Archives of Clinical Neuropsychology Advance Access published August 18, 2016

OXFORD
UNIVERSITY PRESS

Archives
of
CLINICAL
NEUROPSYCHOLOGY

# Long-term Stability and Reliability of Baseline Cognitive Assessments in High School Athletes Using ImPACT at 1-, 2-, and 3-year Test–Retest Intervals

Benjamin L. Brett[1,2,*], Nathan Smyk[3], Gary Solomon[4,5], Brandon C. Baughman[1,6], Philip Schatz[3]

[1]*Semmes Murphey Neurologic and Spine Institute, Memphis TN, USA*
[2]*Department of Counseling, Educational Psychology and Research, The University of Memphis, Memphis TN, USA*
[3]*Department of Psychology, Saint Joseph's University, Philadelphia, PA, USA*
[4]*Department of Neurological Surgery, Vanderbilt University School of Medicine, Nashville, TN, USA*
[5]*Vanderbilt Sports Concussion Center, Vanderbilt University School of Medicine, Nashville, TN, USA*
[6]*Department of Neurosurgery, University of Tennessee Health Science Center, Memphis, TN, USA*

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

Accepted 19 June 2016

**Abstract**

**Objective:** The ImPACT (Immediate Post-Concussion Assessment and Cognitive Testing) neurocognitive testing battery is a widely used tool used for the assessment and management of sports-related concussion. Research on the stability of ImPACT in high school athletes at a 1- and 2-year intervals have been inconsistent, requiring further investigation. We documented 1-, 2-, and 3-year test–retest reliability of repeated ImPACT baseline assessments in a sample of high school athletes, using multiple statistical methods for examining stability.
**Methods:** A total of 1,510 high school athletes completed baseline cognitive testing using online ImPACT test battery at three time periods of approximately 1- ($N = 250$), 2- ($N = 1146$), and 3-year ($N = 114$) intervals. No participant sustained a concussion between assessments.
**Results:** Intraclass correlation coefficients (ICCs) ranged in composite scores from 0.36 to 0.90 and showed little change as intervals between assessments increased. Reliable change indices and regression-based measures (RBMs) examining the test–retest stability demonstrated a lack of significant change in composite scores across the various time intervals, with very few cases (0%–6%) falling outside of 95% confidence intervals.
**Conclusion:** The results suggest ImPACT composites scores remain considerably stability across 1-, 2-, and 3-year test–retest intervals in high school athletes, when considering both ICCs and RBM. Annually ascertaining baseline scores continues to be optimal for ensuring accurate and individualized management of injury for concussed athletes. For instances in which more recent baselines are not available (1–2 years), clinicians should seek to utilize more conservative range estimates in determining the presence of clinically meaningful change in cognitive performance.

*Keywords:* Head injury; Traumatic brain injury; Practice effects/reliable change; Assessment; Childhood brain insult; Norms/normative studies; Test construction

## Introduction

Neuropsychological testing is accepted as a major component in a multifaceted approach towards effective concussion assessment and management (Aubry et al., 2002; Guskiewicz et al., 2004; McCrory, et al., 2009). Such testing methods are often used by sports medicine professionals in the evaluation of concussion, as well as in making return-to-play decisions (American Academy of Neurology, 1997; Covassin, Elbin & Stiller-Ostrowski, 2009), which, given the ethical considerations, can be a challenging undertaking for clinicians (Kirschen, Tsou, Nelson, Russell & Larriviere, 2014). In order to accurately determine whether significant changes in cognitive functioning have actually occurred, performance on neuropsychological


doi:10.1093/arclin/acw055

assessment following a concussive injury is often compared to the athlete's "baseline" performance. This common practice of using an athlete as his/her own control when measuring change in scores from baseline, known as serial assessment, is currently regarded as best practice in the effective management of sports concussion (Covassin et al., 2009; Helibronner et al., 2010; Lovell, Collins, Fu, Burke & Podell, 2001).

According to the American Medical Society for Sports Medicine (2013), inclusion of standardized assessment tools provides a helpful structure for the evaluation of concussion. Furthermore, tracking the variety of symptoms associated with concussion through serial evaluations is also strongly recommended (Harmon et al., 2013). However, when scores or performance on baselines measures are inaccurate representations of an athlete's true neurocognitive abilities, changes detected following head injury may be unreliable and may not necessarily indicate difference in cognitive functioning (Lichtenstein, Moser & Schatz, 2014). Inaccurate representations of an athlete's true neurocognitive abilities during baseline testing can occur for a multitude of reasons, including sandbagging (intention to perform poorly), confusion, or misidentification in some aspect of the test, distractions in group test performance environments, and poor sleep the previous night (Iverson & Schatz, 2015; McClure, Zuckerman, Kutscher, Gregory & Solomon, 2014). For these instances in which obtaining suboptimal scores on a baseline measure are recorded, decreased sensitivity in detecting declines within the postinjury assessment is possible, diminishing the effectiveness of the serial method when assessing cognitive impairment following concussion.

Among the most commonly used tools for measuring neurocognitive functioning in the assessment and management of sport-related concussion in athletes is the Immediate Post-Concussion Assessment and Cognitive Testing (Covassin et al., 2009; ImPACT, 2012). As a well-validated measure (Allen & Gfeller, 2011; Maerlender et al., 2010), ImPACT minimizes practice effects through the use of alternate forms and is widely used in the clinical management of concussion in athletes across various points in the developmental spectrum (e.g., youth, adolescent, high school, college, professional). Furthermore, ImPACT has demonstrated effective diagnostic utility for concussion with sensitivity of 91.4% and specificity of 69.1%, ultimately aiding medical providers with valuable information regarding return-to-play decisions (Schatz & Sandel, 2013).

