table 2. Measures of central tendency, reliability, and effect sizes for ImPACT for all cadets

	n	Mean (SD)	Median (Q1–Q3)	ICCc (lower, upper)	ICCa (lower, upper)	Bias (lower, upper)	Cohen's *d*
Verbal memory							
Year 1	4220	90.05 (9.24)	92 (84−98)	0.42 (0.39, 0.44)	0.42 (0.39, 0.44)	0.75* (−18.68, 20.18) λ	0.08
Year 2	4220	90.8 (9.1)	93 (85−99)				
Year 1	202	89.37 (10.17)	92 (84−98)	0.40 (0.28, 0.51)	0.40 (0.28, 0.54)	2.55* (−17.52, 22.62) λ	0.27
Year 3	202	91.92 (8.51)	95 (88−99)				
Visual memory							
Year 1	4220	82.22 (11.89)	84 (75−91)	0.52 (0.50, 0.54)	0.52 (0.50, 0.68)	1.29* (−21.37, 23.96)	0.11
Year 2	4220	83.51 (11.68)	85 (77−92)				
Year 1	202	81.77 (13.09)	83.5 (75−92)	0.44 (0.33, 0.55)	0.44 (0.33, 0.48)	3.26 (−22.36, 28.88) λ	0.26
Year 3	202	85.03 (11.64)	87 (78−94)				
Visual motor speed							
Year 1	4219	43.06 (5.95)	43.58 (38.88−47.7)	0.67 (0.65, 0.68)	0.67 (0.65, 0.44)	0.89 (−8.54, 10.32)	0.15
Year 2	4219	43.96 (5.84)	44.67 (40.22−48.65)				
Year 1	202	42.46 (6.36)	43.03 (38.3−47.47)	0.57 (0.46, 0.65)	0.57 (0.46, 0.54)	2.64 (−8.25, 13.52)	0.44
Year 3	202	45.1 (5.54)	46.36 (41.63−49.55)				
Reaction time							
Year 1	4220	0.57 (0.09)	0.56 (0.52−0.61)	0.45 (0.43, 0.48)	0.45 (0.43, 0.68)	0.003 (−0.17, 0.18)	0.12
Year 2	4220	0.58 (0.08)	0.57 (0.53−0.61)				
Year 1	202	0.62 (0.15)	0.58 (0.53−0.66)	0.24 (0.10, 0.36)	0.24 (0.1, 0.48)	−0.05 (−0.33, 0.24)	−0.44
Year 3	202	0.57 (0.06)	0.56 (0.52−0.61)				

ICCa, Intraclass Correlation Coefficient-Agreement; ICCc, Intraclass Correlation Coefficient-Consistency. λ Levels of agreement change with size of mean. *Bias confidence interval does not cover 0 mean difference

https://doi.org/10.1017/S1355617720000594 Published online by Cambridge University Press

