OXFORD
UNIVERSITY PRESS

Archives of Clinical Neuropsychology 29 (2014) 625–632

Archives
of
CLINICAL
NEUROPSYCHOLOGY

Comparing Post-Concussive Neurocognitive Test Data to Normative Data Presents Risks for Under-Classifying “Above Average” Athletes

Philip Schatz*, Stacey Robertshaw

Department of Psychology, Saint Joseph’s University, Philadelphia, PA, USA

*Corresponding author at: Department of Psychology, Saint Joseph’s University, 222 Post Hall, Philadelphia, PA 19131, USA. Tel.: +1-610-660-1804; fax: +1-610-660-1819.
E-mail address: pschatz@sju.edu (P. Schatz).

Accepted 1 August 2014

Abstract

We compared classification accuracy of post-concussion test data against baseline and normative data, accounting for baseline level of performance. Athletes (N = 250) completed baseline and post-concussion ImPACT assessments, within 7 days of concussion (verified by sports medicine professionals and self-reported symptoms). Athletes were classified as “below average,” “average,” or “above average” at baseline. Change from baseline was calculated using reliable change indices (RCIs) and regression-based measures (RBz), and comparison to normative data was achieved using z-scores. Normative comparisons identified fewer symptomatic, concussed athletes than RCIs and RBz. Both RCIs and RBz consistently identified “impairment” at 1 and 1.5 *SD*, regardless of baseline level, whereas normative comparisons identified 46–48% fewer athletes performing “above average” at baseline using a cut-off of 1 *SD* and 36-38% fewer using a cut-off of 1.5 *SD*. The use of normative comparisons may differentially classify concussed, symptomatic athletes who are outside the “average” range at baseline.

Keywords: Concussion; mTBI; Neuropsychological assessment; Baseline assessment

Over the past decade, the topic of assessment and management of concussion has received increased attention in the media, in legislative actions in the United States, and in the sports medicine and neuropsychology literature. The use of preseason neurocognitive test data as a baseline comparison against post-concussive test data was introduced by Barth and colleagues in the late 1980s (Barth et al., 1989), and this paradigm was soon extended to professional sports (Lovell & Collins, 1998). Serial testing of athletes raised questions regarding test-retest reliability, especially with respect to practice effects, and “deficits” were often identified in concussed athletes in the context of failure to benefit from repeated administrations, when compared with non-concussed controls (Macciocchi, Barth, Alves, Rimel, & Jane, 1996).

Researchers have since identified methods for calculating “reliable” change from baseline (Barr, 2002; Chelune, Naugle, Lüders, Sedlak, & Awad, 1993; Hinton-Bayre, Geffen, Geffen, McFarland, & Friis, 1999; Iverson, Lovell, & Collins, 2003; Jacobson & Truax, 1991). In contrast to concussion testing, interpretation of data from traditional neuropsychological test batteries typically involves comparison to normative data. Researchers have only recently begun evaluating the utility of comparing post-concussion test data to normative versus baseline data. Echemendia and colleagues (2012) documented the utility of comparing post-concussion neuropsychological data with normative data (i.e., in the absence of baseline data), and concluded that the majority of college athletes who experience clinically meaningful post-concussion cognitive decline can be identified without baseline data. Schmidt, Register-Mihalik, Mihalik, Kerr, and Guskiewicz (2012) found that baseline comparisons identified more impaired athletes on a simple reaction time task, and normative comparisons identified more impaired athletes on a mathematical processing task. In the absence of superiority of one method over the other, they ultimately concluded that clinicians may consider using comparisons to normative data in lieu of comparisons to baseline, especially when resources are limited, or valid baseline data cannot be obtained.

Basing decision-making criteria on deviation from the mean (e.g., normative data), rather than change from a previous (or premorbid) level of functioning has been regarded as “likely to result in missing significant changes in very high functioning

doi:10.1093/arclin/acu041 Advance Access publication on 1 September 2014

individuals while suggesting that low functioning individuals have acquired impairments that they do not have" (Lezak, 2004, p. 148). Comparisons to baseline data are typically made using a reliable change index (RCI) or regression-based (RBz) approach, which identify *z*-score-based cut-offs representing specific confidence intervals. More specifically, the use of two-tailed confidence intervals of 80%, 90%, and 95% correspond to *z*-score cut-offs of 1.28, 1.64, and 1.96, which, in turn, represent 10%, 5%, and 2.5% chance of Type I error in each "tail." In contrast, comparisons to normative data are typically made using set *z*-score cut-offs, which represent specific percentile ranks. In this regard, two-tailed cut-offs of 1 or 1.5 *SD* represent 8% and 3% chance of Type I error in each tail. Given that these approaches utilize different cut-offs, it is difficult to make comparisons between the two.