In order to determine whether neurocognitive and neurobehavioral changes occur through serial testing, reliability of the assessment must be established in order to accurately attribute any observed changes as effects of head injury, rather than extraneous or confounding factors. With time between administrations as one of the greatest confounds in test reliability, ImPACT has demonstrated a stable range of test–retest reliability across various intervals, ranging from 7 days (Iverson, Lovell & Collins, 2003), 1 month (Schatz, Pardini, Lovell, Collins & Podell, 2006; Schatz & Ferris, 2013), 45–50 days (Nakayama, Covassin, Schatz, Nogle & Kovan, 2014), 1 year (Bruce, Echemendia, Meeuwisse, Comper & Sisco, 2014; Elbin, Schatz & Covassin, 2011; Miller Adamson, Pink & Sweet, 2007), 2 years (Schatz, 2010), and even possibly up to 3 years (cautiously interpreted due to the highly extensive exclusion criteria of the study, omission of ICCs and RCI as part of the analysis, and an aggregate analysis for the entire sample as a collective group, rather than assessing for significant change for each athlete at each year difference; Maerlender & Molfese, 2015). In contrast, other investigations of ImPACT have yielded contradictory results, demonstrating lower bound estimates of test–retest reliability for ImPACT composite scores and symptom scale (Broglio, Ferrara, Macciocchi, Baumgartner & Elliot, 2007; Cole et al., 2013; Resch et al., 2013).

Contradictory findings regarding the stability of ImPACT are especially prevalent within the high school population. Reliability of ImPACT over a 1-year test–retest interval has previously been examined for high school athletes through one particular study (Elbin et al., 2011), which yielded intraclass correlation coefficients (ICCs; Chicchetti, 1994) of composite scores as follows: Verbal Memory (0.62), Visual Memory (0.70), Visual Motor Speed (0.85), Reaction Time (0.76), and Symptom Scale (0.57). Based upon a previously established reliability classification system (≥0.90 = very high; 0.80–0.89 = high; 0.70–0.79 = adequate; 0.60–0.69 = marginal; <0.60 = low; Slick, 2006), these 1-year reliability estimates were somewhat variable and borderline respectable. The use of Pearson's correlations, as well as average measures ICCs have been criticized (Alsalaheen, Stockdale, Pechumer & Broglio, 2015), whereas the use of reliable change indices (RCIs; Iverson, Lovell & Collins, 2003) and regression-based measures (RBM; Barr, 2002) have been recommended. Suitable test–retest reliability within the same study was also demonstrated through RBM (McSweeny, Naugle, Chelune, Gordon & Lüders, 1993) using 80% and 95% confidence intervals (CIs), with 88%–91% of follow-up baseline scores within the 80% range and only a minimal number of scores from Visual Memory (1.0%) and Reaction Time (0.2%) falling outside the 95% CI cutoff (Elbin et al., 2011). Similarly, RCI (Iverson, 2001; Jacobson & Truax, 1991) also demonstrated considerable stability, with a limited percentage of cases falling outside the cutoffs associated with 80% and 95% CIs. Conversely, a recent investigation by Tsushima et al. (2016) examining ImPACT stability over a 2-year interval in a sample of high school athletes demonstrated contradictory findings to that of Elbin et al. (2011) with significantly lower ICC for Verbal Memory (0.21), Visual Memory (0.49), Visual Motor Speed (0.72), Reaction Time (0.46), and Symptom Scale (0.40). RCI and RBM using 80% and 95% CI revealed similar and generally acceptable reliability among the five composite scores.

Psychometric development has been identified as a key issue in the use of computerized neuropsychological assessment devices (Bauer et al., 2012), and researchers have recommended caution when using baseline neurocognitive testing as a

comparative criterion for return-to-play decision-making, primarily due to marginal test–retest reliability (Alsalaheen et al., 2015). Although some researchers have recommended the use of normative comparisons (for post-concussion data; Echemendia et al., 2012), others have been argued that normative comparisons may differentially classify concussed, symptomatic athletes who fall outside the "average" range (Schatz & Robertshaw, 2014). The use of multivariate base rates has been proposed, using a criterion of two or more ImPACT composite scores in the "impaired range" (using RCI; Iverson & Schatz, 2015).

Given these contradictory findings, further clarification into the stability of ImPACT within high school athletes, most especially at different time intervals is necessary. The aim of this study was to examine and obtain further clarification regarding previously demonstrated inconsistencies of test–retest reliability in a sample of high school athletes. Given that differences in neurocognitive functioning and performance can occur at this time period due to developmental changes, examining the stability of ImPACT baseline scores in a population sample of high school athletes is imperative and of value (Choudhury, Blakemore & Charman, 2006; Crone & Elzinga, 2015). Furthermore, this study aims to investigate the reliability of baseline ImPACT scores beyond 2 years, as well as the stability of scores across three interval levels or groups of high school athletes, examining meaningful change at 1, 2, and 3 years between baselines.

## Methods

### *Participants*

Participants were English-speaking high school athletes, aged 13–18 years, enrolled in 30 high schools in a southern region of the US who participated in multiple sports and were supported by a regional neurocognitive sports medicine testing program (Table 1). 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, and these data have not been presented previously in any ImPACT reliability studies. As part of standard program protocol, data were acquired through preseason cognitive baseline assessments by the athletic departments at each high school from 2010 to 2015. Athletes in the sample completed the online version of ImPACT, Version 2.1. Baseline tests were administered in group settings and proctored by a certified athletic trainer who had been trained in the administration of ImPACT. Due to the multisite nature of the data, identification of group size was not possible. Athletes who sustained concussion during the test–retest interval were excluded. Additional exclusion criteria for the study consisted of those who were not clearly indicated as high school athletes in the database (15; <0.01%) and non-native English speakers or the presence of a language discrepancy between test language and primary language (23; 0.01%). Based on the criteria for invalid baselines provided by ImPACT (2012), 11 (<0.01%) invalid baseline scores at Time 1, and 34 (0.02%) invalid baselines scores at Time 2 were removed from the final analyses. Additionally, 3 (<0. 01%) athletes were also removed if their second baseline was obtained more than 3 years after their initial

**Table 1.** Demographics

| | *n* (%) | | *n* (%) |
|---|---|---|---|
| **Sex** | | **Sport** | |
| Male | 988 (65.4) | Football | 605 (40.1) |
| Female | 522 (34.6) | Soccer | 311 (20.6) |
| Total | 1510 | Volleyball | 135 (8.9) |
| Race | | Lacrosse | 104 (6.9) |
| Caucasian | 1162 (77.0) | Basketball | 103 (6.8) |
| Black or African American | 202 (13.4) | Baseball | 92 (6.1) |
| Hispanic or Latino | 33 (2.2) | Cheerleading | 47 (3.1) |
| Biracial | 19 (1.3) | Wrestling | 43 (2.8) |
| Asian | 18 (1.2) | Softball | 38 (2.5) |
| | | X-Country | 13 (0.9) |
| American Indian or Alaskan Native | 12 (0.8) | Tennis | 9 (0.6) |
| | | Track & Field | 4 (0.3) |
| Native Hawaiian or Pacific Islander | 3 (0.2) | Rugby | 3 (0.2) |
| Undeclared | 61 (4.0) | Golf | 2 (0.1) |
| **Age**[a] | 15.1 ± 1.9 | Road Biking | 1 (0.1) |
| **Diagnosis** | | **Time between BL (months)**[a] | |
| ADHD | 102 (6.8) | 1-year interval | 11.6 ± 2.1 |
| Dyslexia | 26 (1.7) | 2-year interval | 23.2 ± 1.2 |
| Autism | 3 (0.2) | 3-year interval | 33.8 ± 2.5 |

