Two-year Test–Retest Reliability in High School Athletes Using the Four- and Two-Factor ImPACT Composite Structures: The Effects of Learning Disorders and Headache/Migraine Treatment History

Benjamin L. Brett[1,2,*], Gary S. Solomon[2,3], Jennifer Hill[4], Philip Schatz[4]

[1]*Department of Counseling, Educational Psychology and Research, The University of Memphis, Memphis, TN, USA*
[2]*Vanderbilt Sports Concussion Center, Nashville, TN, USA*
[3]*Department of Neurological Surgery, Vanderbilt University School of Medicine, Nashville, TN, USA*
[4]*Department of Psychology, Saint Joseph's University, Philadelphia, PA, USA*

*Corresponding author at: Department of Counseling, Educational Psychology and Research, The University of Memphis, 100 Ball Hall, Memphis, TN 38152, USA. Tel.: +1-901-678-2841; fax: +1-901-678-5114.
E-mail address: blbrett@memphis.edu (B.L. Brett).

Abstract

Objective: This study examined the test–retest reliability of the four- and two-factor structures (i.e., Memory and Speed) of ImPACT over a 2-year interval across multiple groups with premorbid conditions, including those with a history of special education or learning disorders (LD; $n = 114$), treatment history for headache/migraine ($n = 81$), and a control group ($n = 792$).
Methods: Nine hundred and eighty seven high school athletes completed baseline testing using online ImPACT across a 2-year interval. Paired-samples *t*-tests documented improvement from initial to follow-up assessments. Test stability was examined using Regression-based measures (RBM) and Reliable change indices (RCI). Reliability was examined using intraclass correlation coefficients (ICC).
Results: Significant improvement on all four composites were observed for the control group over a 2-year interval; whereas significant differences were observed only on Visual Motor Speed for the LD and headache/migraine treatment history groups. ICCs ranges were similar across groups and greater or comparable reliability was observed for the two-factor structure on Memory (0.67–0.73) and Speed (0.76–0.78) composites. RCIs and RBMs demonstrated stability for the four- and two-factor structures, with few cases falling outside the range of expected change within a healthy sample at the 90% and 95% CIs.
Conclusion: Typical practices of obtaining new baselines every 2 years in the high school population can be applied to athletes with a history of special education or LD and headache/migraine treatment. The two-factor structure has potential to increase test–retest reliability. Further research regarding clinical utility is needed.

Keywords: Childhood brain insult; Developmental and learning disabilities; Assessment; Practice effects/reliable change; Norms/normative studies; Sports concussion

Introduction

Serial neuropsychological assessment is commonly included as part of the comprehensive management of sport-related concussion (SRC; Covassin, Elbin, & Stiller-Ostrowski, 2009; Harmon et al., 2013; McCrory et al., 2013), although the effective practice of the serial assessment method is inextricably tied to the reliability of the instrument being used (Bauer et al., 2012). When using athletes' baseline performance as the reference for meaningful change, inadequate test–retest (TrT) reliability can be problematic, as scores not reflecting an athlete's true neurocognitive abilities can result in decreased sensitivity in the detection of cognitive changes associated with SRC (Schatz & Sandel, 2013). The use of baseline neurocognitive data for comparison against post-concussion performance necessitates having an accurate measure of an athlete's baseline abilities,

which can present challenges when youth, high school, and collegiate athletes are still experiencing cognitive maturation. The CDC has recommended that baseline assessments be updated/repeated every 2 years for youth athletes (CDC, 2015), although researchers have recognized that baselines may need to be updated more frequently for younger athletes (Buzzini & Guskiewicz, 2006). Repeating or updating baseline assessments every 2 years is a common practice in both high school (Wojtowicz et al., 2017a, 2017b) and collegiate (Schatz, 2010) concussion management programs. As such, the need to establish consistently stable reliability estimates for neurocognitive measures used in the evaluation of SRC is crucial. Among the most commonly used tools for measuring neurocognitive functioning in the serial assessment and management of SRC in athletes is the Immediate Post-Concussion Assessment and Cognitive Testing (Covassin et al., 2009; ImPACT, 2012; Maerlender et al., 2010).

Based upon an established classification system (Slick, 2006; reliability estimate ranges include ≥.90 = very high; .80–.89 = high; .70–.79 = adequate; .60–.69 = marginal; <.60 = low), previous investigations into the test–retest reliability of ImPACT have yielded variable results. In many instances, ImPACT has demonstrated a stable range of test–retest reliability coefficients across various intervals, ranging from 7 days to 3 years (see Table 1; Brett, Smyke, Solomon, Baughman, & Schatz, 2016; Bruce, Echemendia, Meeuwisse, Comper, & Sisco, 2014; Elbin, Schatz, Covassin, 2011; Iverson, Lovell, Collins, 2003; Maerlender & Molfese, 2015; Miller Adamson, Pink, & Sweet, 2007; Nakayama, Covassin, Schatz, Nogle, & Kovan, 2014; Schatz, 2010; Schatz, Pardini, Lovell, Collins, & Podell, 2006). In contrast, other investigations of ImPACT have yielded contradictory results, demonstrating lower bound estimates of test–retest reliability for ImPACT composite and symptom scale scores (Broglio, Ferrara, Macciocchi, Baumgartner, & Elliot, 2007; Cole et al., 2013; Resch et al., 2013). For a comprehensive review, see Alsalaheen et al. (2016) and Nelson et al. (2016). A summary of these studies is presented in Table 1.

Although time between test administrations is a primary factor affecting test–retest reliability (Dikmen, Heaton, Grant, & Temkin, 1991; McCaffrey, Ortega, & Hasse, 1993), additional sources contributing to variation have been identified. For instance, age-related differences have been observed as influencing reliability, with significant improvement observed on CogSport in children between the ages of 9 and 15 (McCrory, Collie, Anderson, & Davis, 2004) and college-aged athletes outperforming high school athletes on ImPACT Visual Motor Speed (Register-Mihalik et al., 2012). However, variations in reliability have been demonstrated within studies examining the high school population, suggesting that additional biopsychosocial factors may play a role in the stability of neurocognitive test scores. One such factor includes the presence of a learning disability (LD)/history of special education, as significant differences on neurocognitive testing have been observed in this population when compared to a control group (Elbin et al., 2013; Zuckerman, Lee, Odom, Solomon, & Sills, 2013). Specifically, those with LD demonstrated a lower performance on ImPACT in the areas of Verbal Memory, Visual Memory, Visual Motor Speed, and Reaction Time (Elbin et al., 2013; Zuckerman et al., 2013). Additionally, athletes with LD are significantly more likely to score below at least one pre-established threshold indicating invalid performance on ImPACT (Schatz, Moser, Solomon, Ott, & Karpf, 2012).