The purpose of this study was to document the utility of comparing post-concussion test data to either baseline data or normative data, while accounting for baseline level of performance. We hypothesized that, using post-concussion comparisons to normative data, athletes demonstrating "below average" performance at baseline would be disproportionately classified as "below average" following a concussion, when compared with those demonstrating "above average" performance at baseline. In addition, we hypothesized that comparison of post-concussion performance to baseline performance, using Reliable Change and Regression-based Measures, would classify concussed athletes proportionally, regardless of their level of baseline performance. In order to compare the utility of the "normative data" and "baseline data" approaches using the same metric, we utilized identical *z*-score cut-offs of 1, 1.5, 2, 2.5, and 3 *SD*.

Methods

Participants

Participants were 250 athletes, predominantly men (73%), ages 13–21 (mean 15.8, *SD*. 1.9), competing in football (49%), soccer (13%), basketball (12%), lacrosse (6%), volleyball (6%), and other sports (i.e., cheerleading, field hockey, ice hockey, softball, and wrestling; 14%). All athletes completed a valid baseline neurocognitive assessment using the ImPACT test battery, as indicated by baseline test scores which fell within pre-established ranges by the test developers (for more information, refer to the ImPACT manual; Lovell, 2011). Data were extracted from a larger database, using the following criteria, as well as those listed in the "Procedures" section. All athletes completed baseline assessments during the 2010 through 2013 athletic seasons, and subsequently sustained a concussion that was (1) identified by a certified athletic trainer or sports medicine professional, (2) supported by subjective endorsement of post-concussion symptoms by the athlete, and (3) assessed using ImPACT within 7 days (mean 3.5, *SD* 1.8).

Materials

The ImPACT test (Lovell, 2011) comprises a demographic section, symptom inventory, and six subtests measuring attention, memory, processing speed, and reaction time, yielding composite scores in the areas of Verbal Memory, Visual Memory, Visual Motor Processing Speed, and Reaction Time. The reliability (Elbin, Schatz, & Covassin, 2011; Schatz, 2009; Schatz & Ferris, 2013) and validity (Iverson, Gaetz, Lovell, & Collins, 2005; Maerlender et al., 2010, 2013; Schatz, Pardini, Lovell, Collins, & Podell, 2006; Schatz & Sandel, 2012) of the ImPACT test have been documented and also debated (Lovell, 2006; Mayers & Redick, 2012; Randolph, 2011; Randolph, Lovell, & Laker, 2011; Randolph, McCrea, & Barr, 2005; Schatz, Kontos, & Elbin, 2012) in the literature.

Procedures

Athletes completed pre-season and post-concussion testing using the ImPACT test as a function of their participation in the school or university athletics program. Baseline testing was administered by ATCs and/or sports medicine personnel, as was post-concussion testing within 1 week of sustaining a concussion. Baseline testing was administered in groups of 10–20 athletes, depending on the school or university, and post-concussion testing was administered in a private test session. All assessments were conducted using the online version of ImPACT, which employs different sets of stimuli for baseline and post-concussion testing. The ImPACT test includes a demographic section and history of previous concussion and history of attention deficit and learning disability were self-reported by athletes. All baseline tests utilized "Baseline" stimuli sets, whereas post-concussion tests utilized "Post-injury 1" stimuli. The retrospective analysis of de-identified data was approved by the Saint Joseph's University Institutional Review Board.

Analyses

Differences between baseline and post-concussion scores were calculated using RCIs and regression-based methods. Note that both formula require test-retest reliability coefficients, which were abstracted from 1-year test-retest reliability data (0.45, Verbal Memory; 0.55 Visual Memory; 0.74, Visual Motor Speed, 0.62, Reaction Time; Elbin et al., 2011), as the average interval from baseline to post-concussion assessment was 168 days. These test-retest data fall within the range of reliability coefficients (Pearson's *r*) documented at 45 days (Nakayama, Covassin, Schatz, Nogle, & Kovan, 2014) and 2 years (Schatz, 2009).