[a]Data presented as mean and standard deviation.

baseline. In total, 86 participants (0.06%) were excluded from the study. The final sample was composed of 1510 athletes who completed baseline assessments at an initial session (mean age = 15.11, *SD* = 1.85) and a second baseline later in their high school tenure (mean age = 17.00, *SD* = 1.86).

Repeat baselines were classified into Groups 1, 2 and 3, as interval times between baselines were 1 year, 2 years, and 3 years, respectively. Intervals of 1, 2, and 3 years were selected as measurement points in accordance with typical practices that involve test administration at the beginning of the year on an annual basis. Those in Group 1 (*N* = 250) completed repeat baselines between 6 and 18 months (mean interval in months = 11.6, *SD* = 2.11), Group 2 (*N* = 1146) between 19 and 30 months (mean interval in months = 23.21, *SD* = 1.18), and Group 3 (*N* = 114) between 31 and 42 months (mean interval in months = 33.83, *SD* = 2.489). Gender, racial, sport participation, age, and additional diagnostic distribution for this study are provided in Table 1.

### *Materials*

ImPACT is a computer-based program used to assess neurocognitive function and concussion symptoms. It consists of six individual subtests that yield composite scores in the areas of Verbal Memory, Visual Memory, Visual Motor Speed, Reaction Time, and Impulse Control (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). Additionally, there is a self-report total symptom scale, comprised 22 common symptoms, each rated on a 0–6 Likert scale, with 0 = none and 6 = severe, reported by the athlete.

### *Measures of Test–Retest Reliability*

Pearson product moment correlations (*r*): As a general measure of linear association, Pearson product moment correlations (*r*) provide indications as to the strength of relationship of outcome variables at Times 1 and 2. However, limitations are present with Pearson's *r*, as it is simply a measure of general correlation and contains an inherent weakness as a measure of test–retest reliability (Vaz, Falkmer, Passmore, Parsons & Andreou, 2013).

*ICCs and* t*-tests.* Given this limitation of Pearson's *r*, the use of "two-way mixed" type "consistency" ICCs are more advantageous and considered a better measure of association for investigations of test–retest, due to its ability to provide an unbiased estimate of reliability based on the consistency of baseline assessments from test to retest within subjects, but also in average performance of all participants as a group (Vaz et al., 2013). Although effective for several other forms of inquiry, paired samples *t*-tests do not possess optimal utility in the assessment of test–retest reliability, especially in the domain of neuropsychological testing, as they are unable to account for regression to the mean and relevant covariance among the two assessments (Barr, 2002; Bonate, 2000).

*Reliable change indices.* The use of RCIs have been utilized in the evaluation of change when attempting to determine whether a change between repeated assessments is reliable and meaningful and has been established as applicable within computerized measures of concussion (Duff, 2012; Parsons, Notebaert, Shields & Guskiewicz, 2009). Through the calculation of an estimate of measurement error within the test–retest difference scores, the RCI determines whether changes in test scores are due to reliable change or individually inherent factors (Jacobson & Truax, 1991). Recently, this method has been demonstrated to be efficacious in assessing for reliable change between test–retest scores with ImPACT in identifying/classifying meaningful change, ultimately assisting in the detection and management of sports concussion (Iverson, Lovell & Collins, 2003). Additionally, the detection of meaningful change within serial assessment can be enhanced further through the use of formulas intended to account for practice effects due to repetitive exposure to a particular assessment (Chelune, Naugle, Lüders, Sedlak & Awad, 1993; Collie, Maruff, Darby & McStephen, 2009).

*Regression-based measures.* Finally, as a more individualized approach, regression-based measures (RBMs) are typically used in order to assess meaningful change within test–retest differences (McSweeny et al., 1993). RBM incorporates scores from the initial assessment in order to better predict a participant's level of performance on a neuropsychological instrument at retest and examines meaningful change based on differences between a predicted Time 2 score and their actual performance. Thus, RBM helps control for regression to the mean, and adjustment of score difference is made based upon initial performance (Duff, 2012; Iverson et al., 2003).

*Procedures and Data Analyses*

A within-subjects design, consisting of participants completing two baseline assessments, allowed for a comparison between Time 1 and Time 2 baseline scores. The dependent measures for this study included the ImPACT composite scores (Verbal Memory, Visual Memory, Visual Motor Speed, Reaction Time, and Impulse Control), as well as the total symptom score described earlier. In order to assess general measure of linear association and correlation, Pearson *r* and ICCs were calculated in order to evaluate the strength of relationship of composite scores/symptom scores at Times 1 and 2. Following the methodology of Maerlender and Molfese (2015), paired samples *t*-tests were applied to the data as an attempt to replicate previous findings cited earlier. Bonferroni correction for multiple comparisons resulted in a required alpha level of 0.003. RCIs, including adjustments for practice effects, were calculated to assess whether changes between repeated baseline assessments represented meaningful change. Additionally, RBMs were also used in order to assess whether participants performance on repeat assessments meaningfully deviated from predicted scores based on initial baseline testing scores.