Table 1. Variation in Pearson's *r* and intraclass correlation coefficients (ICC) from studies of ImPACT test–retest reliability

Variable	Test–retest Interval								
	7 days[a]	30 days[b]	30 days[c]	45 days[d]	45 days[e]	45 days[f]	6 Months[g]	1 year[h]	2 years[i]
Study *N*	56	25	215	73	46/45	85	200	369	95
Verbal Memory (*ICC*)	0.70	0.79	0.60	0.23	0.56/45	0.77	0.71/0.52	0.62	0.46
(*Pearson r*)	*	0.66	0.61	*	*	0.63	0.57/0.35	0.45	0.30
Visual Memory (*ICC*)	0.67	0.60	0.50	0.32	0.26/0.52	0.72	0.70/0.75	0.70	0.65
(*Pearson r*)	*	0.43	0.49	*	*	0.58	0.54/0.61	0.55	0.49
Visual Motor (*ICC*)	0.86	0.88	0.53	0.38	0.78/0.76	0.86	0.77/0.86	0.82	0.74
(*Pearson r*)	*	0.78	0.53	*	*	0.74	0.63/0.75	0.74	0.60
Reaction Time (*ICC*)	0.79	0.77	0.86	0.39	0.84/0.57	0.71	0.70/0.76	0.71	0.68
(*Pearson r*)	*	0.63	0.83	*	*	0.52	0.54/0.61	0.62	0.52

[a]Iverson et al. (2003).
[b]Schatz and Ferris (2013).
[c]Cole et al. (2013).
[d]Broglio et al. (2007).
[e]Resch et al. (2013).
[f]Nakayama et al. (2014).
[g]Womble, Reynolds, Schatz, Shah, and Kontos (2016).
[h]Elbin, Schatz, and Covassin (2011).
[i]Schatz (2010).

A history of headache/migraine treatment may also contribute to variation observed in ImPACT test–retest reliability values. The relationship between a history of headache treatment and differences in baseline neurocognitive performance has been previously documented in professional athletes, as those with a history of headache treatment exhibited lower Verbal Memory scores and significantly lower Visual Memory scores than others in the sample (Solomon & Haase, 2008). A treatment history for migraines may also potentially influence test–retest reliability, as those with migraine disorders who were symptomatic at one of two test administrations exhibited a higher frequency of reliable change on a computerized neurocognitive measure (Roebuck-Spencer, Sun, Cernich, Farmer, & Bleiberg, 2007); however, these differences in the aforementioned study may have been due to confounding variables listed in the limitations (e.g. sample size and selection, the absence of an independent measure of impairment other than the neuropsychological test themselves, etc.). Further, lower scores on ImPACT (Verbal Memory, Visual Memory, and Visual Motor Speed) have been recorded in athletes with a history of migraine treatment and additional comorbidities (Wojtowicz et al., 2017a, 2017b). Considered as a modifying factor of SRC (McCrory et al., 2017), a history of migraines has also been associated with prolonged recovery following injury (Lau, Lovell, Collins, & Pardini, 2009; Tator et al., 2016). Given this relationship between headache/migraine treatment history and baseline neurocognitive testing, as well as recovery, it is possible that a history headache/migraine treatment may also be a source of variation in ImPACT test–retest reliability metrics.

Factor structures provided in the original version of ImPACT (Version 1.2) included three factors, or composite scores, including Memory, Reaction Time, and Visual Motor Speed, as well as Impulse Control (the latter was used for test validity interpretation; Lovell, 2007). The most recent version of ImPACT includes a four-factor structure of composite scores, including Verbal Memory, Visual Memory, Visual Motor Speed, Reaction Time, and Impulse Control. For a more comprehensive review of the evolution of ImPACT, as well as related factor analytic studies, see Schatz and Maerlender (2013). Recently, evidence for the validity of a two-factor ImPACT structure has been demonstrated through two confirmatory factor analytical studies (Gerrard et al., 2017; Schatz & Maerlender, 2013), documenting shared variance between current composite scores of Verbal Memory and Visual Memory into a single "Memory" composite, as well as Reaction Time and Visual Motor Speed composites into an integrated composite of "Speed."

Preliminary studies have demonstrated increased temporal stability of ImPACT using the two-factor structure (Bruce et al., 2016; Schatz & Maerlender, 2013). Specifically, levels of test–retest reliability for Memory and Speed factors have been observed as acceptable in a sample of undergraduate students at 30 days (.81/.88) and 2 years (.74/.76), respectively (Schatz & Maerlender, 2013). Acceptable levels of reliability have also been demonstrated at a 1-year test–retest interval for Memory (.76) and Speed (.85) in a sample of high school athletes. In addition to enhanced reliability estimates, the two-factor theory has yielded improved diagnostic utility through higher levels of sensitivity (89%) and specificity (70%), as compared to the sensitivity (80%) and specificity (62%) calculated for the same sample using the established four factor composite scores (Schatz & Maerlender, 2013). While these findings contain promise for the application of the two-factor structure of ImPACT, further validation of these outcomes should be demonstrated through replication.

Due to the fact that the original normative sample provided by ImPACT (2012) does not provide precise means (50th percentile) and standard deviations (SDs; 16th and 84th percentile), one challenge facing researchers is where to obtain appropriate means and SDs for calculating two-factor scores. Although age- and gender-based percentile rankings are provided, there is no available method for extrapolating the appropriate means and SDs. For example, for Males Ages 16–18, a Verbal Memory score of 84 corresponds to a percentile rank of 48, and a Verbal Memory score of 85 corresponds to a percentile rank of 52, which presents the challenge of precisely determining the sample mean (corresponding to the 50th percentile). Similarly, given that the 16th and 84th percentiles correspond to 1 *SD* from the mean, these reflect either a score of 74 (16th percentile) or 94 (84th percentile). While these appear to be 10 points on either side of the mean, a middle-score of 84 is not an accurate representation of the mean. Further, Reaction Time values (within the same gender/age) are provided which correspond to the 49th (0.58) and 54th (0.57) percentiles, as well as the 16th (0.67) and the 83rd (0.51). Even if one were to "accept" 0.58 as reflective of the 50th percentile and 0.51 as the 84th percentile, the scores are not uniformly dispersed about the mean, with a 0.09 difference between the 16th and 50th percentile scores, and 0.07 between the 50th and 84th percentile scores. As a result, when calculating two-factor scores, researchers have utilized baseline means and (SD) which have, to-date, been obtained from sample-specific or regional databases. In the Schatz & Maerlender (2013) study, which first presented the two-factor method, sample-specific means and SD were utilized from four different studies (one-month test–retest [college], 1-year test–retest [high school], and 2-year test–retest [college] samples, and an acute concussion sample [college]). Bruce and colleagues (2016) utilized sample-specific baseline means and SD from professional hockey players. Most recently, Gerrard and colleagues (2017) utilized sample-specific means and SD from a large, regional database, in order to calculate two-factor scores in a high school sample.