(1) RCIs (Jacobson & Truax, 1991) were calculated to assess whether a change between repeated assessments was reliable and meaningful. The RCI provides an estimate of the probability that a given difference in scores would not be obtained as a result of measurement error (Iverson, Sawyer, McCracken, & Kozora, 2001). A modified RCI formula (Chelune et al., 1993), which includes an adjustment for practice effects, was also calculated.

(2) Regression-based methods (RBz) were also calculated. In accordance with McCrea and colleagues (2005), we employed linear regression using baseline scores from the baseline (i.e., healthy control) data (Time 1) to generate a formula for predicting follow-up (i.e., post-concussion) scores (Time 2). The regression coefficient and the intercept of the regression line were then used with the baseline score to compute a predicted score for each participant at Time 2. This approach provides an empirical method for detecting meaningful change while also providing correction for practice effects and regression to the mean. Participants were considered to have undergone a meaningful change in test performance if the difference between the obtained and predicted score, divided by the standard error of prediction, was larger than a specific criterion value (e.g., a *z*-score of 1.28, 1.64, and 1.96) which translates to a specific confidence interval (e.g., 80%, 90%, and 95% confidence intervals, representing, two-tailed 10%, 5%, and 2.5% chance of Type I error). The resulting equation provides an empirical method for detecting "meaningful change" while adjusting for practice effects as well as controlling for regression to the mean (Hsu, 1995; McCrea et al., 2005).

Comparison to normative data (Normative) was examined using *z*-scores, which were calculated by subtracting the athlete's age- and gender-based mean (found in the ImPACT manual (Lovell, 2011)) from their post-concussion test score, and dividing by their age- and gender-based standard deviation.

Cut-offs of 1 and 1.5 *SD* are commonly employed by neuropsychologists to identify possible deficits on cognitive tests (Barr, 2003; Heaton, Grant, & Matthews, 1991), and *z*-score cut-offs of 1.28, 1.64, and 1.96 are commonly employed for RCI and RBz (representing 80%, 90%, and 95% confidence intervals). In order to allow for cross-comparison between RCI, RBz, and Normative approaches, we compared the percentage of athletes surpassing (i.e., falling below) cut-offs of 1, 1.5, 2, 2.5, and 3 *SD*.

In order to determine classification differences for below-average, average, and above-average athletes, baseline ImPACT composite scores (Verbal Memory, Visual Memory, Visual Motor Processing Speed, Reaction Time) were all converted to *z*-scores by subtracting the age- and gender-based mean from the post-concussion test score, and dividing the result by the age- and gender-based standard deviation. Thus, the *z*-scores yielded for the Verbal Memory, Visual Memory, Visual Motor Speed, and Reaction Time were averaged, and participants were coded as "Below Average" (<-1 *SD*), "Average" (-1 to $+1$ *SD*), or "Above Average" ($>+1$ *SD*) on the basis of this average baseline *z*-score. Cases were sampled to approximate a normal distribution, such that there was adequate representation in the "below average" and "above average" groups (or "tails"). As such,

Table 1. Group Demographics

	Below average	Average	Above average
Gender			
Male	40 (80%)	105 (70%)	37 (74%)
Female	10 (20%)	45 (30%)	13 (26%)
ADD/LD	7 (14%)	6 (4%)	5 (10%)
Concussion history			
0	31 (65%)	106 (76%)	33 (67%)
1	13 (27%)	24 (17%)	12 (25%)
2+	4 (8%)	9 (7%)	4 (8%)
	Individual	Group	Total
Age	15.6 (1.8)	15.9 (1.9)	15.3 (2.1)
Days since concussion	4.3 (2.0)	3.2 (2.0)	3.3 (1.6)

cases were sampled until 150 were represented in the "average" group, and then cases were sampled until 50 athletes were represented in the "below average" and "above average" groups (i.e., 20%, 60%, 20%, respectively) (Table 1).

Between-method classification likelihoods (i.e., above or below the specific cut-off for RCI vs. RBz vs. Normative comparisons) were analyzed using χ^2 analyses, using an α-level of $p < .003$ to control for family-wise error. To evaluate agreement between the three classification methods, Cohen's κ statistic with bootstrapping using 500 re-samplings was used, with values below 0.20 representing slight agreement, between 0.21 and 0.40 representing fair agreement, between 0.41 and 0.60 representing moderate agreement, between 0.61 and 0.80 representing substantial agreement, and above 0.80 representing almost perfect agreement (Landis & Koch, 1977). Given that all athletes were concussed (or observed and presumed to be), we also calculated weighted specific-category κ as a measure of evaluate agreement on this specific classification (Kvalseth, 2003).