## Results

Minimal variation was observed in mean ImPACT composite and symptom scale scores of two baselines assessments across all three time intervals of testing (Tables 2–5). Improvement was observed on Time 2 baseline assessment for all composite scores, at all three time intervals. Practice effects can also be considered as a potential source of improvement at repeat testing, as repeat test exposure can result in score increased due to learned strategies, memory for test items, and test sophistication (Calamia, Markon & Tranel, 2012). Paired samples *t*-test demonstrated significant increases in performance across 2-year intervals for Verbal Memory ($p < .001$), Visual Memory ($p < .001$), and Reaction Time ($p < .001$) composite scores. Additionally, significant increases were observed in Visual Motor Speed scores at intervals of 1 ($p < .001$), 2 ($p < .001$), and 3 years ($p < .001$). Given the limitations of *t*-tests in examining test–retest reliability (Bonate, 2000), as well as the increased probability of obtaining significant results with larger sample sizes (Sullivan & Feinn, 2012), the meaning and magnitude of these results should be interpreted with extreme caution, if at all. This is further supported by the relatively minor effect sizes produced through each paired samples *t*-test, as the majority of *t*-tests yielded Cohen's *d* values that fall within the small effect range, save Visual Motor Speed Composite scores, which produced large effect sizes for 1–3 year intervals ($d = 1.08$, $d = 1.43$, and $d = 2.09$, respectively). Observed improvement in visual motor speed, as demonstrated by high effect sizes at all three different time intervals may reflect normal psychomotor development (Cioni & Sgandurra, 2013).

**Table 2.** Test–retest reliability[a]

| Variable (years) | Time 1 | Time 2 | *r* | ICC | 95% CI Lower | 95% CI Upper | *t*[b] | Sig. | *d*[c] |
|---|---|---|---|---|---|---|---|---|---|
| **Verbal Memory** | *M* (*SD*) | *M* (*SD*) | | | | | | | |
| 1 | 84.0 (9.8) | 85.4 (10.2) | 0.30 | 0.464 | 0.313 | 0.582 | −1.95 | 0.052 | 0.25 |
| 2 | 84.6 (9.6) | 86.0 (9.9) | 0.36 | 0.528 | 0.47 | 0.58 | −4.43 | 0.001 | 0.26 |
| 3 | 84.7 (9.5) | 84.8 (11.2) | 0.21 | 0.358 | 0.069 | 0.557 | −0.12 | 0.90 | 0.023 |
| **Visual Memory** | | | | | | | | | |
| 1 | 74.6 (12.6) | 76.4 (14.6) | 0.50 | 0.665 | 0.57 | 0.739 | −2.21 | 0.028 | 0.28 |
| 2 | 74.7 (12.5) | 77.2 (12.7) | 0.50 | 0.664 | 0.623 | 0.701 | −6.64 | 0.000 | 0.39 |
| 3 | 73.9 (13.0) | 77.0 (15.2) | 0.47 | 0.631 | 0.465 | 0.745 | −2.28 | 0.025 | 0.42 |
| **Visual Motor Speed** | | | | | | | | | |
| 1 | 36.1 (6.2) | 38.5 (6.2) | 0.76 | 0.865 | 0.827 | 0.895 | −8.50 | 0.001 | 1.08 |
| 2 | 35.1 (6.2) | 38.6 (6.3) | 0.71 | 0.828 | 0.807 | 0.847 | −24.20 | 0.001 | 1.43 |
| 3 | 34.3 (6.8) | 38.8 (7.1) | 0.81 | 0.896 | 0.85 | 0.929 | −11.11 | 0.001 | 2.09 |
| **Reaction Time** | | | | | | | | | |
| 1 | 0.63 (0.08) | 0.61 (0.07) | 0.51 | 0.670 | 0.576 | 0.742 | 2.83 | 0.005 | 0.36 |
| 2 | 0.63 (0.09) | 0.60 (.08) | 0.42 | 0.585 | 0.534 | 0.63 | 9.31 | 0.001 | 0.55 |
| 3 | 0.63 (.092) | 0.61 (.09) | 0.54 | 0.703 | 0.57 | 0.795 | 2.20 | 0.030 | 0.41 |
| **Symptom Scale** | | | | | | | | | |
| 1 | 3.3 (7.7) | 3.6 (7.4) | 0.55 | 0.708 | 0.626 | 0.773 | −0.49 | 0.62 | 0.062 |
| 2 | 3.1 (6.8) | 2.8 (6.1) | 0.36 | 0.527 | 0.469 | 5.79 | 1.55 | 0.12 | 0.091 |
| 3 | 2.7 (5.6) | 2.9 (5.7) | 0.39 | 0.560 | 0.363 | 0.696 | −0.33 | 0.74 | 0.062 |

[a]ICC, intraclass correlation coefficient; CI, confidence interval; Sig, significance (*p*).
[b]*df* = 1 year (249), 2 years (1145), and 3 years (113); Bonferroni-corrected alpha, $p < .003$.
[c]Cohen's *d* effect size ranges; small ($d = 0.2$), medium ($d = 0.5$), and large ($d \geq 0.8$).

Pearson's *r* correlations and ICCs were relatively stable for each retest time interval and as expected, ICCs produced higher estimates of reliability (Baumgartner & Chung, 2001). Verbal Memory Composite scores produced the lowest levels of stability, Visual Motor Processing Speed scores were consistently the most stable, and other composite score ICCs tended to demonstrate stability of 0.59–0.70 across all 3 years (see Table 2 for the full range of *t*-test, Pearson *r,* ICC values across all three time intervals).

RCIs were calculated for composite and total symptom scores, which are presented along with 80%, 90%, and 95% CIs (Table 3). This method assumes that for composite and symptom scores to demonstrate stability, 90% and 95% of cases will fall within their respective ranges. In other words, lower than anticipated percentages indicate decreased reliability of a Time 2 assessment. Most generally, the anticipated 90% of composite scores fell within the 90% CI on follow-up baselines assessments (Table 4), with the exception of minimal percentages falling outside the anticipated range for Verbal Memory Scores at 2 years (0.3%), Visual Memory scores at 3 years (0.5%), Reaction Time at 1 year (0.4%), and Total Symptoms at 3 years (0.5%), which fell outside the cutoff. Greater variation of within range scores was observed for 95% CI; however, the degree of outlier values are minimal (0.2%–1.6%).

Similarly, RBMs were calculated for composite and symptom scores and examined at the 90% and 95% CIs level (Table 5). The majority of composite and symptom scores across all three time intervals fell within expected boundaries of the 90% CI. In a small number of instances, scores of fell outside the expected CIs at Year 3, which include Visual Motor Speed (3.1%) and total symptom scores (0.5%). Similar to RCIs, greater variation was observed within the 95% CI, with lower percentages of outliers for each individual composite score exceeding the acceptable cutoff (Table 4). Rates of bidirectional change, including both improvement and impairment, are presented within Table 4 for RCI and RBM. As with previous investigations of this nature (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. Overall, ImPACT composite and symptom scores displayed minimal meaningful change through the use of RCI and RBM.