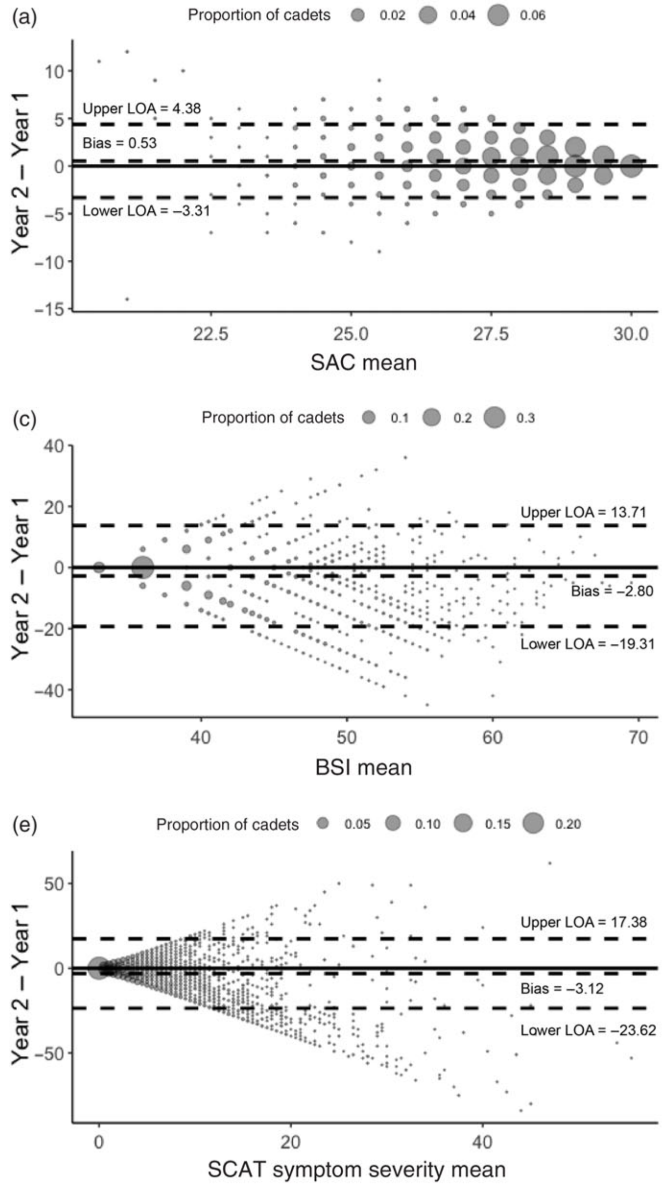

Fig. 1. Bland-Altman Plots of Year 1 and Year 2 Clinical Scores. Bland-Altman plots showing average level of agreement and bias between Year 1 and Year 2 for (a) Standardized Assessment of Concussion, (b) Balance Error Scoring System, (c) Brief Symptom Inventory-Total, (d) SCAT Symptom Number, and (e) SCAT Symptom Severity scores.

(0.44), were observed for the BSI-18, ImPACT visual motor speed, and ImPACT reaction time. Cohen's *d* effect sizes for the varsity and non-varsity analyses appear in Tables S1–S2 and S4–S5. Medium effects were observed for the BSI-18 (0.58) and SCAT Symptom scales (0.55) in varsity cadets. Effect sizes for the freshman versus upperclassman comparison appear in Tables S7–S8 and S10–S11. Medium effects were observed for the BSI-18 (0.68) and SCAT Symptom Evaluation scales (0.72) in the freshmen from Y1 to Y2.

Change score estimates from 75% to 99% confidence are reported for each assessment in Table 3. Change scores for varsity cadets and non-varsity cadets were calculated separately and are available in Supplementary Tables S3 and S6, respectively. Freshmen and upperclassmen change scores are presented in Supplementary Tables S9 and S12, respectively.

DISCUSSION

The objective of this investigation was to establish test–retest reliabilities and clinical interpretation ranges for the annual concussion baseline test battery currently implemented at three U.S. Service Academies participating in the CARE Consortium. Annual baseline testing or testing every other year is a common clinical practice, thus both 1- and 2-year test intervals were examined in this study. Overall, the reliability for these instruments was less than optimal with none