Results

χ^2 analyses revealed no between-groups differences for gender ($p = .38$), previous history of concussion ($p = .53$), with a significant difference noted in likelihood of athletes in the "below average" (14%) and "above average" (10%) groups having a diagnosed learning disorder or attention deficit disorder than those in the "average" group (4%; $p = .048$).

Analysis of variance revealed significant between-groups differences for all four ImPACT Composite scores at baseline and post-concussion (Table 2). However, groups showed no differences on Total Symptom scores at baseline, or post-concussion. The *post hoc* Tukey analyses of baseline data identified the "Below Average" group as scoring significantly worse than the "Average" and "Above Average" groups on all four composites scores. The *post hoc* Tukey analysis of post-concussion data identified the "Below Average" group scoring significantly worse than both the "Average" and "Above Average" groups on Visual Motor Speed and Reaction Time, and the "Below Average" and "Average" groups scoring significantly worse than the "Above Average" group on Verbal and Visual Memory.

Using a 1 *SD* cut-off for comparison to normative data yielded a sensitivity of 72%, when compared with 94% using RCI and RBz (Table 3, "All Cases" column). Using a cut-off of 1.5 *SD*, comparison to normative data yielded a sensitivity of 62%, when compared with 80% using RCI and 83% using RBz. Using cut-offs of 2–3 *SD* yielded ~10–20% lower sensitivity for the "comparison to normative data" method than comparison to baseline using RCI or RBz.

When accounting for the athletes' baseline level of performance, RCI and RBz consistently identified "impairment" at 1 and 1.5 *SD*, regardless of the athlete's baseline level of performance (Table 3, 1 *SD* and 1.5 *SD* rows). In contrast, comparison to normative data classified 46–48% fewer "above average" athletes than did RCI or RBz, using a cut-off of 1 *SD*, and 36–38% fewer using a cut-off of 1.5 *SD*.

Across all three methods, using a cut-off of 2 *SD* or higher, far fewer "above average" athletes at baseline were identified as impaired than "average" or "below average" athletes. In addition, comparison to normative data identified a greater number of "below average" athletes falling below cut-offs for impairment, than "average" or "above average" athletes.

χ^2 comparisons between RCI, RBz, and Normative approaches (e.g., 94% RCI, 92% RBz, 46% Normative) revealed significantly fewer "above average" athletes classified using the Normative approach than RCI or RBz, at all 5 *SD* cut-offs ($p < .002$).

Table 2. Group ImPACT Data

	Below average	Average	Above average	*d*[a]
Verbal memory				
Baseline	80.8 (11.0)	85.4 (9.1)	86.1 (8.1)	0.48, 0.08
Post-injury	63.6 (12.3)	72.1 (16.4)	79.1 (12.9)	0.36, 0.45
Visual memory				
Baseline	69.7 (13.9)	74.0 (11.5)	76.5 (13.9)	0.35, 0.21
Post-injury	55.7 (10.8)	60.9 (15.6)	67.7 (15.8)	0.36, 0.43
Visual motor speed				
Baseline	33.4 (8.4)	37.2 (6.6)	36.8 (7.5)	0.53, 0.06
Post-injury	25.9 (7.5)	32.5 (8.8)	34.3 (7.4)	0.77, 0.21
Reaction time				
Baseline	0.65 (.12)	0.58 (0.07)	0.61(0.08)	0.82, 0.41
Post-injury	0.80 (.15)	0.70 (0.18)	0.66 (0.12)	0.58, 0.24
Total symptoms				
Baseline	3.6 (6.1)	4.6 (6.9)	2.5 (3.7)	0.15, 0.33
Post-injury	35.9 (20.9)	32.5 (16.9)	28.5 (18.6)	0.19, 0.23

Notes: [a]*d*, denotes Cohen's *d*, representing the effect size difference between the Below Average and Average groups, and the Average and Above Average groups.