## Discussion

This study was conducted in an attempt to further investigate the test–retest reliability of baseline cognitive performance of the online ImPACT test battery in a sample of high school athletes at 1-, 2-, and 3-year test intervals. Findings from the

**Table 3.** RCIs: 1–3 year intervals[a]

| Variable (years) | *r* | SE[1] | SE[2] | Sdiff[b] | 80% CI[c,d] | 90% CI[c,d] | 95% CI[c,d] |
|---|---|---|---|---|---|---|---|
| **Verbal Memory** | | | | | | | |
| 1 | 0.30 | 8.19 | 8.53 | 11.83 | 15.10 | 19.40 | 23.20 |
| 2 | 0.36 | 7.65 | 7.91 | 11.01 | 14.10 | 18.10 | 21.60 |
| 3 | 0.21 | 8.36 | 9.85 | 12.92 | 16.50 | 21.20 | 25.30 |
| **Visual Memory** | | | | | | | |
| 1 | 0.50 | 8.91 | 9.65 | 13.13 | 16.80 | 21.50 | 25.70 |
| 2 | 0.50 | 8.83 | 9.02 | 12.62 | 16.10 | 20.60 | 24.70 |
| 3 | 0.47 | 9.50 | 11.09 | 14.61 | 18.70 | 23.90 | 28.60 |
| **Visual Motor Speed** | | | | | | | |
| 1 | 0.76 | 3.01 | 3.23 | 4.41 | 5.60 | 7.20 | 8.60 |
| 2 | 0.71 | 3.37 | 3.39 | 4.78 | 6.10 | 7.80 | 9.40 |
| 3 | 0.81 | 2.93 | 3.06 | 4.24 | 5.40 | 7.00 | 8.30 |
| **Reaction Time** | | | | | | | |
| 1 | 0.51 | 0.057 | 0.052 | 0.077 | 0.099 | 0.127 | 0.152 |
| 2 | 0.42 | 0.066 | 0.060 | 0.080 | 0.103 | 0.132 | 0.157 |
| 3 | 0.54 | 0.062 | 0.058 | 0.085 | 0.108 | 0.139 | 0.166 |
| **Impulse control** | | | | | | | |
| 1 | 0.52 | 3.27 | 2.86 | 4.34 | 5.60 | 7.10 | 8.50 |
| 2 | 0.40 | 3.61 | 3.36 | 4.93 | 6.30 | 8.10 | 9.70 |
| 3 | 0.42 | 3.72 | 3.74 | 5.27 | 6.70 | 8.60 | 10.30 |
| **Symptom Scale** | | | | | | | |
| 1 | 0.55 | 5.20 | 4.97 | 7.19 | 9.20 | 11.80 | 14.10 |
| 2 | 0.36 | 5.40 | 4.87 | 7.27 | 9.30 | 11.90 | 14.20 |
| 3 | 0.39 | 4.39 | 4.43 | 6.24 | 8.00 | 10.20 | 12.23 |

[a]*r*, Pearson correlation between Time 1 and Time 2 scores; *SE*, standard error of measure at Time 1 ($SE^2$) and Time 2 ($SE^2$) ($SD \times \sqrt{[1-r_{xy}]}$); *SD*, standard deviation.

[b]Sdiff, standard error of difference scores (Iverson, 2001): $\sqrt{([SEM1^2]+[SEM2^2])}$.

[c]Reliable change index based on Chelune et al. (1993).

[d]CI, confidence interval; numbers represent reliable change scores at 80% (1.28), 90% (1.65), and 95% (1.96) CIs.

**Table 4.** Rates of impairment using RCIs versus RBMs[a]

| Variable (years) | RCI[b] | | | | | | RBM | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 90% CI[c] | | | 95% CI[c] | | | 90% CI[c] | | | 95% CI[c] | | |
| | Impr | Decl | Tot | Impr | Decl | Tot | Impr | Decl | Tot | Impr | Decl | Tot |
| **Verbal Memory** | | | | | | | | | | | | |
| 1 | 4.8 | 4.8 | 9.6 | 4.4 | 2.0 | 6.2 | 2.4 | 5.6 | 8.0 | 0.4 | 4.0 | 4.4 |
| 2 | 5.2 | 5.1 | 10.3 | 2.5 | 2.8 | 5.3 | 2.0 | 6.9 | 8.9 | 0.3 | 3.8 | 4.1 |
| 3 | 7.0 | 2.6 | 9.6 | 4.4 | 1.8 | 6.2 | 0.9 | 5.3 | 6.2 | 0.0 | 3.5 | 3.5 |
| **Visual Memory** | | | | | | | | | | | | |
| 1 | 4.4 | 5.6 | 10.0 | 3.2 | 2.8 | 6.0 | 3.2 | 6.4 | 9.6 | 1.2 | 4.4 | 5.6 |
| 2 | 4.8 | 5.1 | 9.9 | 2.4 | 3.0 | 5.4 | 3.1 | 5.9 | 9.0 | 1.4 | 3.7 | 5.1 |
| 3 | 3.5 | 7.0 | 10.5 | 3.5 | 3.5 | 7.0 | 1.8 | 7.9 | 9.7 | 1.8 | 3.5 | 5.3 |
| **Visual Motor Speed** | | | | | | | | | | | | |
| 1 | 4.8 | 4.8 | 9.6 | 1.6 | 4.4 | 5.8 | 5.2 | 4.0 | 9.2 | 1.2 | 4.4 | 5.6 |
| 2 | 4.8 | 4.3 | 9.1 | 2.4 | 2.4 | 4.8 | 4.6 | 3.8 | 8.4 | 2.2 | 1.9 | 4.1 |
| 3 | 4.4 | 4.4 | 8.8 | 3.5 | 1.8 | 5.3 | 7.0 | 6.1 | 13.1 | 3.5 | 1.8 | 5.3 |
| **Reaction Time** | | | | | | | | | | | | |
| 1 | 4.8 | 5.6 | 10.4 | 2.8 | 2.4 | 5.2 | 6.8 | 1.6 | 8.4 | 4.8 | 0.8 | 5.6 |
| 2 | 4.5 | 3.9 | 9.4 | 3.0 | 3.3 | 6.6 | 6.2 | 1.7 | 7.9 | 4.5 | 0.6 | 5.1 |
| 3 | 3.5 | 4.4 | 7.9 | 3.5 | 1.8 | 5.3 | 6.1 | 0.0 | 6.1 | 5.3 | 0.0 | 5.3 |
| **Impulse Control** | | | | | | | | | | | | |
| 1 | 4.4 | 5.6 | 10.0 | 1.6 | 3.6 | 5.2 | 8.4 | 2.0 | 10.2 | 0.0 | 4.4 | 4.4 |
| 2 | 4.0 | 5.4 | 9.4 | 2.8 | 2.7 | 5.5 | 6.5 | 1.1 | 7.6 | 4.4 | 0.6 | 5.0 |
| 3 | 4.4 | 3.5 | 7.9 | 3.5 | 3.5 | 7.0 | 4.4 | 0.0 | 4.4 | 0.0 | 3.5 | 3.5 |
| **Symptom Scale** | | | | | | | | | | | | |
| 1 | 3.6 | 3.6 | 7.2 | 2.8 | 3.2 | 6.0 | 4.8 | 2.0 | 6.8 | 3.6 | 0.8 | 4.4 |
| 2 | 3.6 | 4.2 | 7.8 | 3.1 | 3.2 | 6.3 | 5.3 | 0.6 | 5.9 | 4.7 | 0.3 | 5.0 |
| 3 | 7.0 | 3.5 | 10.5 | 3.5 | 2.6 | 6.1 | 9.6 | 0.9 | 10.5 | 8.8 | 0.9 | 9.7 |