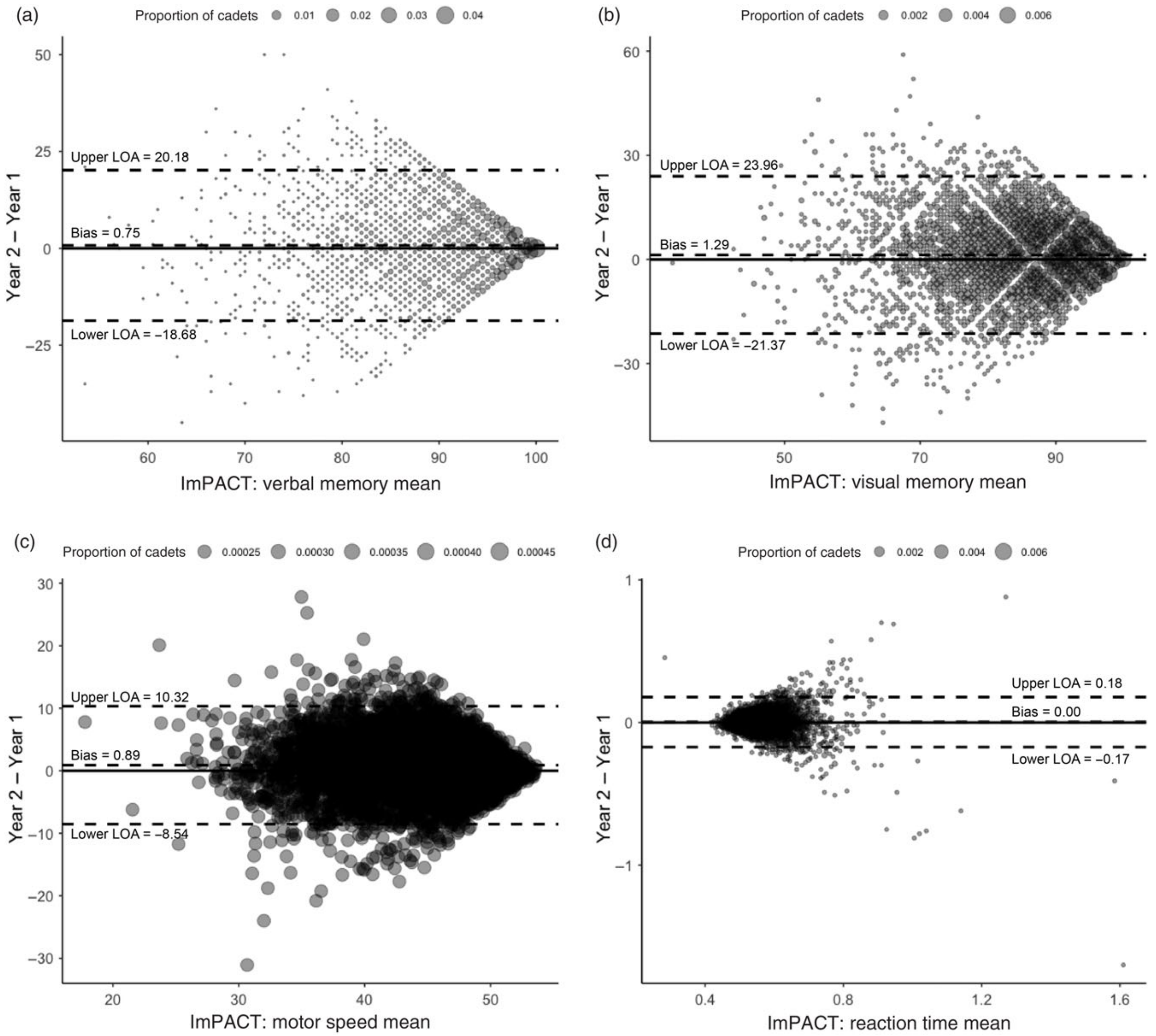

Fig. 2. Bland-Altman Plots of Year 1 and Year 2 ImPACT Scores. Bland-Altman plots showing average level of agreement and bias along between Year 1 and Year 2 for (a) ImPACT Verbal Memory, (b) ImPACT Visual Memory, (c) ImPACT Visual Motor Speed, and (d) ImPACT Reaction Time.

of the metrics nearing an ICC of 0.75, the clinical threshold to suggest optimal stability over time. The findings from this study are generally consistent with the CARE Consortium data that were previously published for NCAA student-athletes from 29 institutions (Broglio et al., 2018), as well as, previous reports for the SAC and BESS (Chin et al., 2016), SCAT-Symptom Evaluations (Register-Mihalik et al., 2013), BSI-18 (Lancaster et al., 2016), and ImPACT (Broglio et al., 2007; Cole et al., 2013; Nelson et al., 2016; Resch et al., 2013). However, the ICC values reported from the current study are much lower than those previously reported.

Cohen's *d* effect sizes were also calculated to evaluate change between baseline test administrations. No effect (Cohen's $d < 0.2$) or small effects (Cohen's $d = 0.2$–0.5) were observed for all assessments for both 1- and 2-year test intervals indicating a substantial overlap in test performance. Even statistically significant effects that have medium or (0.2–0.5) or small (<0.2) effect sizes represent a considerable 80% to 92% overlap on scores between baseline test administrations. Thus, despite the less than optimal ICC values, the limited range of effect sizes (0.01–0.44) suggests substantial overlap, and overall stability, between assessments. While this may seem counter intuitive to the ICC value interpretation, the tightly clustered values may have skewed the ICCs downward.

The effect sizes observed in the current study for SCAT and BSI-18 symptom scores and ImPACT (ES = 0.08–0.44) subscales were slightly greater than the effect sizes from the original CARE Consortium findings that were limited to

https://doi.org/10.1017/S1355617720000594 Published online by Cambridge University Press