Table 3. Sensitivity to impairment by method and standard deviation

Cut-off (*SD*)	Baseline performance group			
	Below average	Average	Above average	Total
Comparison to baseline (Reliable Change Indices)				
1	45 (98%)	131 (92%)	40 (94%)	216 (94%)
1.5	42 (90%)	107 (78%)	33 (78%)	182 (80%)
2	37 (78%)	88 (61%)	26 (62%)	151 (65%)
2.5	32 (66%)	67 (46%)	18 (38%)	117 (48%)
3	25 (52%)	53 (36%)	8 (16%)	86 (35%)
Comparison to baseline (Regression-Based Methods)				
1	48 (98%)	134 (93%)	42 (92%)	224 (94%)
1.5	47 (96%)	121 (83%)	31 (72%)	199 (83%)
2	46 (94%)	97 (66%)	26 (54%)	169 (69%)
2.5	38 (76%)	86 (58%)	21 (42%)	145 (58%)
3	34 (70%)	68 (45%)	16 (32%)	118 (48%)
Comparison to normative data				
1	48 (96%)	110 (73%)	23 (46%)	181 (72%)
1.5	44 (88%)	93 (62%)	18 (36%)	155 (62%)
2	36 (72%)	74 (49%)	13 (26%)	123 (49%)
2.5	27 (54%)	62 (41%)	6 (12%)	95 (38%)
3	16 (32%)	53 (35%)	2 (4%)	71 (28%)

Table 4. Agreement (κ) between comparison to baseline using RCI, comparison to baseline using RBz, and comparison to normative data

Cut-off (*SD*)	Comparison method		
	RCI/RBz	RCI/Norm	RBz/Norm
1	0.55/0.76	0.37/0.64	0.44/0.68
1.5	0.60/0.77	0.48/0.69	0.56/0.74
2.0	0.71/0.82	0.58/0.73	0.62/0.75
2.5	0.68/0.79	0.51/0.66	0.57/0.70
3	0.69/78	0.31/0.47	0.28/0.43

Notes: Jackknifed κ and weighted specific-category κ are presented.

However, $\chi2$ comparisons between RCI, RBz, and Normative approaches revealed significantly fewer "average" athletes classified using the Normative approach than RCI or RBz, only at the 1 and 1.5 *SD* cut-offs ($p < .002$). Finally, χ^2 comparisons between RCI, RBz, and Normative approaches revealed significantly fewer "below average" athletes classified using the Normative approach than RCI or RBz, only at the 3 *SD* cut-off ($p < .002$).

κ and weighted specific-category κ were calculated as a measure of agreement between classification methods (Table 4). The highest agreements (κ) were seen between the two "comparison to baseline" approaches (RCI and RBz), in the range of 0.55 (moderate agreement) to 0.71 (substantial agreement). Weighted specific-category measurements reflected substantial agreement for all cut-offs, with one in the "near perfect" agreement range. Agreement between "comparison to baseline" and "comparison to normative methods" were lower at all cut-offs, ranging from "fair" (0.31) to "moderate" (0.58) for the RCI/Normative agreements and "fair"(0.28) to "substantial" (0.62) for the RBz/Normative agreements. Weighted specific-category measurements reflected "moderate" (0.47) to "substantial" (0.73) for the RCI/Normative agreements, and "moderate" (0.43) to "substantial" (0.75) for the RBz/Normative agreements.

Discussion

We documented different rates of classification of impairment following a concussion, using RCIs, Regression-Based Measures, and comparing post-concussion scores to normative data, when accounting for baseline level of performance. Overall, results showed a greater number of cases identified as "impaired" using RCI and RBz, versus the use of comparison to normative data, regardless of the z-score cut-off used. However, when accounting for baseline level of performance, concussed, "above average" athletes, were consistently under-classified as impaired when post-concussion data were compared with normative data.

Similar percentages of concussed athletes classified as performing “below average” at baseline were identified as impaired using cut-offs of 1 and 1.5 *SD*, regardless of whether comparisons were made to normative data or baseline data (using either RCI or RBz), but this pattern did not continue using cut-offs of 2, 2.5, or 3 *SD*. In addition, fewer percentages of concussed athletes classified as “average” or “above average” were identified as impaired using comparisons to normative data versus comparison to baseline data (using either RCI or RBz) across the full range of SDs. Overall (e.g., in the “total” columns in Table 3), comparison to normative data appears to “lag behind” the comparison to baseline data (using either RCI or RBz) by ~0.5 *SD*. In other words, in the absence of baseline data, the use of normative data represents a wider confidence interval and greater likelihood of Type I error.