The aim of the current study was to compare the test–retest reliability of ImPACT four- and two-factor scores across a 2-year interval in samples of high school athletes with: (1) a history of special education or LD, (2) a history of treatment for

headache/migraine, (3) neither of these conditions (Control group). Given the previously demonstrated higher rates of invalidity indicators and lower scores on neurocognitive testing in the LD population, it is expected that lower test–retest reliabilities will be observed within the LD group, as compared to controls. Given the previously observed effect of migraine symptoms and treatment on test stability (Roebuck-Spencer et al., 2007), it is similarly expected that reliability estimates will be lower within the headache/migraine treatment group, as compared to the control group. In addition, based on previous research, it was expected that two-factor composite scores would demonstrate higher test–retest reliability estimates versus the four-factor factor structure across all groups within the study.

Materials and Methods

Participants

Data were collected through pre-season cognitive baseline assessments by the athletic departments at 30 high schools from middle Tennessee from 2010 to 2016. Anonymous, deidentified data were obtained for psychometric assessment from the Lead Programmer at ImPACT, who was blind to the purpose of the study. Institutional Review Board approval (exemption) was obtained for this study. Baseline tests, obtained using the online version of ImPACT, Version 2.1, were administered in groups by a certified athletic trainer who had been trained in the group administration of ImPACT. Participants were high school athletes, ages 13–18 years who completed two pre-season baseline assessments (Table 2). All assessments were administered in groups, and proctored by a certified athletic trainer who had been trained in the administration of ImPACT. The exact group size of administration across the different participating schools was unknown.

A priori exclusion criteria included those who: (1) obtained an invalid baseline (as denoted by the "Baseline++" on ImPACT test results; ImPACT, 2012), (2) sustained a concussion during the test–retest interval, and (3) were non-native English-speaking students. Additional exclusion criteria included a self-reported history of (1) epilepsy/seizures, (2) brain surgery, (3) meningitis, (4) depression or anxiety, and (5) substance/alcohol abuse.

Athletes were classified into the LD group if they self-reported a history of having received special education services, or a diagnosis of a learning disability ($n = 114$). Athletes were classified into the headache/migraine treatment history group if they self-reported a history of treatment for headaches and/or migraines ($n = 81$). Athletes within the control group ($n = 792$) did not report a history of special education/LD or headache/migraine treatment. The final sample was composed of 987 athletes who completed two baseline assessments between 19 and 30 months (mean interval in months = 23.21, $SD = 1.19$) while attending high school.

Table 2. Demographics

	n (%)	Sport	n (%)
Sex			
Male	615 (62.3)	Football	385 (39.0)
Female	372 (37.7)	Soccer	216 (21.9)
		Volleyball	99 (10.0)
Age[a]	15.00 ± 0.71	Lacrosse	69 (7.0)
		Basketball	60 (6.1)
Condition		Baseball	54 (5.5)
Control	792 (80.2)	Cheerleading	37 (3.7)
		Wrestling	29 (2.9)
Learning disorder	114 (11.6)	Softball	28 (2.8)
		X-Country	4 (0.4)
Headache/migraine treatment[b]	81 (8.2)	Tennis	2 (0.2)
Headache	65	Track & Field	2 (0.2)
Migraine	44	Other	2 (0.2)
Time between baseline[a]	23.21 ± 1.19		

[a]Data presented as mean & standard deviation of months between baseline.
[b]Number of conditions endorsed exceeds total participants due to multiple conditions being endorsed.

Materials

ImPACT is a computer-based program used to assess neurocognitive function and concussion symptoms. The test consists of six individual subtests that yield composite scores mentioned previously (see Iverson, Franzen, Lovell, & Collins, 2004 for more detail on the subscales and Schatz et al., 2006 for information regarding the psychometric properties of ImPACT), as well as a self-report total symptom scale, composed of 22 common symptoms, each rated on a 0–6 Likert scale, with 0 = none and 6 = severe, reported by the athlete.

Procedures and Data Analyses

Participants completed two baseline assessments, allowing for a comparison between Time 1 and Time 2 baseline scores. The dependent measures for this study included the four ImPACT clinical composite scores. Following the precedent provided by the studies described earlier, means and standard deviations (SD) for age and gender were calculated using normative data obtained from the aforementioned athletic departments. These, in turn, were utilized in calculating two-factor composite factors, for each group (see Table 3). The two-factor structure for each participant was calculated through z-scores (i.e., Memory and Speed), by subtracting each of the athletes' four composite scores from the mean of a normative data sample, and dividing by the SD of the sample. The z-scores for Memory were then calculated by taking the average of the Verbal Memory and the Visual Memory z-scores. The Speed composite was obtained by taking the average of the z-scores for Visual Motor Speed and the Reaction Time.

Paired samples *t*-tests were also utilized as a means of assessing significant differences between scores at baseline (Time 1) and 2-year follow-up (Time 2; Bonferroni corrected alpha level = 0.002). Intraclass correlation coefficients (ICCs) were calculated as a measure of test–retest reliability, reflecting of the strength of relationship of composite scores at Times 1 and 2. ICCs included "2-way mixed" type "consistency" analyses, which is a more advantageous measure of association for investigations of test–retest over Pearson's *r* correlation, due to its ability to provide an unbiased estimate of reliability based on the consistency of baseline assessments from test to retest within subjects (Vaz, Falkmer, Passmore, Parsons, & Andreou, 2013). Interpretation of reliability estimate ranges were based upon the classification system described earlier (Slick, 2006). Reliable change indices (RCIs; Jacobson & Truax, 1991), including adjustments for practice effects (Iverson, 2001), were calculated to assess whether changes between repeated baseline assessments represented meaningful change for the four factor composite