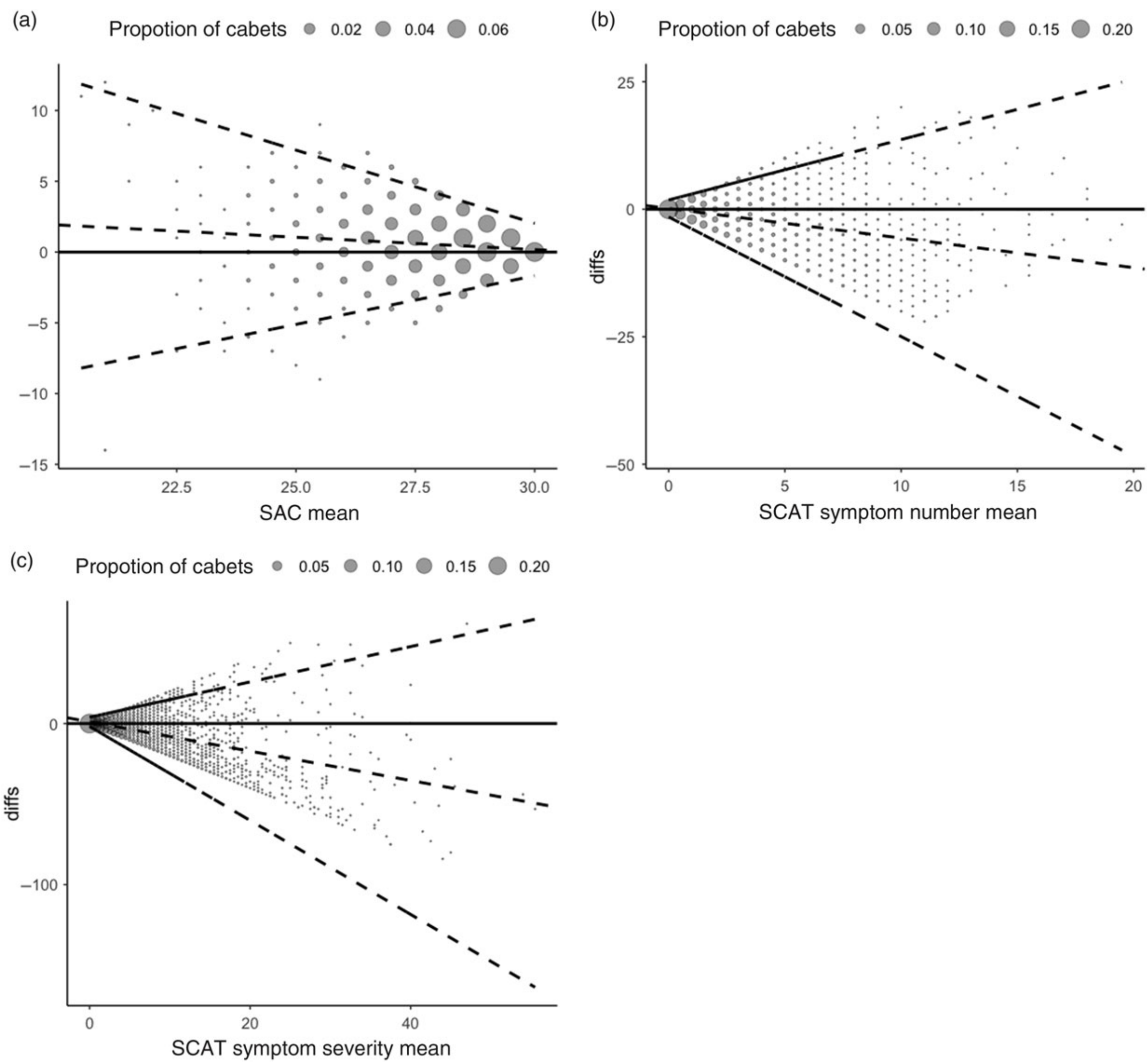

Fig. 3. Bland-Altman Plots of Year 1 and Year 2 for Clinical Scores Where the Assumption of Homoscedasticity Was Not Met. Clinical assessments pictured include: (a) Standardized Assessment of Concussion, (b) SCAT Symptom Number, and (c) SCAT Symptom Severity scores.

NCAA student-athletes (ES = 0.05–0.23) (Broglio et al., 2018). Thus, there may have been more variability in the cadets' symptom and ImPACT scores between baseline test administrations than NCAA athletes at civilian universities. Although very slight, this difference may be attributed to the cadets' ever-changing environments with additional stressors. However, a similar trend was observed for BSI-18 and SCAT symptom scores between varsity cadet-athletes and non-varsity cadet-athletes. While the ICC values were very similar, effect sizes for varsity cadet-athlete symptom scores (ES = 0.57–0.60) were much larger than non-varsity cadet-athletes (ES = 0.28–0.30). This suggests that there may be less overlap and more variability in varsity athlete symptom scores from year-to-year. The effect sizes observed for the SAC, BESS, and ImPACT subscales were fairly consistent in varsity and non-varsity cadet-athletes. The same trend was noted for symptom scores between cadets first baselined as freshmen year and cadets first baselined as an upperclassman. Effect sizes from Y1 to Y2 for freshman symptom scores (ES = 0.68–0.72) were substantially larger than those of upperclassman symptom scores in Y1 and Y2 (ES = 0.01–0.10). This is not surprising as this is most likely the first time many of these first year cadets are exposed to a military environment with added stressors, including basic training, limited contact with friends and family, low sleep, and a rigorous physical and academic schedule. From Y1 to Y2, BSI-18 and SCAT Symptom scores decreased approximately four points in those baselined as freshmen and <1 point in those baselined as an upperclassmen. Similar to the athlete comparison, the effect sizes and scores observed for the clinical assessments were fairly consistent regardless of class.

Based on the effect sizes, we can conclude that there was roughly 80% overlap across clinical assessment distributions between Y1 and Y2 (ES = 0.04–0.31) and Y1 and Y3

https://doi.org/10.1017/S1355617720000594 Published online by Cambridge University Press