The overall average classifications obtained in the current results are not in agreement with recent findings documenting similar efficacy comparing post-concussion data to normative versus baseline data (Echemendia et al., 2012; Schmidt et al., 2012), and accounting for an athlete’s baseline level of performance yields substantially different rates of impairment for athletes in the above-average, average, and below-average ranges. Researchers have recognized that factors such as ADHD and low/high intelligence may moderate post-injury performance (Echemendia et al., 2012; Rabinowitz & Arnett, 2012). Lezak (2004) postulated that decision-making criteria based on comparisons to normative data would likely overlook significant change (e.g., from baseline) in high functioning individuals while over-classifying lower functioning individuals. Similarly, estimated base rate of impairment on neuropsychological testing was found to lose considerable accuracy in individuals with low or high intelligence (Brooks & Iverson, 2010). The current results suggest that comparison to normative data, in the absence of any data on premorbid or baseline level of functioning, tend to under-classify above-average athletes.

The current sample reflected a significantly higher representation of athletes with ADD/LD in the “below average” (14%) and “above average” (10%) groups. As “comparison to baseline” yielded higher rates of identification of impairment than “comparison to normative data,” the current results may have implications for determining which student athletes should be given highest priority for receiving baseline assessments, especially in situations where resources are limited. However, given that “comparison to normative data” yielded lower rates of identification of impairment than “comparison to baseline,” even for athletes in the “average” group, baseline data appear to be a more useful and sensitive measure of post-injury neurocognitive functioning than normative data.

At present, there are no “built-in” measures of crystallized intelligence or pre-morbid functioning in concussion screening measures (e.g., ImPACT, Headminder’s CRI, Axon). Although the use of pre-season neurocognitive test data would seem a likely comparator, research has yielded variable results with respect to the reliability of baseline assessments across a range of days (Iverson et al., 2003), weeks (Broglio, Ferrara, Macciocchi, Baumgartner, & Elliott, 2007; Resch et al., 2013; Schatz & Ferris, 2013), and years (Elbin et al., 2011; Schatz, 2009). The current results suggest that until a valid and reliable measure of pre-morbid functioning can be identified to classify athletes as above-average, average, or below-average, comparison of post-concussion data solely to normative data may present a risk of under-classifying athletes in the above-average range.

Consensus experts on sports-related concussion recognize that neurocognitive screening measures should not be used as a “stand-alone tool” for the management of sports concussions (McCrory et al., 2013), but rather, should be used in tandem with objective measures of balance and concussion symptom scales. In this context, regardless of whether an athlete completed a baseline evaluation, their post-concussion test scores have “returned to baseline,” or their post-concussion test scores are “within normal limits,” the use of baseline and/or post-concussion test scores should only be one “tool” or indicator in the return to play decision-making process. Ultimately, concussed athletes must be asymptomatic, progress through the widely used “graded return to play protocol” (McCrory et al., 2013), and be cleared by team physicians and/or independent medical experts. Neurocognitive testing remains a widely used tool that has been shown to provide diagnostic value beyond symptom reporting alone (Van Kampen, Lovell, Pardini, Collins, & Fu, 2006). Neurocognitive testing is also helpful in identifying athletes who may be purposefully denying concussion-related symptoms (Schatz & Sandel, 2012). In this context, individuals interpreting test results should be sensitive to individual differences inherent in athletes, and the implications of normative comparisons for those individuals who fall outside the “average” range.