Table 3. Regional ImPACT normative data, by age and gender

Composite	Age	*N* (8134)	Female	Male
Verbal Memory	13	427	85.55 (9.3)	81.95 (10.2)
	14	2788	85.78 (9.8)	83.16 (10)
	15	1877	85.32 (9.8)	83.48 (10.1)
	16	1214	86.67 (9.9)	83.16 (10)
	17	989	86.64 (9.6)	84.14 (10.3)
	18	839	88.03 (9.8)	85.94 (9.3)
Visual Memory	13	427	73.71 (12.4)	71.56 (13.2)
	14	2788	74.88 (12.5)	73.97 (13)
	15	1877	73.61 (13.1)	73.85 (12.9)
	16	1214	75 (12.9)	74.22 (13.2)
	17	989	75.79 (13.6)	73.97 (13.4)
	18	839	75.15 (11.8)	77.4 (12.9)
Visual Motor Speed	13	427	34.43 (5.5)	31.57 (5.7)
	14	2788	35.8 (5.8)	33.68 (6.2)
	15	1877	36.68 (6.1)	35.12 (6.3)
	16	1214	38.41 (6.4)	36.79 (6.8)
	17	989	39.48 (6.3)	37.43 (7)
	18	839	40.61 (6.3)	40.07 (6.6)
Reaction Time	13	427	0.68 (0.1)	0.67 (0.1)
	14	2788	0.64 (0.09)	0.64 (0.09)
	15	1877	0.63 (0.09)	0.63 (0.09)
	16	1214	0.61 (0.09)	0.61 (0.09)
	17	989	0.61 (0.09)	0.61 (0.08)
	18	839	0.59 (0.1)	0.59 (0.09)

Numbers represent Mean (SD).

scores. Regression-based measures (RBMs; McSweeny, Naugle, Chelune, Gordon, & Lüders, 1993) were also used in order to assess whether participants' performance on repeat assessments meaningfully deviated from predicted scores based on initial baseline testing scores. Ultimately, RBM and RCI are sensitive methods of meaningful change, as differences in test performance are uniquely compared to the baseline scores of that individual (Duff, 2012; Iverson, 2001; McSweeny, et al., 1993). Given this fact, RBM and RCI are likely to be a more sensitive method in detecting significant group differences based on individual change, rather than simple score discrepancies used in t-tests and other statistics dependent on mean differences (Duff, 2012).

Results

Significant improvement was seen within the Control group between Time 1 and Time 2 for Verbal Memory, $t(791) = −4.3$, $p < .001$, $d = 0.31$, Visual Memory, $t(791) = −6.24$, $p < .001$, $d = 0.44$, Visual Motor Speed $t(791) = −20.4$, $p < .001$, $d = 1.45$, and Reaction Time $t(791) = 7.88$, $p < .001$, $d = 0.56$ (see Table 4). Significant improvement in Visual Motor Speed between Time 1 and Time 2 was also observed for the LD group $t(113) = −8.17$, $p < .001$, $d = 1.54$, and headache/migraine treatment group $t(80) = −3.97$, $p < .001$, $d = 0.89$. Two factor composites yielded comparable results and significant differences were not observed between Time 1 and Time 2 on Memory composites for the LD and headache/migraine treatment groups. Controls significantly increased in performance on the Memory composite score $t(791) = −6.93$, $p < .001$, $d = 0.49$, as well as on the Speed composite $t(791) = −17.54$, $p < .001$, $d = 1.25$. Significant improvement of Speed composite scores between

Table 4. Test–retest reliability of groups

Variable	Time 1 Mean (SD)	Time 2 Mean (SD)	ICC[a]	95% CI[b]	95% CI[c]	t[d]	p[d]	Effect size[e]
Four factor composites								
Verbal Memory								
Control	85.07 (9.53)	86.70 (9.52)	0.54	0.48	0.60	−4.3	<.001	0.31
LD	83.58 (8.66)	83.24 (10.34)	0.46	0.22	0.63	−0.17	.87	*
Headache/migraine	84.24 (9.99)	85.56 (10.17)	0.42	0.1	0.63	−0.97	.33	*
Visual Memory								
Control	75.19 (12.20)	78.03 (12.40)	0.63	0.57	0.68	−6.24	<.001	0.44
LD	72.10 (12.48)	73.92 (12.27)	0.68	0.54	0.78	−1.61	.11	*
Headache/migraine	74.63 (13.28)	74.90 (14.04)	0.74	0.59	0.83	−0.2	.85	*
Motor Speed								
Control	35.38 (6.07)	38.76 (6.02)	0.83	0.8	0.85	−20.4	<.001	1.45
LD	34.47 (6.42)	37.81 (6.41)	0.87	0.81	0.91	−8.17	<.001	1.54
Headache/migraine	36.15 (6.72)	38.81 (7.56)	0.79	0.67	0.86	−3.97	<.001	0.89
Reaction Time								
Control	0.623 (0.080)	0.600 (0.076)	0.61	0.56	0.66	7.88	<.001	0.56
LD	0.637 (0.105)	0.609 (0.071)	0.54	0.34	0.69	3.04	.003	*
Headache/migraine	0.630 (0.085)	0.617 (0.095)	0.59	0.36	0.74	1.26	.21	*
Total Symptom Score								
Control	2.45 (5.83)	2.29 (5.09)	0.48	0.4	0.55	0.71	.48	*
LD	3.81 (7.48)	3.64 (6.92)	0.47	0.23	0.63	0.21	.83	*
Headache/migraine	5.14 (7.98)	5.0 (9.06)	0.74	0.6	0.83	0.16	.88	*
Two-factor composites								
Memory composite								
Control	0.08 (0.78)	0.08 (0.78)	0.67	0.62	0.71	−6.93	<.001	0.49
LD	0.16 (0.65)	0.16 (0.65)	0.73	0.6	0.81	−1.2	.23	*
Headache/migraine	0.15 (0.85)	0.15 (0.85)	0.73	0.58	0.83	−0.84	.41	*
Speed composite								
Control	−0.39 (0.76)	0.01 (0.76)	0.78	0.75	0.81	−17.54	<.001	1.25
LD	−0.52 (0.61)	−0.13 (0.9)	0.78	0.68	0.85	−5.8	<.001	1.09
Headache/migraine	−0.43 (0.94)	−0.13 (0.85)	0.76	0.63	0.85	−3.39	.001	0.76

[a]ICC, intraclass correlation coefficient.
[b]95% lower bound confidence intervals for ICC.
[c]95% upper bound confidence intervals for ICC.
[d]Paired samples t-test degrees of freedom = control (791), LD (113), and Headache/migraine (80); Bonferroni-corrected alpha, $p < .002$.
[e]Cohen's d effect size interpretations: small = 0.2, medium = 0.5, large = 0.8; Cohen (1988).