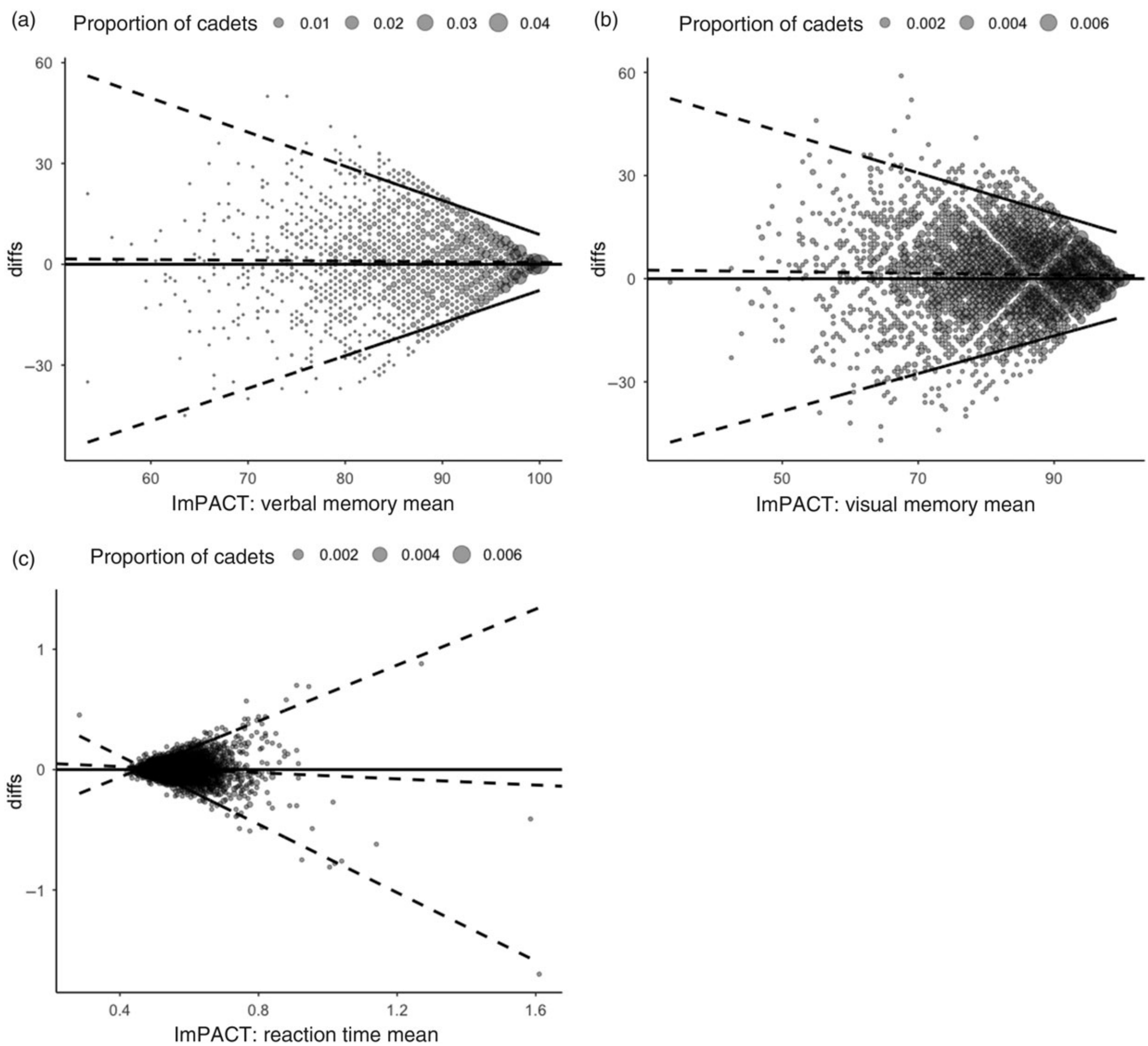

Fig. 4. Bland-Altman Plots of Year 1 and Year 2 ImPACT Scores Where the Assumption of Homoscedasticity Was Not Met. Clinical assessments pictured include: (a) ImPACT Verbal Memory, (b) ImPACT Visual Memory, and (c) ImPACT Reaction Time.

(ES = 0.01–0.44). Despite multiple raters, the BESS scores exhibited >92% overlap from year-to-year exhibiting very small effect sizes (ES = 0.01–0.04). This overlap suggests that the costs associated (i.e., staffing, supplies, time) with annual baseline testing may outweigh the benefits, particularly with the overlap observed in this data from year-to-year. However, those baselined as freshmen demonstrated greater variability between Y1 and Y2 with effect sizes ranging from 0.68 to 0.72 on the BSI-18 and SCAT Symptom assessments, dropping the overlap in assessment scores to roughly 70%. Therefore, re-baselining after freshmen year might be worth the cost benefit analysis.

Re-baselining after freshman year may also add value to post-concussion management. Despite significant overlap between assessments from year-to-year, using the most recent baseline data provided the greatest sensitivity when diagnosing concussions (Broglio et al., 2019). While there is not much variability in assessments from year-to-year, there appears to be value added by having annual baseline assessments. Additionally, other annual concussion efforts, like athlete education, may have a positive impact on concussion disclosure with similar time investment. Thus, moving forward clinicians should consider at least a single administration of these baseline assessments at the time of enrollment.

Following a suspected concussion, established ranges of change scores can provide confidence to the practitioner when interpreting performance on concussion assessments post-injury. To assist with clinical interpretation, change scores were calculated with confidence intervals to express an associated level of certainty that change (e.g., concussion) has occurred. For example, if a patient commits more than 19 errors on the BESS compared to their baseline, the clinician can have 99% confidence that the change is related to something other than normal test–retest variability (see Table 3). Similarly, a 2-point decrease on the SAC would carry 90% confidence. This approach should always be overlaid with clinical presentation and clinician expertise. However, these are the first published values for a unique service academy population.