The current results are tempered by the limitations of the study. First, only one computer-based neurocognitive measure (ImPACT) was used, so the results may not generalize to other assessment tools. Secondly, while data were collected prospectively, for the purpose of concussion management and return-to-play decision-making, study design and data analyses were conducted retrospectively. As such, experimental controls associated with more formal experimental procedures (e.g., random assignment, use of control group) were not present at baseline. Thirdly, although athletes were identified as having sustained concussions by qualified sports medicine professionals, and diagnoses were corroborated by objective self-reported concussion symptoms, there remains no “gold standard” for definitive diagnosis of concussion. Although post-concussion assessment is typically performed when an athlete is clinically asymptomatic, consensus experts note that cognitive recovery may occasionally precede or (more commonly) follow clinical symptom resolution (McCrory et al., 2013). As athletes often deny or under-report concussions and concussion symptoms (McCrea, Hammeke, Olsen, Leo, & Guskiewicz, 2004), we opted to include data from athletes who were symptomatic in order to insure the sample was represented data from athletes in the acute stages of concussion. In addition, while ImPACT data reflected decreased performance following concussion, and RCI and RBz classification methods yielded greater sensitivity, there were no corroborative neurocognitive data available to confirm diagnostic accuracy. Fourthly, while baseline assessments were classified as "valid" using built-in parameters in ImPACT, athletes were assessed in computer laboratories in groups of 10–20 individuals. As group administration has been documented to result in decreased performance, when compared with an individualized setting (Moser, Schatz, Neidzwski, & Ott, 2011), this may have contributed to decreased performance at baseline, as well as increased variability in baseline scores. Fifthly, RCI and RBz statistics were calculated using test-retest data from a 1-year interval (Elbin et al., 2011) from a sample of 369 high school athletes which may or may not be representative of the current sample of high school and collegiate athletes. Although researchers have documented lower reliability data using intraclass correlation coefficients (Broglio et al., 2007; Resch et al., 2013), Pearson's *r* data were not documented for use in RCI or RBz analyses. In this regard, the use of higher test-retest coefficients may inflate the sensitivity of RCI and RBz data. Finally, the sample was composed 80% men and ~50% football high school and collegiate athletes, so the results may not generalize well to female athletes and males playing other sports. Despite these limitations, the current findings suggest that the use of normative data for comparison to post-concussion test data may fail to classify a significant percentage of "above average" athletes.

Conflict of Interest

Dr P.S. has received funding to study the effects of concussion in high school and collegiate athletes from the International Brain Research Foundation and the Sports Concussion Center of New Jersey. He has also served as a consultant to ImPACT Applications, Inc. However, these entities had no role in the conceptualization of the study, the collection or analysis of data, the writing of the article, or the decision to submit it for publication. Ms S.R. has no Conflict of Interest to declare.

References

Barr, W. B. (2002). Neuropsychological testing for assessment of treatment effects: Methodologic issues. *CNS Spectrum*, *7 (4)*, 300–302, 304–306.

Barr, W. B. (2003). Neuropsychological testing of high school athletes. Preliminary norms and test-retest indices. *Archives of Clinical Neuropsychology*, *18 (1)*, 91–101.

Barth, J. T., Alves, W. M., Ryan, T. V., Macciocchi, S. N., Rimel, R. W., Jane, J. A., et al. (1989). Mild head injury in sports: Neuropsychological sequelae and recovery of function. In: H. S. Levin, H. M. Eisenberg, & A. L. Benton (Eds.), *Mild Head Injury* (pp. 257–275). New York: Oxford University Press.

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

Brooks, B. L., & Iverson, G. L. (2010). Comparing actual to estimated base rates of "abnormal" scores on neuropsychological test batteries: Implications for interpretation. *Archives of Clinical Neuropsychology*, *25 (1)*, 14–21.

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.

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.

Elbin, R. J., Schatz, P., & Covassin, T. (2011). One-year test-retest reliability of the online version of ImPACT in high school athletes. *The American Journal of Sports Medicine*. doi:10.1177/0363546511417173.

Heaton, R. K., Grant, I., & Matthews, C. G. (1991). *Comprehensive norms for an expanded Halstead-Reitan battery: Demographic corrections, research findings, and clinical applications*. Odessa, FL: Psychological Assessment Resources.

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

Hsu, L. M. (1995). Regression toward the mean associated with measurement error and the identification of improvement and deterioration in psychotherapy. *Journal of Consulting and Clinical Psychology*, *63 (1)*, 141–144.

Iverson, G. L., Gaetz, M., Lovell, M., & Collins, M. (2005). Validity of ImPACT for measuring processing speed following sports-related concussion. *Journal of Clinical and Experimental Neuropsychology*, *27*, 683–689.

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

Iverson, G. L., Sawyer, D. C., McCracken, L. M., & Kozora, E. (2001). Assessing depression in systemic lupus erythematosus: Determining reliable change. *Lupus*, *10 (4)*, 266–271.

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 (1)*, 12–19.

Kvalseth, T. O. (2003). Weighted specific-category kappa measure of interobserver agreement. *Psychological Reports*, *93 (3 Pt 2)*, 1283–1290.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33 (1)*, 159–174.

Lezak, M. D. (2004). *Neuropsychological assessment*. New York, NY: Oxford University Press.

Lovell, M. R. (2006). Letters to the Editor. *Journal of Athletic Training*, *41 (2)*, 137–140.