Table 5. Rates of change using reliable change indices (RCI) and regression-based measures (RBM)

Variable	RCI[a]						RBM[b]					
	90% CI[c]			95% CI[c]			90% CI[c]			95% CI[c]		
	Decline	Improve	Total	Decline	Improve	Total	Decline	Improve	Total	Decline	Improve	Total
Verbal Memory												
Control	4.4	4.8	9.2	2.8	2.7	5.3	6.0	1.8	7.8	3.7	0.5	4.2
LD	5.3	4.4	9.7	2.6	2.6	5.2	4.8	1.6	6.4	1.6	0.0	1.6
Headache/migraine	6.2	6.2	12.4	2.5	2.5	5.0	3.4	0.0	3.4	1.7	0.0	1.7
Visual Memory												
Control	6.7	4.9	11.6	3.7	2.7	6.4	5.5	2.1	7.6	3.1	0.8	3.9
LD	4.4	6.1	10.5	2.6	3.5	6.1	5.6	3.2	8.8	4.0	1.6	5.6
Headache/migraine	2.5	4.9	7.4	2.5	2.5	5.0	1.7	2.6	4.3	0.9	1.7	2.6
Visual Motor Speed												
Control	3.9	4.2	8.1	2.3	2.3	4.6	2.5	5.0	7.5	1.0	1.4	2.4
LD	5.3	5.3	10.6	4.4	3.5	7.9	5.6	4.0	9.6	2.4	2.4	4.8
Headache/migraine	7.4	2.5	8.5	3.7	1.2	4.9	2.6	1.7	8.5	0.9	0.0	0.9
Reaction Time												
Control	3.8	4.2	8.0	2.8	2.7	5.5	1.3	3.7	4.9	0.7	2.9	3.6
LD	2.6	4.4	7.0	2.6	2.6	5.2	2.4	8.1	10.5	0.0	4.8	4.8
Headache/migraine	4.9	3.7	8.6	2.5	3.7	6.2	0.0	5.1	5.1	0.0	3.4	3.4
Total Symptom Score												
Control	3.4	3.2	6.6	3.0	2.8	5.8	5.6	1.9	7.5	4.3	0.6	4.9
LD	4.4	6.1	10.5	2.6	2.6	5.2	7.5	0.0	7.5	5.6	0.0	5.6
Headache/migraine	4.9	2.5	7.4	3.7	1.2	4.9	2.5	0.6	3.1	1.9	0.0	1.9
Two-factor composite												
Memory												
Control	6.8	3.0	9.8	3.7	1.5	5.2	4.9	1.9	6.8	2.9	0.5	3.4
LD	0.9	0.3	1.1	0.2	0.3	0.5	4.0	4.0	8.0	2.4	1.6	4.0
Headache/migraine	0.5	0.5	1.0	0.3	0.4	0.7	4.3	3.7	7.7	1.7	2.6	4.3
Speed												
Control	11.3	1.6	12.9	6.2	1.2	7.3	4.7	3.8	8.5	3.1	1.6	4.7
LD	1.2	0.3	1.5	0.7	0.1	0.8	6.5	3.2	9.7	3.2	1.6	4.8
Headache/migraine	0.3	0.9	1.2	0.3	0.5	0.8	5.1	2.6	7.7	3.4	0.0	3.4

[a]RCI based on Chelune, Naugle, Luders, Sedlak, and Awad (1993) and adapted for practice effects (Iverson, 2001).
[b]Regression-based measure based on McSweeny et al. (1993).
[c]CI, Confidence interval; numbers represent percentage of participants scoring beyond cutoff values(improve or decline), the respective CI ranges, 90% (1.65) and 95% (1.96) CI.

Time 1 and Time 2 was also observed for the LD $t(113) = -5.8$, $p < .001$, $d = 1.09$ and headache/migraine treatment groups $t(80) = -3.39$, $p < .001$, $d = .76$.

The ICCs were relatively stable for each condition across the 2-year time interval. Differences in ICC values were relatively minor across each condition. The greatest degree of difference in ICC values across groups was observed for Verbal memory (0.42–0.54), with the control group yielding the higher ICC value above the headache/migraine treatment group. Verbal Memory Composites also produced the lowest levels of reliability (0.42–0.54), whereas Visual Motor Speed scores demonstrated the highest estimates of reliability and produced adequate to high ICC values (0.79–0.87; see Table 4 for the full range of ICC values for each condition). Two factor ICCs of Memory were higher for the control group (ICC = 0.67, 95% CI = 0.62, 0.71) and LD group (ICC = 0.73, 95% CI = 0.6, 0.81). Regarding the headache/migraine treatment history group, ICCs for Memory (ICC = 0.73, 95%CI = 0.58, 0.83) were higher than Verbal Memory (ICC = 0.42, 95% CI = 0.1, 0.63) and comparable to Visual Memory (ICC = 0.74, 95% CI = 0.59, 0.83). Speed composite scores also demonstrated ICC values within the adequate range (0.76–0.78) for all three groups.

RCIs for each condition were calculated for composite scores and are presented along with 90% and 95% CI ranges (Table 5). This method assumes that for composite scores to demonstrate stability, 90% and 95% of cases will fall within their respective ranges and 90% of change scores in the sample will fall within normal range. In other words, it is expected that the percentage of cases demonstrating meaningful difference between assessments will not exceed the amount of what would be expected within a healthy or uninjured sample (90% CI = 10% change and 95% CI = 5% change). Should the percentage of cases exceed the expected changes within a healthy sample, reduced stability of Time 1 scores across a 2-year interval is indicated. Most generally, the anticipated 90% of composite scores fell within the 90% CI range on follow-up baselines

assessments, with the exception of minimal percentages falling outside the anticipated 10% change range for Verbal Memory in the headache/migraine treatment group (12.4%) and Visual Memory in the control (11.6%) and LD group (10.5%). The LD group also fell slightly outside the anticipated percentages in Visual Motor Speed (10.6%) and Total Symptom Score (10.5%). Reaction Time scores fell within the expected range for all groups. Greater variation of within range scores was observed for 95% CI, ranging from 4.6% to 7.9%. RCIs for two-factor composite scores all fell within the expected range for 90% CI, with the exception of Speed composite scores for the control group (12.9%). RCIs at the 95% CI also fell within the expected range, save Memory composite scores (5.2%) and Speed composite scores (7.3%) for the control group.

Similarly, RBM were calculated for composites and examined at the 90% and 95% CIs level. The majority of composites fell within expected boundaries of the 90% CI across all conditions, with the exception of the LD group on Reaction Time (10.5%). Similar to RCIs, greater variation was observed within the 95% CI range, with outliers for the LD group on Visual Memory (5.6%) and Total Symptom Score (5.6%). Rates of bidirectional change, including both improvement and impairment, are presented within Table 5 for RCI and RBM. As with previous investigations of this nature (Brett et al., 2016; Elbin et al., 2011; Schatz, 2010), RBM, compared with RCIs, proved to be a more conservative measure of change, as fewer instances of follow-up scores fell within the impaired range or demonstrated meaningful change.