[a]Impr, improved; Decl, declined; Tot, total.
[b]RCI, (Chelune et al., 1993).
[c]CI, confidence interval; numbers represent percent of participants scoring beyond cutoff values, 90% (1.65) and 95% (1.96) CI.

**Table 5.** Regression-based method: 1-, 2-, and 3- year intervals[a]

| Variable (years) | Time 1 | Time 2 | $\alpha$ | $\beta$ | Sxy | 90% CI[b] | 95% CI[b] |
|---|---|---|---|---|---|---|---|
| **Verbal Memory** | *M* (*SD*) | *M* (*SD*) | | | | | |
| 1 | 84.0 (9.8) | 85.4 (10.2) | 58.98 | 0.315 | 9.749 | 92.0 | 95.6 |
| 2 | 84.6 (9.6) | 86.0 (9.9) | 54.61 | 0.317 | 9.223 | 91.1 | 95.9 |
| 3 | 84.7 (9.5) | 84.8 (11.2) | 63.81 | 0.26 | 10.933 | 93.9 | 96.5 |
| **Visual Memory** | | | | | | | |
| 1 | 74.6 (12.6) | 76.4 (14.6) | 36.05 | 0.541 | 11.845 | 90.4 | 94.4 |
| 2 | 74.7 (12.5) | 77.2 (12.7) | 39.22 | 0.508 | 11.032 | 91.0 | 94.9 |
| 3 | 73.9 (13.0) | 77.0 (15.2) | 36.83 | 0.544 | 13.487 | 90.4 | 94.7 |
| **Visual Motor Speed** | | | | | | | |
| 1 | 36.1 (6.2) | 38.5 (6.2) | 8.89 | 0.82 | 4.292 | 90.8 | 95.2 |
| 2 | 35.1 (6.2) | 38.6 (6.3) | 13.57 | 0.711 | 4.428 | 91.6 | 95.9 |
| 3 | 34.3 (6.8) | 38.8 (7.1) | 9.53 | 0.851 | 4.141 | 86.8 | 94.7 |
| **Reaction Time** | | | | | | | |
| 1 | 0.63 (0.08) | 0.61 (0.07) | 0.327 | 0.454 | 0.063 | 91.6 | 94.4 |
| 2 | 0.63 (0.09) | 0.60 (0.08) | 0.367 | 0.378 | 0.071 | 92.1 | 94.9 |
| 3 | 0.63 (0.092) | 0.61 (0.09) | 0.294 | 0.505 | 0.072 | 93.9 | 94.7 |
| **Symptom Scale** | | | | | | | |
| 1 | 3.3 (7.7) | 3.6 (7.4) | 1.82 | 0.525 | 6.202 | 90.4 | 95.2 |
| 2 | 3.1 (6.8) | 2.8 (6.1) | 1.77 | 0.324 | 5.678 | 94.9 | 94.9 |
| 3 | 2.7 (5.6) | 2.9 (5.7) | 1.86 | 0.393 | 5.245 | 89.5 | 90.4 |

[a]$\alpha$, intercept; $\beta$, slope; Sxy, standard error of the estimate (13).
[b]CI, confidence interval; numbers represent the percent of participants with change scores within cutoff (90% CI = 1.65; 90% = 1.96).

investigation indicate a range of stability estimates of ImPACT composite scores at each time interval from the original baseline assessment. ICCs ranged from 0.36 to 0.90 and showed little change from year to year among composite scores, with the largest variation noted for the Verbal Memory composite scale and post-concussion symptom scale. Additionally, the large majority of repeat baseline scores fell within the expected 90% and 95% CI ranges for RCI and RBM, with the exception of a low percentage of scores in select composite scores (Table 5). Ultimately, these results suggest the stability of ImPACT composite scores (i.e., Reaction Time) and clinical utility in applying RCI and RBM when assessing for meaningful change.