Table 3. Confidence ranks by change score for all cadets

	75%	80%	87.50%	90%	92.50%	95%	97.50%	99%
SAC								
Year 1–Year 2	−1	−1	−2	−2	−2	−3	−3	−4
Year 1–Year 3	−1	−1	−2	−2	−2	−3	−4	−4
BESS								
Year 1–Year 2	6	7	9	10	12	13	16	19
Year 1–Year 3	6	7	10	11	12	14	17	20
BSI-18								
Year 1–Year 2	3	4	7	8	9	11	14	17
Year 1–Year 3	2	3	6	7	8	9	12	14
SCAT symptom number								
Year 1–Year 2	2	2	4	4	5	6	7	9
Year 1–Year 3	1	1	2	3	3	4	5	6
SCAT symptom severity								
Year 1–Year 2	4	6	9	10	12	14	17	21
Year 1–Year 3	2	3	5	5	6	7	9	11
Verbal memory								
Year 1–Year 2	−6	−8	−11	−12	−14	−16	−19	−22
Year 1–Year 3	−4	−6	−9	−11	−12	−14	−18	−21
Visual memory								
Year 1–Year 2	−7	−8	−12	−14	−15	−18	−21	−26
Year 1–Year 3	−6	−8	−12	−13	−16	−18	−22	−27
Visual motor speed								
Year 1–Year 2	−2	−3	−5	−5	−6	−7	−9	−10
Year 1–Year 3	−1	−2	−4	−4	−5	−6	−8	−10
Reaction time								
Year 1–Year 2	−0.06	−0.07	−0.1	−0.11	−0.13	−0.14	−0.17	−0.21
Year 1–Year 3	−0.15	−0.17	−0.21	−0.23	−0.26	−0.29	−0.33	−0.38

BESS, Balance Error Scoring System; BSI-18, Brief Symptom Inventory-18; SAC, Standardized Assessment of Concussion; SCAT, Sport Concussion Assessment Tool. Scores are rounded to nearest integer.

Our study is not without limitations. The findings discussed in this paper are specific to military service academy cadets and may have application to active duty military service members of comparable age. However, they may not be applicable to younger athletes undergoing brain growth and development nor professional athletes who have likely attained brain maturation. It was also assumed that all participants provided honest effort and accurate answers during testing. Although ImPACT has a validity check, the other assessments do not and some athletes may have intentionally underperformed to hide poor post-concussion performance if injured (Leahy, 2011). It is also possible that the participants may have become apathetic after multiple years of testing. Lastly, we did not document if participants had completed any portion of the assessments prior to attending the service academy, so some may have had prior exposure performing the baseline assessment tasks that influenced their scores.

The SAC, BESS, ImPACT, SCAT – Symptom Evaluation, and BSI-18 scores demonstrated less than optimal reliability in service academy cadets. The ImPACT scores from Y1 to Y2 were the closest to an acceptable level for clinical utility with visual motor speed demonstrating the most stable scores. The SCAT Symptom Evaluation demonstrated the least stable scores from year-to-year. While none of the assessments met or exceeded the accepted clinical threshold (ICC $\geq$ 0.75), the effect sizes were relatively small suggesting an overlap in performance from year-to-year. Thus, the lack of stability in the scores combined with the weak effect sizes suggests that the clinical assessments are likely representative of state function of overt traits and will continue to vary with more testing. As such, baseline assessments beyond the initial evaluation in service academy cadets are not essential but could aid concussion diagnosis. Until a sound clinical tool is developed, we recommend at least an initial baseline assessment upon entry to a U.S. Service Academy. The baseline scores can be used in combination with the change scores to improve practitioner confidence when managing concussions.

ACKNOWLEDGMENTS

The authors would like to thank Bonnie Campbell, Lisa Campbell, Megan Jackson, Jennifer Miley, Joel Robb, and Kim Robb (United States Air Force Academy), Robin Miller and Jarrett Headley (United States Coast Guard Academy), Stephanie Carminati, Steven Malvasi, Story Miraldi, Jamie Reilly, Sean Roach, and Jesse Trump (United States Military Academy) for data acquisition, as well as, the research and medical staff that assisted with baseline data collection at each of the service academy sites. The authors would also like to thank Jaroslaw Harezlak, Jody Harland, Janetta Matesan, Larry Riggen (Indiana University), Ashley Rettmann, Nicole L'Heureux (University of Michigan), Melissa Koschnitzke (Medical College of Wisconsin), Michael Jarrett, Vibeke Brinck, and Bianca Byrne (Quesgen), Thomas Dompier, Erin B. Wasserman, Melissa Niceley Baker, and Sara Quetant (Datalys Center for Sports Injury Research and Prevention). This publication was made possible, in part, with support from the Grand Alliance Concussion Assessment, Research, and Education Consortium, funded by the National Collegiate Athletic Association and the Department of Defense. The U.S. Army Medical Research Acquisition Activity, 820 Chandler Street, Fort Detrick, MD 21702-5014, USA is the awarding and administering acquisition office. This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Psychological Health and Traumatic Brain Injury Program under Award no. W81XWH-14-2-0151. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense (Defense Health Program funds).