Discussion

Results supported the null hypothesis for differences in test–retest reliability estimates across groups, suggesting test–retest reliability estimates did not significantly differ among high school athletes with a history of headache/migraine treatment history or LD, as compared to controls. Results also suggested that two-factor composite scores demonstrated higher, or comparable, test–retest reliability estimates compared to the four-factor structure across all groups within the study. Regarding clinical practice, these findings imply the uniformity in test–retest serial assessment practices for all athletes within these conditions. In other words, clinicians should seek to comply with the recommended practice of obtaining a new baseline for high school athletes every 2 years (CDC, 2015; ImPACT, 2012).

Although ICCs were generally consistent across all conditions for the various ImPACT composite scores, significant variation was observed within confidence intervals between the different groups. Specifically, those with LD and a history of headache/migraine treatment tended to demonstrate a wider range of scores, even when their ICC estimates were similar to healthy controls. As an example, consider the difference between the control group and those with a history of headache/migraine treatment on the Reaction Time composite. Although relatively similar ICCs were recorded for the control group and those with a history of headache/migraine treatment, 95% CI ranges differed substantially for Reaction Time as 0.56–0.66 and 0.36–0.74, respectively, suggesting a wider variation of scores among the history of headache/migraine treatment group. Although some studies have demonstrated some utility in the use of normative data in the evaluation of SRC (Echemendia et al., 2012; Schmidt, Register-Mihalik, Mihalik, & Guskiewicz, 2012), these findings suggest the importance of the serial testing method of using athletes as their own control (baseline methodology), especially for individuals with a diagnosis of LD or history of headache/migraine treatment. Future research should seek to identify additional biopsychosocial factors in which the serial assessment method is also preferred to normative comparisons. Prior research has demonstrated the advantages of the serial testing method in the "above average athlete," ultimately decreasing the risk of false negatives when classifying athlete's performances following SRC (Schatz & Robertshaw, 2014). The advantages of the serial method over the use of normative data have also been demonstrated within a military service member population using a different computer based cognitive assessment (Automated Neuropsychological Assessment Matrix; ANAM), with higher levels of false-positives generated through normative data (Roebuck-Spencer, Vincent, Schlegel, Gilliand, 2013).

In the current study, high school athletes in the control group improved significantly in on Verbal Memory, Visual Memory, Visual Motor Speed, and Reaction Time between Time 1 and Time 2. This is not surprising given findings from previous studies, demonstrating the influence of developmental changes within this population as a source of variation in stability estimates. Specifically, developmental changes in performance on computerized cognitive test paradigms, especially between the ages of 9 and 15 (McCrory et al., 2004), result in score discrepancies comparable to changes of post-concussive impairments observed on computerized cognitive assessment in adults following injury (Iverson et al., 2003). Interestingly, the LD and headache/migraine treatment groups only improved significantly on Visual Motor Speed. Future research should investigate this trend further, examining a potential interaction effect of these conditions, development at this time period, and neurocognitive test performance (i.e., verbal memory, visual memory, and reaction time).

As a means to enhance test–retest reliability, use of the two-factor structure of ImPACT continues to show promise as a viable method (Echemendia et al., 2012; Schatz & Maerlender, 2013). As described by Schatz and Maerlender (2013), the two-factor structure contains potential for increased clinical utility, in that it "represents fewer, more tangible constructs with

better reliability, and similar to improved reliability." Conceptually, the increased reliability may be a product of reducing the number of metrics, and thus reducing inherent variance. Given the increased reliability estimates, clinicians may seek to calculate two factor scores when assessing for meaningful change, using standard deviation cutoffs. While it may not be a practical solution for all ImPACT end-users, the procedures for calculating and implementing factor scores only require calculating Z-scores using age- and gender-appropriate mean from the athlete's test score, and dividing by the commensurate standard deviation. Future studies should look to improve clinical effectiveness of two-factor composites by examining diagnostic accuracy of different cutoff scores (i.e., 1 *SD*, 1.25 *SD*, and 1.5 *SD*), as well as identifying whether the findings of increased stability in Speed Composite scores also applies to other special populations.

This study is not without limitations. Firstly, while results displayed similarities in test–retest reliability across the populations examined in the study, results should not be generalized to populations outside the high school athlete. Second, although standardized training in baseline test administration was provided to the test administrators, experimental control and uniformity of test administration procedures could not be controlled with precision. Although all assessments were administered in groups and proctored by a certified athletic trainer who had been trained in the administration of ImPACT, variation in group sizes or administration procedures may account for aspects of variation within the data. For instance, the degree of supervision, which has been demonstrated as having an effect on speed-based task performance (Visual Motor Speed and Reaction Time; Kuhn & Solomon, 2014), is a factor that was not standardized between Time 1 and Time 2 in the current study. As a result, changes in athletes' scores and test–retest reliability estimates may be due to differences in the level of supervision provided at test administration. Additionally, group size at administration, which has been associated with significant differences on all ImPACT composite scores (Moser, Schatz, Neidzwski, & Ott, 2011), is unknown in the current study. The uncertain nature of group size at Time 1 and Time 2 across settings may also have had an effect on changes in athletes' scores and test–retest reliability estimates. Third, group membership classification of athletes was based upon self-reported conditions and was not verified via medical records. As such, athletes may have been misclassified due to factors such as, "self-diagnosis," inadequate medical understanding of a particular disorder, or simply forgetting to report a condition. In this context, although athletes may be more willing to disclose symptoms via a computer (i.e., as compared to guided interviews; Elbin et al., 2016), it is possible that athletes may have been "unwilling" to acknowledge their condition. Fourth, limitations also exist within the methodology of this study, as invalid baseline scores were excluded from the analysis. Given that those with a history of LD tend to exhibit higher rates of invalid performance on ImPACT (Schatz et al., 2012), this exclusion criterion contains the possibility of eliminating participants with the highest potential for variation between assessments. In this context, including invalid baseline scores would likely increase variability and decrease reliability. However, although athletes with LD are more likely to trigger an invalidity indicator on ImPACT, their performance may be a valid representation of their true neurocognitive abilities. Thus, theoretically, "invalid" scores at two time points would be consistent as their true effort at Time 1 and Time 2, similar to a non-injured athlete who put forth a valid effort. Additionally, the application of increasingly stringent validity criteria has been found to have no effect on test–retest reliability estimates for ImPACT, as similar test–retest reliability estimates were recorded with increasingly rigorous criteria applied to the same dataset (Brett & Solomon, 2016).