Previous studies of collegiate athletes have demonstrated acceptable long-term test–retest reliability and stability of the online version of the ImPACT test battery for intervals of up to 2 years (Maerlender & Molfese, 2015; Schatz, 2010). Other studies have provided mixed findings on the stability of ImPACT in a sample of high school athletes, with respectable reliability over a period of 1 year in between baselines (Elbin et al., 2011) and poor reliability over a period of 2 years (Tsushima et al., 2016). Interestingly, this study produced significantly lower ICC reliability estimates for all ImPACT composite scores, with the exception of visual motor speed, which was approximately similar (ranges 0.82–0.85). Additionally, lower test–retest reliability estimates at Years 1, 2, and 3 intervals within this study were all lower than those established previously by Elbin et al. (2011) in a sample of high school athletes. The influence of developmental changes within this population should also be considered as a possible source of variation instability estimates across and within studies. As previously shown, developmental changes in performance on computerized cognitive test paradigms, especially between the ages of 9 and 15 (McCrory, Collie, Anderson & Davis, 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). Lower reliability estimates and significant changes from pretest to posttest on computerized neurocognitive test paradigms as attributable to developmental changes is further supported by previous studies examining similar age-related changes using paper and pencil measures. Within the area of visual motor processing speed, consistent improvement in performance on paper and pencil measures of visual motor speed (e.g., TMT and a two-letter cancellation task) have been observed annually in subjects between the ages of 12 and 17 years old (Kumar Sharma, Kumar Subramanian, Sarah, Balasubramaniam, Velkumary, 2014). Additionally, continued improvements in visual motor speed are expected through development, as collegiate athletes have performed significantly better than high school counterparts on computerized and paper and pencil neurocognitive measures (Register-Mihalik et al., 2012). These developmental changes are less expected for verbal memory within a similar sample (Maril et al., 2010; Schenider, Knopf & Stefanek, 2002), suggesting lower levels of reliability being due to other factors. Although developmental trajectories of visual memory is limited, previous investigations suggest that recognition memory begins to plateau around the age of 13 (Flin, 1985), where visual working memory continues to develop through the age of 16 (Isbell, Fukuda, Neville & Vogel, 2015).

Given the considerably large sample, it is very possibly for significant variation in the test–retest individual scores to account for the lower than expected reliability estimates. This notion is further supported by the rather large CIs generated for each composite score. Considering the conventional ICC cutoff of 0.70 (Baumgartner & Chung, 2001; Slick, 2006), reliability estimates trended towards the lower bound of acceptable ranges. However, composite scores did demonstrate relative consistency across the 3 years when methods such as RCI and RBM were utilized. Consistent with previous studies, the lowest level of reliability was established for verbal memory and the higher levels of reliability included Visual Motor Speed and Reaction Time (Broglio et al., 2007; Bruce et al., 2014; Elbin et al., 2011; Resch et al., 2013; Schatz, 2010; Tsushima et al., 2016).

To date, this study contained the largest test–retest sample of athletes, and with increased sample size, the findings yielded from this investigation can be interpreted with greater confidence. Given the range of reliability estimates demonstrated within this study, data suggests that stability of ImPACT varies depending on the particular composite score and interpretations of score change should be made based upon these differences. Additionally, little variation in the reliability of composite scores was observed between intervals of 1, 2, or 3 years. The sizeable CIs generated by the data indicate that significant deviation among participants' scores does in fact occur at the acquisition of baseline. Furthermore, we observed a general trend of increasing standard errors of the estimate (for RBM) and standard errors of the difference (for RCI) values as test–retest intervals were expanded. These two values largely influence the CI and determine the range of allowable change in scores before they are considered to be statistically different. The wider the CI range, the greater the possibility of a false negative for any one score. Considering the minimal variation in reliability across the 3 years and increase in CI range with time, results from this study suggest that it does not matter if you wait 1, 2, or 3 years to update a baseline, as the variation between groups demonstrates minimal variation. However, individuals may vary greatly, and the large CIs required to capture clinically meaningful change may not be all that useful in detecting more mild deficits. The data also suggest that the greater the duration of time, the increased likelihood of a false-negative error in attempting to identify reliable change in composite scores, or in other words, failing to detect concussion.

This study is not without limitations. Firstly, although the present sample is the largest to date examining the test–retest reliability of the test battery, results should not be generalized to populations outside the high school athlete. Limitation within

experimental control and uniformity of procedures is also limited. Although all assessments were administered in group settings and proctored by a certified athletic trainer who had been trained in the administration of ImPACT, variation among administration procedures by trainer or by school may exist and account for aspects of variation within the data. As previously noted (Moser, Schatz, Neidzwski & Ott, 2011), examinees perform more poorly when tested in groups as compared to individual administration, and these effects are magnified when younger athletes (Moser, Schatz & Lichtenstein, 2015) are being tested, especially those with attention-deficit/hyperactivity disorder (Vaughan, Gerst, Sady, Newman & Gioia, 2014). In this regard, the current results should not be generalized to instances in which individual administration was performed. In addition, experimental control in the form of group size standardization was not achieved, as differences across schools and trainers may vary in regards to the number of students included in group administration sessions. Had administration procedures been in alignment with current recommendation that examinees be tested in small groups of ≤5 (Echemendia et al., 2013), results of this investigation could be interpreted with greater confidence.

Future investigations should look to resolve discrepancies of 1-year test–retest stability between this study and that of Elbin et al. (2011). Future studies should also further investigate discrepancies among test–retest reliability of the instrument across shorter and longer time periods, as some studies have demonstrated estimates of reliability for the ImPACT over 45 days (Broglio et al., 2007) to be lower than those of 2–3 years within this study. As noted by Schatz (2010), these scores were obtained following an extensive list of test batteries and lower estimates of reliability obtained may reflect more fatigue than instability of the instrument. Given the limitations of the ICC in assessing test–retest reliability, as well as the lack of investigations currently on the topic, future research should focus on evaluating the ImPACT test battery using RCIs and RMBs across a variety of time intervals (i.e., 30 days, 45 days, 6 months, 1 year, 2 years, etc.).

In summary, establishing test stability is essential for the effective use of neurocognitive testing in the objective management of sport-related concussion. Findings from this study provide test–retest reliability of ImPACT composite scores at 1-, 2-, and 3-year intervals, which can assist with the interpretation of change at different time intervals for each individual composite score. Although there is little effect of obtaining baselines at 1-, 2-, or 3-year intervals in terms of reliability, individual variation in determining clinically meaningful change may be difficult to detect with large CIs generated by such wide variation. Additionally, the greater the time interval between retest periods, the increased likelihood of false-negative errors when attempting to identify meaningful change in testing scores. As such, annually ascertaining baseline scores continues to be optimal for ensuring accurate and individualized management of injury for athletes who sustain a concussion. For instances in which annual or biennial baselines are not available, clinicians should seek to utilize more conservative estimates (CI) when determining the presence of clinically meaningful change.

## Conflict of Interest

Dr. Schatz is a consultant to ImPACT Applications Inc. and Drs. Schatz and Solomon are members of the Scientific Advisory Board of ImPACT, and receive reimbursement for expenses to Board meetings. However, ImPACT had no input into the decision to conduct the study, its design, or execution, and no funding was obtained for this project.