Lovell, M. R. (2011). ImPACT technical manual: Online ImPACT 2007–2012 *(Vol. 2013)*. Pittsburgh, PA: ImPACT Applications, Inc.

Lovell, ., & Collins, M. W. (1998). Neuropsychological assessment of the college football player. *Journal of Head Trauma and Rehabilitation*, *13 (2)*, 9–26.

Macciocchi, S. N., Barth, J. T., Alves, W., Rimel, R. W., & Jane, J. A. (1996). Neuropsychological functioning and recovery after mild head injury in collegiate athletes. *Neurosurgery*, *39 (3)*, 510–514.

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 (8)*, 1309–1325.

Maerlender, A., Flashman, L., Kessler, A., Kumbhani, S., Greenwald, R., Tosteson, T., et al. (2013). Discriminant construct validity of ImPACT: A companion study. *The Clinical Neuropsychologist*, *27 (2)*, 290–299.

Mayers, L. B., & Redick, T. S. (2012). Clinical utility of ImPACT assessment for postconcussion return-to-play counseling: Psychometric issues. *Journal of Clinical and Experimental Neuropsychology*, *34 (3)*, 235–242.

McCrea, M., Barr, W. B., Guskiewicz, K., Randolph, C., Marshall, S. W., Cantu, R., et al. (2005). Standard regression-based methods for measuring recovery after sport-related concussion. *Journal of the International Neuropsychological Society*, *11 (1)*, 58–69.

McCrea, M., Hammeke, T., Olsen, G., Leo, P., & Guskiewicz, K. (2004). Unreported concussion in high school football players: Implications for prevention. *Clinical Journal of Sport Medicine*, *14 (1)*, 13–17.

McCrory, P., Meeuwisse, W. H., 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 (5)*, 250–258.

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*. doi:10.1177/0363546511417114.

Nakayama, Y., Covassin, T., Schatz, P., Nogle, S., & Kovan, J. (2014). Examination of the test-retest reliability of a computerized neurocognitive test battery. *The American* Journal of Sports Medicine. doi:10.1177/0363546514535901.

Rabinowitz, A. R., & Arnett, P. A. (2012). Reading based IQ estimates and actual premorbid cognitive performance: Discrepancies in a college athlete sample. *Journal of the International Neuropsychological Society*, *18 (1)*, 139–143.

Randolph, C. (2011). Baseline neuropsychological testing in managing sport-related concussion: Does it modify risk? *Current Sports Medicine Reports*, *10 (1)*, 21–26.

Randolph, C., Lovell, M., & Laker, S. R. (2011). Neuropsychological testing point/counterpoint. *PM& R: The Journal of Injury, Function, and Rehabilitation*, *3 (10 Suppl 2)*, S433–S439.

Randolph, C., McCrea, M., & Barr, W. B. (2005). Is neuropsychological testing useful in the management of sport-related concussion? *Journal of Athletic Training*, *40 (3)*, 139–152.

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 (4)*, 506–511.

Schatz, P. (2009). Long-term test-retest reliability of baseline cognitive assessments using ImPACT. *The* American Journal of Sports Medicine, *38 (1)*, 47–53.

Schatz, P., & Ferris, C. S. (2013). One-month test-retest reliability of the ImPACT test battery. *Archives of Clinical Neuropsychology*. doi: 10.1093/arclin/act034.

Schatz, P., Kontos, A., & Elbin, R. (2012). Response to Mayers and Redick: “clinical utility of ImPACT assessment for postconcussion return-to-play counseling: Psychometric issues.” *Journal of Clinical and Experimental Neuropsychology*, *34 (4)*, 428–434; discussion 435–442.

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 (1)*, 91–99.

Schatz, P., & Sandel, N. (2012). Sensitivity and specificity of the online version of ImPACT in high school and collegiate athletes. *The* American Journal of Sports Medicine. doi:10.1177/0363546512466038.

Schmidt, J. D., Register-Mihalik, J. K., Mihalik, J. P., Kerr, Z. Y., & Guskiewicz, K. M. (2012). Identifying impairments after concussion: Normative data versus individualized baselines. *Medicine and Science in Sports and Exercise*, *44 (9)*, 1621–1628.

Van Kampen, D. A., Lovell, M. R., Pardini, J. E., Collins, M. W., & Fu, F. H. (2006). The “value added” of neurocognitive testing after sports-related concussion. The *American Journal of Sports Medicine*, *34 (10)*, 1630–1635.