In summary, findings suggested minimal differences in the test–retest reliability of baseline cognitive performance of the online ImPACT test battery over a 2-year interval in a sample of high school athletes with a history of special education/LD, previous treatment for headache/migraine, and healthy controls. RBM and RCI were highly useful in the calculation of individualized reliable change and 90% CI were more effective in classifying the expected amount of cases as "unchanged." Additionally, RBM and RCI are highly recommended when evaluating reliable change scores as meaningful in those with a history of LD or treatment for headache/migraine. Given that the two-factor structure of ImPACT demonstrated comparable or higher reliability estimates over the four-factor structure for all three groups, further research is needed to determine its clinical utility.

References

Alsalaheen, B., Stockdale, K., Pechumer, D., & Broglio, S. P. (2016). Measurement error in the Immediate Postconcussion Assessment and Cognitive Testing (ImPACT): Systematic review. *Journal of Head Trauma Rehabilitation*, *31*, 242–251.

Bauer, R. M., Iverson, G. L., Cernich, A. N., Binder, L. M., Ruff, R. M., & Naugle, R. I. (2012). Computerized neuropsychological assessment devices: Joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *Archives of Clinical Neuropsychology*, *27*, 362–373.

Brett, B. L., Smyke, N., Solomon, G. S., Baughman, B. C., & Schatz, P. (2016). Long-term stability and reliability of baseline cognitive assessments in high school athletes using ImPACT at 1, 2, and 3-year test-retest intervals. *Archives of Clinical Neuropsychology*. doi:10.1093/arclin/acw055.

Brett, B. L., & Solomon, G. S. (2016). The influence of validity criteria on ImPACT test-retest reliability among high school athletes. *Journal of Clinical and Experimental Neuropsychology*. doi:10.1080/13803395.2016.1224322.

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

Bruce, J., Echemendia, R., Meeuwisse, W., Comper, P., & Sisco, A. (2014). 1 year test-retest reliability of ImPACT in professional ice hockey players. *The Clinical Neuropsychologist*, *28*, 14–25.

Bruce, J., Echemendia, R., Tangeman, L., Meeuwisse, W., Comper, P., Hutchison, M., et al. (2016). Two baselines are better than one: Improving the reliability of computerized testing in sports neuropsychology. *Applied Neuropsychology. Adult*, *23*, 336–342.

Buzzini, S. R., & Guskiewicz, K. M. (2006). Sport-related concussion in the young athlete. *Current Opinion in Pediatrics*, *18*, 376–382.

CDC. (2015). FAQs about Baseline Testing. https://www.cdc.gov/headsup/basics/baseline_testing.html. Accessed February 3 2017.

Chelune, G. J., Naugle, R. I., Luders, H., Sedlak, J., & Awad, I. A (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology*, *7*, 41–52.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

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*, 732–742.

Covassin, T., Elbin, R. J., & Stiller-Ostrowski, J. L. (2009). Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT) practices of sports medicine professionals. *Journal of Athletic Training*, *40*, 639–644.

Dikmen, S., Heaton, R., Grant, I., & Temkin, N. (1991). Test–retest reliability and practice effects of Expanded Halstead–Reitan Neuropsychological Test Battery. *Journal of International Neuropsychological Society*, *5*, 346–356.

Duff, K. (2012). Evidence-based indicators of neuropsychological change in the individual patient: Relevant concepts and methods. *Archives of Clinical Neuropsychology*, *27*, 248–261.

Echemendia, R. J., Bruce, J. M., Bailey, C. M., Sanders, J. F., Arnett, P., & Vargas, G. (2012). The utility of post-concussion neuropsychological data in identifying cognitive change following sports-related MTBI in the absence of baseline data. *The Clinical Neuropsychologist*, *26*, 1077–1091.

Elbin, R. J., Schatz, P., & Covassin, T. (2011). One-year test-retest reliability of the online version of ImPACT in high school athletes. *American Journal of Sports Medicine*, *39*, 2319–2324.

Elbin, R. J., Kontos, A. P., Kegel, N., Johnson, E., Burkhart, S., & Schatz, P. (2013). Individual and combined effects of LD and ADHD on computerized neurocognitive concussion test performance: Evidence for separate norms. *Archives of Clinical Neuropsychology*, *8*, 476–484.

Elbin, R. J., Knox, J., Kegel, N., Schatz, P., Lowder, H., French, J., et al. (2016). Assessing symptoms in adolescents following sport-related concussion: A comparison of four different approaches. *Applied Neuropsychology: Child*, *5*, 294–302.

Gerrard, P. B., Iverson, G. L., Atkins, J. E., Maxwell, B. A., Zafonte, R., Schatz, P., et al. (2017). Factor structure of ImPACT® in adolescent student athletes. *Archives of Clinical Neuropsychology*, *32*, 117–122.

Harmon, K. G., Drezner, J. A., Gammons, M., Guskiewicz, K. M., Halstead, M., Herring, S. A., et al. (2013). American Medical Society for Sports Medicine position statement: Concussion in sport. *British Journal of Sports Medicine*, *47*, 15–26.

Immediate Post-Concussion Assessment and Cognitive Testing (ImPACT). (2012). Immediate post-concussion assessment testing (ImPACT) test: Technical manual. Retrieved from https://www.impacttest.com/pdf/ImPACTTechnicalManual.pdf

Iverson, G. L. (2001). Interpreting change on the WAIS-III/ WMS-III in clinical samples. *Archives of Clinical Neuropsychology*, *16*, 183–191.

Iverson, G. L., Franzen, M., Lovell, M. R., & Collins, M. W. (2004). Construct validity of computerized neuropsychological screening in athletes with concussion. *Archives of Clinical Neuropsychology*, *19*, 961–962.

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

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.

Kuhn, A. W., & Solomon, G. S. (2014). Supervision and computerized neurocognitive baseline testing performance in high school athletes: An initial investigation. *Journal of Athletic Training*, *49*, 800–805.

Lau, B., Lovell, M. R., Collins, M. W., & Pardini, J. (2009). Neurocognitive and symptom predictors of recovery in high school athletes. *Clinical Journal of Sport Medicine*, *19*, 216–221.

Lovell, M. R. (2007). ImPACT Version 6.0 Clinical User's Manual. Retrieved from http://www.impacttest.com/

Maerlender, A., Flashman, L., Kessler, A., Kumbhani, S., Greenwald, R., Tosteson, T., et al. (2010). Examination of the construct validity of ImPACT computerized test, traditional, and experimental neuropsychological measures. *The Clinical Neuropsychologist*, *24*, 1309–1325.

Maerlender, A., & Molfese, D. L. (2015). Repeat baseline assessment in college-age athletes. *Developmental Neuropsychology*, *40*, 69–73.

McCaffrey, R. J., Ortega, A., Orsillo, S. M., Nelles, W. B., & Haase, R. F. (1993). Practice effects in repeated neuropsychological assessments. *The Clinical Neuropsychologist*, *6*, 32–42.