## References

Allen, B. J., & Gfeller, J. D. (2011). The immediate post-concussion assessment and cognitive testing battery and traditional neuropsychological measures: A construct and concurrent validity study. *Brain Injury*, *25*, 179–191.

Alsalaheen, B., Stockdale, K., Pechumer, D., & Broglio, S. P. (2015). Measurement error in the immediate postconcussion assessment and cognitive testing (ImPACT): Systematic review. *J Head Trauma Rehabil*. [Epub ahead of print].

American Academy of Neurology. (1997). Practice parameter: The management of concussion in sports (summary statement). Report of the Quality Standards Subcommittee. *Neurology*, *48*, 581–585.

Aubry, M., Cantu, R., Dvorak, J., Graf-Baumann, T., Johnston, K., & Kelly, J., et al. & Concussion in Sport Group. (2002). Summary and agreement statement of the 1st International Symposium on Concussion in Sport, Vienna 2001. *Clinical Journal of Sports Medicine*, *12*, 6–11.

Barr, W. B. (2002). Neuropsychological testing for assessment of treatment effects: Methodologic issues. *CNS Spectrums*, *7*, 300–302, 304–306.

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. *Arch Clin Neuropsychol*, *27* (3), 362–373.

Baumgartner, T. A., & Chung, H. (2001). Confidence limits for intraclass reliability coefficients. *Measurement in Physical Education and Exercise Science*, *5*, 179–188.

Bonate, P. L. (2000). *Analysis of pretest-posttest designs*. Boca Raton, FL: Chapman and Hall/CRC.

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.

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*, 543–570.

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

Chicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.

Choudhury, S., Blakemore, S., & Charman, T. (2006). Social cognitive development during adolescence. *Social Cognitive and Affective Neuroscience*, *1*, 165–174.

Cioni, G., & Sgandurra, G. (2013). Normal psychomotor development. *Handbook of Clinical Neurology*, *111*, 3–15.

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.

Collie, A., Maruff, P., Darby, D., & McStephen, M. G. (2009). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. *Journal of the International Neuropsychological Society*, *9*, 419–428.

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.

Crone, E. A., & Elzinga, B. M. (2015). Changing brains: How longitudinal functional magnetic resonance imaging studies can inform us about cognitive and social-affective growth. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*, 53–63.

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* (7), 1077–1091.

Echemendia, R. J., Iverson, G. L., McCrea, M., Macciocchi, S. N., Gioia, G. A., & Putukian, M., et al. (2013). Advances in neuropsychological assessment of sport-related concussion. *British Journal of Sports Medicine*, *47*, 294–298.

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.

Flin, R. H. (1985). Development of visual memory: An early adolescent regression. *The Journal of Early Adolescence*, *5*, 259–266.

Guskiewicz, K. M., Bruce, S. L., Cantu, R., Ferrara, M. S., Kelly, J. P., & McCrea, M., et al. (2004). Recommendations on management of sport-related concussion: Summary of the National Athletic Trainers' Association position statement. *Neurosurgery*, *55*, 891–895.

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.

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

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.

Isbell, E., Fukuda, K., Neville, H. J., & Vogel, E. K. (2015). Visual working memory continues to develop through adolescence. *Frontiers in Psychology*, *6*, 696.

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.

Iverson, G. L., & Schatz, P. (2015). Advanced topics in neuropsychological assessment following sport-related concussion. *Brain Injury*, *29*, 263–275.

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.

Kirschen, M. P., Tsou, A., Nelson, S. B., Russell, D. O., & Larriviere, D. (2014). Legal and ethical implications in the evaluation and management of sports related concussion. *Neurology*, *84*, 352–358.

?Kumar Sharma, V., Kumar Subramanian, S., Vinayathan, A., Sarah, R., Balasubramaniam, S.R., & Velkumary, S.(2014). Study of effect of age and gender related differences on common paper and pencil neurocognitive.

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

Lovell, M. R., Collins, M. W., Fu, F., Burke, C., & Podell, K. (2001). Neuropsychological testing in sports: Past, present, and future. *British Journal of Sports Medicine*, *35*, 367–372.

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.

Maril, A., Davis, P. E., Koo, J. J., Reggev, N., Zuckerman, M., & Ehrenfeld, L., et al. (2010). Developmental fMRI study of episodic verbal memory encoding in children. *Neurology*, *75*, 2110–2116.

McClure, D. J., Zuckerman, S. L., Kutscher, S. J., Gregory, A. J., & Solomon, G. S. (2014). Baseline neurocognitive testing in sports-related concussion: The importance of a prior night's sleep. *American Journal of Sports Medicine*, *42*, 472–478.

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.

McCrory, P., Meeuwisse, W., Johnston, K., Dvorak, J., Aubry, M., & Molloy, M., et al. (2009). Consensus statement on concussion in sport: The 3rd International Conference on Concussion in Sport held in Zurich, November 2008. *Journal of Athletic Training*, *44*, 434–448.

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., Schatz, P., & Lichtenstein, J. (2015). The importance of proper administration and interpretation of neuropsychological baseline and post-concussion computerized testing. *Applied Neuropsychology: Child*, *4*, 41–48.

Moser, R. S., Schatz, P., Neidzwski, K., & Ott, S. D. (2011). Group versus 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.

Parsons, T. D., Notebaert, A. J., Shields, E. W., & Guskiewicz, K. W. (2009). Application of reliable change indices to computerized neuropsychological measures of concussion. *International Journal of Neuroscience*, *119*, 492–507.

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.

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.

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

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.

Schenider, W., Knopf, M., & Stefanek, J. (2002). The development of verbal memory in childhood and adolescence: findings from Munich Longitudinal Study. *Journal of Educational Psychology*, *94*, 751–761.

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

Sullivan, G. M., & Feinn, R. (2012). Using effect size- or why the *p* value is not enough. *Journal of Graduate Medical Education*, *4*, 279–282.

Tsushima, W. T., Siu, A. M., Pearce, A. M., Zhang, G., & Oshiro, R. S. (2016). Two-year test-retest reliability of ImPACT in high school athletes. *Archives of Clinical Neuropsychology*, *31*, 105–111.

Vaughan, C. G., Gerst, E. H., Sady, M. D., Newman, J. B., & Gioia, G. A. (2014). The relation between testing environment and baseline performance in child and adolescent concussion assessment. *Am J Sports Med*, *42* (7), 1716–1723.

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.