CONFLICT OF INTEREST

The authors have nothing to disclose.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit https://doi.org/10.1017/S1355617720000594

REFERENCES

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

Barth, J.T., Alves, W.M., Ryan, T.V., Macciocchi, S.N., Rimel, R.W., Jane, J.A., & Nelson, W.E. (1989). Mild head injury in sports: Neuropsychological sequelae and recovery of function. *Mild Head Injury*, 257–275.

Bell, D.R., Guskiewicz, K.M., Clark, M.A., & Padua, D.A. (2011). Systematic review of the balance error scoring system. *Sports Health*, *3*(3), 287–295.

Broglio, S.P., Cantu, R.C., Gioia, G.A., Guskiewicz, K.M., Kutcher, J., Palm, M., & Valovich McLeod, T.C. (2014). National Athletic Trainers' Association position statement: Management of sport concussion. *Journal of Athletic Training*, *49*(2), 245–265.

Broglio, S.P., Ferrara, M.S., Macciocchi, S.N., Baumgartner, T.A., & Elliott, R. (2007). Test-retest reliability of computerized concussion assessment programs. *Journal of Athletic Training*, *42*(4), 509–514.

Broglio, S.P., Ferrara, M.S., Sopiarz, K., & Kelly, M.S. (2008). Reliable change of the sensory organization test. *Clinical Journal of Sport Medicine*, *18*(2), 148–154.

Broglio, S.P., Harezlak, J., Katz, B., Zhao, S., McAllister, T., McCrea, M., & Investigators, C.C. (2019). Acute sport concussion assessment optimization: A prospective assessment from the CARE consortium. *Sports Medicine*, *49*(12), 1977–1987.

Broglio, S.P., Katz, B.P., Zhao, S., McCrea, M., & McAllister, T. (2018). Test-retest reliability and interpretation of common concussion assessment tools: Findings from the NCAA-DoD CARE consortium. *Sports Medicine*, *48*(5), 1255–1268.

Broglio, S.P., McCrea, M., McAllister, T., Harezlak, J., Katz, B., Hack, D., & Hainline, B. (2017). A national study on the effects of concussion in collegiate athletes and US Military Service Academy members: The NCAA-DoD concussion assessment, research and education (CARE) consortium structure and methods. *Sports Medicine*, *47*(7), 1437–1451.

Broglio, S.P., Zhu, W., Sopiarz, K., & Park, Y. (2009). Generalizability theory analysis of balance error scoring system reliability in healthy young adults. *Journal of Athletic Training*, *44*(5), 497–502.

Cameron, K.L., Marshall, S.W., Sturdivant, R.X., & Lincoln, A.E. (2012). Trends in the incidence of physician-diagnosed mild traumatic brain injury among active duty U.S. military personnel between 1997 and 2007. *J Neurotrauma*, *29*(7), 1313–1321.

Centers for Disease Control and Prevention. (2016). Basic information about traumatic brain injury and concussion. Retrieved from http://www.cdc.gov/traumaticbraininjury/basics.html

Chan, M., Vielleuse, J.V., Vokaty, S., Wener, M.A., Pearson, I., & Gagnon, I. (2013). Test-retest reliability of the sport concussion assessment tool 2 (SCAT2) for uninjured children and young adults. *British Journal of Sports Medicine*, *47*(5), e1.

Chin, E.Y., Nelson, L.D., Barr, W.B., McCrory, P., & McCrea, M.A. (2016). Reliability and validity of the Sport Concussion Assessment Tool-3 (SCAT3) in high school and collegiate athletes. *The American Journal of Sports Medicine*, *44*(9), 2276–2285.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (Rev. ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cole, W.R., Arrieux, J.P., Schwab, K., Ivins, B.J., Qashu, F.M., & Lewis, S.C. (2013). Test-retest reliability of four computerized neurocognitive assessment tools in an active duty military population. *Archives of Clinical Neuropsychology*, *28*(7), 732–742.

Defense and Veterans Brain Injury Center. (2018). *Department of Defense worldwide numbers for traumatic brain injury.* Retrieved from http://dvbic.dcoe.mil/files/tbi-numbers/WorldwideTotals2000-2017Q1-Q3Nov%2014-2017508.pdf

Department of Defense. (2015). *Department of Defense Instruction 6490.13: Comprehensive policy on traumatic brain injury-related neurocognitive assessments by the military services*. Retrieved from https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/649013p.pdf

Farnsworth, J.L., 2nd, Dargo, L., Ragan, B.G., & Kang, M. (2017). Reliability of computerized neurocognitive tests for concussion assessment: A meta-analysis. *Journal of Athletic Training*, *52*(9), 826–833.