McCrory, P., Meeuwisse, W., Aubry, M., Cantu, B., Dvorak, J., Echemendia, R. J., et al. (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*, 250–258.

McCrory, P., Meeuwisse, W., Dvorak, J., Aubry, M., Bailes, J., Broglio, S., et al. (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*. doi:10.1136/bjsports-2017-09769.

McCrory, P., Collie, A., Anderson, G., & Davis, G. (2004). Can we manage sport related concussion in children the same as adults? *British Journal of Sports Medicine*, *38*, 516–519.

McSweeny, A. J., Naugle, R., Chelune, G. J., Gordon, J., & Lüders, H. (1993). "T scores for change": An illustration of a regression approach to depicting change in clinical neuropsychology. *The Clinical Neuropsychologist*, *7*, 300–312.

Miller, J. R., Adamson, G. J., Pink, M. M., & Sweet, J. C. (2007). Comparison of preseason, midseason, and postseason neurocognitive scores in uninjured collegiate football players. *American Journal of Sports Medicine*, *35*, 1284–1288.

Moser, R. S., Schatz, P., Neidzwski, K., & Ott, S. (2011). Group vs. individual administration affects baseline neurocognitive test performance. *The American Journal of Sports Medicine*, *39*, 2325–2330.

Nakayama, Y., Covassin, T., Schatz, P., Nogle, S., & Kovan, J. (2014). Examination of the test-retest reliability of a computerized neurocognitive test battery. *American Journal of Sports Medicine*, *42*, 2000–2005.

Nelson, L. D., LaRoche, A. A., Pfaller, A. Y., Lerner, E. B., Hammeke, T. A., Randolph, C., et al. (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*, 24–37.

Resch, J., Driscoll, A., McCaffrey, N., Brown, C., Ferrara, M. S., Macciocchi, S., et al. (2013). ImPACT test–retest reliability: Reliably unreliable? *Journal of Athletic Training*, *48*, 506–511.

Register-Mihalik, J. K., Kontos, D. L., Guskiewicz, K. M., Mihalik, J. P., Conder, R., & Shields, E. W. (2012). Age-related differences and reliability on computerized and paper-and-pencil neurocognitive assessment batteries. *Journal of Athletic Training*, *47*, 297–305.

Roebuck-Spencer, T. M., Sun, W., Cernich, A. N., Farmer, K., & Bleiberg, J. (2007). Assessing change with the Automatic Neuropsychological Assessment Metrics (ANAM): Issues and challenges. *Archives of Clinical Neuropsychology*, *22*, 79–87.

Roebuck-Spencer, T. M., Vincent, A. S., Schlegel, R. E., & Gilliand, K. (2013). Evidence for added value of baseline testing in computer-based cognitive assessment. *Journal of Athletic Training*, *48*, 499–505.

Schatz, P. (2010). Long-term test-retest reliability of baseline cognitive assessments using ImPACT. *American Journal of Sports Medicine*, *38*, 47–53.

Schatz, P., & Ferris, C. S. (2013). One-month test-retest reliability of the ImPACT test battery. *Archives of Clinical Neuropsychology*, *28*, 499–504.

Schatz, P., Pardini, J. E., Lovell, M. R., Collins, M. W., & Podell, K. (2006). Sensitivity and specificity of the ImPACT Test Battery for concussion in athletes. *Archives of Clinical Neuropsychology*, *21*, 91–99.

Schatz, P., & Sandel, N. (2013). Sensitivity and specificity of the online version of ImPACT in high school and collegiate athletes. *American Journal of Sports Medicine*, *41*, 321–326.

Schatz, P., Moser, R. S., Solomon, G. S., Ott, S. D., & Karpf, R. (2012). Prevalence of invalid computerized baseline neurocognitive test results in high school and collegiate athletes. *Journal of Athletic Training*, *47*, 289–296.

Schatz, P., & Maerlender, A. (2013). A two-factor theory for concussion assessment using ImPACT: Memory and speed. *Archives of Clinical Neuropsychology*, *28*, 791–797.

Schatz, P., & Robertshaw, S. (2014). Comparing post-concussive neurocognitive test data to normative data presents risks for under-classifying "above average" athletes. *Archives of Clinical Neuropsychology*, *29*, 625–632.

Schmidt, J. D., Register-Mihalik, J. K., Mihalik, J. P., & Guskiewicz, K. M. (2012). Identifying impairments after concussion: Normative data versus individualized baselines. *Medicine & Science in Sports & Exercise*, *44*, 1621–1628.

Slick, D. (2006). Psychometrics in neuropsychological assessment. In E. Strauss, E. Sherman, & O. Spreen (Eds.), *A Compendium of Neuropsychological Tests* (3rd ed., pp. 1–43). New York: Oxford University Press.

Solomon, G. S., & Haase, R. F. (2008). Biopsychosocial characteristics and neurocognitive test performance in National Football League players: An initial assessment. *Archives of Clinical Neuropsychology*, *23*, 563–577.

Tator, C. H., Davis, H. S., Dufort, P. A., Tartagilla, M. C., Davis, K. D., Ebraheem, A., et al. (2016). Postconcussion syndrome: Demographics and predictors in 221 patients. *Journal of Neurosurgery*, *125*, 1206–1216.

Vaz, S., Falkmer, T., Passmore, A. E., Parsons, R., & Andreou, P. (2013). The case for using the repeatability coefficient when calculating test-retest reliability. *PLoS ONE*, *8*, e73990.

Wojtowicz, M., Iverson, G. L., Silverberg, N. D., Mannix, R., Zafonte, R., Maxwell, B., et al. (2017a). Consistency of self-reported concussion history in adolescent athletes. *Journal of Neurotrauma*, *34*, 322–327. doi:10.1089/neu.2016.4412.

Wojtowicz, M., Terry, D., Zafonte, R., Berkner, P. D., Seifert, T., Iverson, G. L. (2017b) *Migraine history and associated comorbidities in middle school athletes*. Poster session presented at the American Academy of Neurology Boston, MA.

Womble, M. N., Reynolds, E., Schatz, P., Shah, K. M., & Kontos, A. P. (2016). Test-retest reliability of computerized neurocognitive testing in youth ice hockey players. *Archives of Clinical Neuropsychology*, *31*, 305–312.

Zuckerman, S. L., Lee, Y. M., Odom, M. J., Solomon, G. S., & Sills, A. K. (2013). Baseline neurocognitive scores in athletes with attention-deficit spectrum disorders and/ or learning disability. *Journal of Neurosurgery. Pediatrics*, *12*, 103–109.