Grilo, L., & Grilo, H. (2012). Comparison of clinical data based on limits of agreement. *Biomedical Letters*, *49*(1), 45–56.

Herring, S.A., Cantu, R.C., Guskiewicz, K.M., Putukian, M., Kibler, W.B., Bergfeld, J.A., … Indelicato, P.A. (2011). Concussion (mild traumatic brain injury) and the team physician: A consensus statement – 2011 update. *Med Sci Sports Exerc*, *43*(12), 2412–2422.

Hinton-Bayre, A.D., Geffen, G.M., Geffen, L.B., McFarland, K.A., & Frijs, P. (1999). Concussion in contact sports: Reliable change indices of impairment and recovery. *Journal of Clinical and Experimental Neuropsychology*, *21*(1), 70–86.

Hoffman, J.I. (2015). Linear regression. In J.I. Hoffman (Ed.), *Biostatistics for medical and biomedical practitioners* (pp. 451–500). Amsterdam: Elsevier.

Iverson, G.L. (2011). Reliable change index. In J.S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of clinical neuropsychology* (pp. 2150–2153). New York, NY: Springer.

Iverson, G.L., Lovell, M.R., & Collins, M.W. (2003). Interpreting change on ImPACT following sport concussion. *The Clinical Neuropsychologist*, *17*(4), 460–467.

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.

Lancaster, M.A., McCrea, M.A., & Nelson, L.D. (2016). Psychometric properties and normative data for the Brief Symptom Inventory-18 (BSI-18) in high school and collegiate athletes. *The Clinical Neuropsychologist*, *30*(2), 338–350.

Langlois, J.A., Rutland-Brown, W., & Wald, M.M. (2006). The epidemiology and impact of traumatic brain injury: A brief overview. *Journal of Head Trauma Rehabilitation*, *21*(5), 375–378.

Leahy, S. (2011, April 27). Peyton Manning admits to tanking NFL’s baseline concussion test. USA Today.

Lichtenstein, J.D., Moser, R.S., & Schatz, P. (2014). Age and test setting affect the prevalence of invalid baseline scores on neurocognitive tests. *The American Journal of Sports Medicine*, *42*(2), 479–484.

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.

McCrory, P., Meeuwisse, W., Dvorak, 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.

McCrory, P., Meeuwisse, W.H., Aubry, M., Cantu, B., Dvorak, J., Echemendia, R.J., Engebretsen, L., . . . Turner, M. (2013). Consensus statement on concussion in sport: The 4th International Conference on Concussion in Sport held in Zurich, November 2012. *British Journal of Sports Medicine*, *47*(5), 250–258.

McLeod, T.C., & Leach, C. (2012). Psychometric properties of self-report concussion scales and checklists. *Journal of Athletic Training*, *47*(2), 221–223.

Meachen, S.J., Hanks, R.A., Millis, S.R., & Rapport, L.J. (2008). The reliability and validity of the brief symptom inventory-18 in persons with traumatic brain injury. *Archives of Physical Medicine and Rehabilitation*, *89*(5), 958–965.

Mehta, S., Bastero-Caballero, R.F., Sun, Y., Zhu, R., Murphy, D.K., Hardas, B., & Koch, G. (2018). Performance of intraclass correlation coefficient (ICC) as a reliability index under various distributions in scale reliability studies. *Statistics in Medicine*, *37*(18), 2734–2752.

Nelson, L.D., LaRoche, A.A., Pfaller, A.Y., Lerner, E.B., Hammeke, T.A., Randolph, C., . . . McCrea, M.A. (2016). Prospective, head-to-head study of three computerized neurocognitive assessment tools (CNTs): Reliability and validity for the assessment of sport-related concussion. *Journal of the International Neuropsychological Society*, *22*(1), 24–37.

Portney, L.G., & Watkins, M.P. (2009). *Foundations of clinical research: Applications to practice*. Upper Saddle River, NJ: Prentice Hall.

Ratcliff, R., Smith, P.L., Brown, S.D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Register-Mihalik, J.K., Guskiewicz, K.M., Mihalik, J.P., Schmidt, J.D., Kerr, Z.Y., & McCrea, M.A. (2013). Reliable change, sensitivity, and specificity of a multidimensional concussion assessment battery: Implications for caution in clinical practice. *Journal of Head Trauma Rehabilitation*, *28*(4), 274–283.

Resch, J., Driscoll, A., McCaffrey, N., Brown, C., Ferrara, M.S., Macciocchi, S., . . . & Walpert, K. (2013). ImPact test-retest reliability: Reliably unreliable? *Journal of Athletic Training*, *48*(4), 506–511.

Riemann, B.L., Guskiewicz, K.M., & Shields, E.W. (1999). Relationship between clinical and forceplate measures of postural stability. *Journal of Sport Rehabilitation*, *8*(2), 71–82.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Snell, F.I., & Halter, M.J. (2010). A signature wound of war: Mild traumatic brain injury. *Journal of Psychosocial Nursing and Mental Health Services*, *48*(2), 22–28.

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.